Les enquêtes longitudinales : de la conception à l'analyse

Recueil

XXV Symposium international de Statistique Canada sur les questions de méthodologie

Du 27 au 30 octobre 2009

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 2009 - Catalogue no. 11-522-XCB

Utilisation finale du contrat de licence

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2012

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

LES ENQUÊTES LONGITUDINALES : DE LA CONCEPTION À L'ANALYSE

TABLE DES MATIÈRES

PRÉFACE **8**

MOT D'OUVERTURE **10**

François Maranda, Statistique Canada

SESSION 1 – DISCOURS PRINCIPAL

01-1 Un programme de recherche méthodologique pour les enquêtes longitudinales **13**

Peter Lynn, Université de Essex, Royaume-Uni

SESSION 2 – COLLECTE DE DONNEES BIOGRAPHIQUES

02-1 Une méthode d'évaluation par triangulation du calendrier rétrospectif de l'étude ELSA **15**

Alice McGee et Hayley Cripps, National Centre for Social Research, Royaume-Uni

Joanne Pascale, U.S. Census Bureau, É.-U.

02-2 L'étude des cycles de vie au moyen d'enquêtes prospectives par panel : leçons tirées de l'étude par panel auprès des familles allemandes **16**

Josef Brüderl, Laura Castiglioni, Ulrich Krieger, Volker Ludwig et Klaus Pforr, Université de Mannheim, Allemagne

02-3 Comparaison des rapports d'enquête obtenus au moyen d'un questionnaire standard et d'un calendrier rétrospectif **17**

Jeffrey Moore, Jason Fields, Joanne Pascale, Gary Benedetto, Martha Stinson et Anna Chan, U.S. Census Bureau, É.-U.

SESSION 3 – BIAIS D'EROSION ET PONDERATION POUR LA NON-REPONSE

03-1 Évaluation et sélection de modèles pour l'ajustement de la non-réponse due à l'attrition **19**

Eric Slud et Leroy Bailey, U.S. Census Bureau, É.-U.

03-2 L'ajustement des poids de non-réponse au moyen de l'imputation multiple dans la Millenium Cohort Study du Royaume-Uni **28**

John W. McDonald et Sosthenes C. Ketende, Université de Londres, Royaume-Uni

03-3 Analyse de la non-réponse dans l'Enquête longitudinale nationale sur les enfants et les jeunes **36**

Mike Tam et Agnes Waye, Statistique Canada

SESSION 4 – COLLECTE ET APPARIEMENT DE DONNEES

04-1 Maintenir le contact avec les familles entre les cycles de la PSID : essai expérimental d'une nouvelle stratégie **46**

Katherine McGonagle, Mick Couper et Robert Schoeni, Université du Michigan, É.-U.

04-2 Les modes de collecte mixtes et multiples : l'expérience de l'enquête HILDA 52

Nicole Watson, Mark Wooden, Université de Melbourne, Australie

04-3 La gestion d'un appariement inexact complexe dans les applications de codage et de couplage 59

Michael J. Wenzowski , Statistique Canada

04-4 La gestion des relations avec les répondants à l'Enquête nationale sur la santé de la population 60

Andrew MacKenzie et Natasha Zaletel, Statistique Canada

SESSION 5 – ANALYSE DES DONNEES D'ENQUETES LONGITUDINALES

05-1 Étude par simulation d'approches par calage pour l'estimation de flux bruts 62

Marcel de Toledo Vieira, Université Fédérale de Juiz de Fora, Brésil

Gad Nathan, Université Hébraïque de Jérusalem, Israël

05-2 Suivis impossibles et modélisation des périodes de chômage de l'EDTR à l'aide du modèle de régression à risques proportionnels de Cox 63

Dagmar Mariaca Hajducek et Jerry Lawless, Université de Waterloo, Canada

05-3 Les enjeux de l'utilisation de la modélisation par équations structurelles dans les fichiers de données longitudinales à grande diffusion 64

Laura Stapleton, Université du Maryland, É.-U.

SESSION 6 – ENJEUX METHODOLOGIQUES DANS LES ENQUETES ECONOMIQUES

06-1 L'élaboration d'un prototype du programme LEED italien fondé sur des données administratives : principaux aspects méthodologiques 74

Andrea Colace, M.Carla Congia et Roberta Rizzi, ISTAT, Italie

06-2 Les plans de sondage des enquêtes sur les prix sont-ils suffisamment robustes au vieillissement des facteurs de pondération? Étude par simulation 75

Zdenek Patak, Statistique Canada

Daniele Toninelli, Université de Bergamo, Italie

06-3 L'ajout d'une composante longitudinale au programme des données fiscales agricoles de Statistique Canada 87

Terri Blanchard et Peter Xiao, Statistique Canada

06-4 Les enquêtes longitudinales auprès de populations difficiles à retracer 93

E.J. Reedy, Kauffman Foundation, É.-U.

SESSION 7 – ABORDER LA CONFIDENTIALITE A L'AIDE DE DONNEES SYNTHETIQUES

07-1 La validité analytique et la protection de la confidentialité dans les systèmes de données statistiques à intégration longitudinale 101

John M. Abowd, Université Cornell, É.-U.

07-2 Sommaire des méthodes et évaluation provisoire de l'ensemble de données SIPP Synthetic Beta, version 5.0 102

Gary Benedetto et Martha Stinson, U.S. Census Bureau, É.-U.

Melissa Bjelland, Université Cornell, É.-U.

07-3 La création de données synthétiques pour le Cross National Equivalent File..... 103

Cynthia Bocci et Jean-François Beaumont, Statistique Canada

SESSION 8 – DISCOURS DU GAGNANT DU PRIX WAKSBERG

08-1 Méthodes de suréchantillonnage de sous-populations rares dans les enquêtes sociales 112

Graham Kalton, Westat, É.-U.

SESSION 9 – DONNEES LONGITUDINALES SUR LA SANTE: ENJEUX ET DEFIS

09-1 Établissement d'une méthode longitudinale de recherche sur la santé communautaire : enjeux et difficultés..... 114

David Marshall, Université du Queensland, Australie

09-2 Analyse de l'approche longitudinale de la santé mise en place en Belgique 121

Ann Ingenbleek, Yves Coppieters et Alain Levêque, Université Libre de Bruxelles, Belgique

Lies Lammens et Patrick Deboosere, Vrije Universiteit Brussel, Belgique

Florence Cols et William D'hoore, Université Catholique de Louvain, Belgique

09-3 Les aspects éthiques de la collecte de données longitudinales aux niveaux individuel et sociétal..... 127

Lies Lammens et Patrick Deboosere, Vrije Universiteit Brussel, Belgique

Florence Cols et William D'hoore, Université Catholique de Louvain, Belgique

Ann Ingenbleek, Yves Coppieters et Alain Levêque, Université Libre de Bruxelles, Belgique

09-4 L'apport des bases administratives et médico-administratives dans la cohorte Constances 133

Gueguen, R. Sitta, JL. Lanoe, M. Goldberg et M. Zins, INSERM, France

L. Bénézet et G. Santin, Institut de veille sanitaire, France

SESSION 10 – ENJEUX LIES A LA COLLECTE DE DONNEES D'ENQUETES LONGITUDINALES

10-1 Maintenir le contact avec les familles mobiles de la Millenium Cohort Study du Royaume-Uni..... 139

Lisa Calderwood, Université de Londres, Royaume-Uni

10-2 Organisation et monitoring du terrain d'enquête : impact sur la qualité des estimateurs pour un panel rotatif de ménages..... 148

Thomas Christin, Stéphane Fleury et Johan Pea, Office fédéral de la statistique, Suisse

10-3 Un plan de sondage réactif pour l'Enquête sur la dynamique du travail et du revenu..... 157

Tracy Tabuchi, François Laflamme, Owen Phillips, Milana Karaganis, et Amélie Villeneuve, Statistique Canada

SESSION 11 – PONDERATION ET ESTIMATION

11-1 L'ajustement de la pondération à l'aide de scores de propension dans un plan à base de sondage double 166

C. Boudreau, M.E. Thompson et M. Iraniparast, Université de Waterloo, Canada

11-2 L'estimation longitudinale dans l'enquête européenne sur le revenu et les conditions de vie 167

Ralf Münnich et Stefan Zins, Université de Trier, Allemagne

11-3 La pondération et l'estimation de la variance dans l'enquête par panel à base de sondage double menée auprès des ménages allemands (PASS) 174

Hans Kiesl, Institute for Employment Research (IAB), Allemagne

SESSION 12 – TENIR COMPTE DES DONNEES MANQUANTES DANS L'ANALYSE DE DONNEES D'ENQUETES LONGITUDINALES

12-1 La modélisation et l'analyse des durées d'après les données d'enquêtes longitudinales 176

Jerry Lawless et Dagmar Mariaca Hajducek, Université de Waterloo, Canada

12-2 L'analyse d'enquêtes longitudinales comportant des réponses manquantes..... 184

Changbao Wu, Université de Waterloo, Canada

Ivan Carrillo Garcia, Statistique Canada

12-3 Les études longitudinales comportant des réponses et des covariables manquantes : application à l'étude ITC 4 185

Baojiang Chen, Université de Washington, É.-U.

Mary Thompson, Université de Waterloo, Canada

SESSION 13 – FACTEURS ET EFFETS DE LA NON-REPONSE

13-1 Les facteurs liés aux différents profils de non-réponse dans la English Longitudinal Study of Ageing (ELSA) 195

Hayley Cheshire et David Hussey, National Centre for Social Research, Royaume-Uni

13-2 Étude empirique du biais de non-réponse dû à l'érosion de l'échantillon de la National Survey of College Graduates (NSCG)..... 196

Donsig Jang, Mathematica Policy Research, É.-U.

John Finamore et David Hall, U.S. Census Bureau, É.-U.

Steve Cohen, Flora Lan et Fan Zhang, National Science Foundation, É.-U.

13-3 Facteurs associés à la participation dans la cohorte GAZEL 197

Marie Zins, Jean François Chastang, Mireille Coeuret-Pellicer, Annette Leclerc, Sébastien Bonenfant, Alice Guéguen, Anna Ozguler et Marcel Goldberg, INSERM, France

13-4 Stratégies pour étudier le biais de non-réponse dans les cohortes Coset (Cohorte santé et travail) et Constances (Cohorte des consultants des centres d'examens de santé) 198

Laetitia Bénézet, Gaëlle Santin, Stéphanie Gauvin, Hélène Sarter et Béatrice Geoffroy-Perez, Institut de veille sanitaire, France

Alice Guéguen, Rémi Sitta, Marie Zins et Marcel Goldberg, INSERM, France

Nicolas Razafindratsima, Institut National d'études démographiques, France

SESSION 14 – ENJEUX METHODOLOGIQUES DIVERS

14-1 Répartition de l'échantillon de la 2010 Decade of the National Survey of College Graduates..... 205

John Finamore et David Hall, U.S. Census Bureau, É.-U.

Donsig Jang, Mathematica Policy Research, É.-U.

Stephen Cohen, Flora Lan et Fan Zhang, National Science Foundation, É.-U.

14-2 Life Pathways Project : conception et enjeux méthodologiques 206

Trivina Kang, Melvin Chan, Tan Teck Kiang et David Hogan, Nanyang Technological University, Singapour

14-3 La conception et l'analyse de données longitudinales à Statistique Nouvelle-Zélande 210

Deborah Brunning, Statistics New Zealand, Nouvelle-Zélande

SESSION 15 – REMANIEMENT D'ENQUETES LONGITUDINALES DE GRANDE ENVERGURE

15-1 La continuité et l'innovation dans la conception de Understanding Society: The UK Household Longitudinal Study 216

Heather Laurie, Université de Essex, Royaume-Uni

15-2 La survie et la reprise de la Survey of Income and Program Participation 225

S. Johnson, U.S. Census Bureau, É.-U.

15-3 Les résultats de l'Enquête pilote par panel auprès des ménages canadiens 226

Andrew Heisz, Statistique Canada

SESSION 16 – MODELISATION DE VARIABLES LATENTES ET ESTIMATION BAYESIENNE

16-1 Évolution de la satisfaction de vivre dans la British Household Panel Survey : modélisation de la courbe de croissance latente 235

Maria de Fátima Salgueiro, ISCTE Business School, Portugal

Marcel de Toledo Vieira, Université Fédérale de Juiz de Fora, Brésil

Peter W. F. Smith, Université de Southampton, Royaume-Uni

16-2 Une méthode d'analyse de transition latente pour modéliser l'évolution de l'hétérogénéité non observée d'une population 241

Andy Ross, National Centre for Social Research, Royaume-Uni

16-3 Les modèles d'appartenances mixtes longitudinaux des données d'enquête sur l'invalidité 242

Daniel Manrique-Vallier et Stephen E. Fienberg, Université Carnegie Mellon, É.-U.

SESSION 17 – ERREURS DE MESURE

17-1 La non-réponse et l'erreur de mesure dans la recherche sur l'emploi..... 250

Frauke Kreuter, JPSM Université du Maryland, É.-U.

Gerrit Mueller et Mark Trappmann, IAB Institute for Employment Research, Allemagne

17-2 Incohérences dans les caractéristiques d'emploi déclarées par les salariés qui conservent leur emploi : données tirées d'une série de panels à deux cycles de l'enquête italienne sur la population active, 1993-2003 251

Francesca Bassi et Ugo Trivellato, Université de Padova, Italie

Alessandra Padoan, ISTAT, Italie

17-3 Le chevauchement des lisières dans l'enquête HILDA : problèmes et observations 252

Nicole Watson, Université de Melbourne, Australie

SESSION 18 – IMPUTATION

18-1 L'utilité de l'imputation dans les enquêtes longitudinales 260

Roberto Gismondi, ISTAT, Italie

18-2 L'imputation aléatoire équilibrée dans les enquêtes..... 266

David Haziza, Université de Montréal, Canada

Guillaume Chauvet et Jean-Claude Deville, Laboratoire de Statistique d'Enquête (CREST/ENSAI), France

18-3 La mise à l'essai de nouvelles méthodes d'imputation des gains dans la Survey of Income and Program Participation 267

Martha Stinson et Gary Benedetto, U.S. Census Bureau, É.-U.

SESSION 19 – VERIFICATION ET IMPUTATION

19-1 Le point sur le projet EU-SILC en Slovénie 269

Rudi Seljak, Statistical Office of the Republic of Slovenia, Slovénie

19-2 La vérification des données longitudinales de l'enquête italienne sur la population active 274

Simona Rosati et Barbara Boschetto, ISTAT, Italie

19-3 Imputation de registres longitudinaux : le cas des ménages 281

D.J. (Jan) van der Laan et Léander Kuijvenhoven, Statistics Netherlands, Pays-Bas

SESSION 20 – APPLICATION : ANALYSE LONGITUDINALE DE DONNEES SUR LA SANTE ET SUR LES ENTREPRISES

20-1 Les enfants des femmes plus âgées qui sont mères pour la première fois au Canada : une analyse longitudinale de leur santé et de leur développement 287

Tracey Bushnik et Rochelle Garner, Statistique Canada

20-2 L'évolution de l'IMC et de la taille au cours du cycle de vie : comparaison de deux cohortes de naissances britanniques..... 293

Leah Li, Rebecca Hardy, Diana Kuh et Chris Power, Collège Universitaire de Londres, Royaume-Uni

20-3 L'impact de la formation sur la productivité des entreprises canadiennes dans un contexte longitudinal : la comparaison d'un modèle additif et d'un modèle interactif 294

Amélie Bernier et Jean-Michel Cousineau, Université de Montréal, Canada

20-4 Mobilité des travailleurs : sommaire et nouveaux résultats de l'Enquête sur le milieu de travail et les employés..... 300

Yves Decady, Statistique Canada

SESSION 21 – **TECHNIQUES D'ANALYSE DE DONNEES LONGITUDINALES**

21-1 L'utilisation de techniques d'analyse exploratoire et confirmative de données longitudinales 308

Marcel de Toledo Vieira, Ronaldo Rocha Bastos et Henrique Steinherz Hippert, Université Fédérale de Juiz de Fora, Brésil

Augusto Carvalho Souza, Université Fédérale de Minas Gerais, Brésil

21-2 Mesures de la qualité de l'ajustement pour les modèles fondés sur une approche des équations d'estimation généralisées 309

Punam Pahwa, Université de Saskatchewan, Canada

SESSION 22 – AJUSTEMENTS POUR LA NON-REPONSE ET L'EROSION

22-1 L'érosion de l'échantillon dans les études de cohorte : tendances, caractéristiques et ajustements..... 311

Ian Plewis, Université de Manchester, Royaume-Uni

Lisa Calderwood, Sosthenes Ketende et Rebecca Taylor, Institute of Education, Royaume-Uni

22-2 Analyse de l'érosion de l'Étude Longitudinale sur le Développement des Enfants du Québec (ÉLDEQ) de 1998 À 2008..... 317

Catherine Fontaine et Robert Courtemanche, Institut de la statistique du Québec, Canada

22-3 Modélisation de la non-réponse pour une enquête longitudinale à l'aide de paradonnées : Application à l'enquête sur la dynamique du travail et du revenu 323

Beatrice Baribeau et Wisner Jocelyn, Statistique Canada

SYMPOSIUM DE STATISTIQUE CANADA – BANQUET DU 25ÈME ANNIVERSAIRE 330

Gordon Brackstone

Préface

Le Symposium 2009 était le 25e Symposium international sur les questions de méthodologie de Statistique Canada. Chaque année, le Symposium se penche sur un thème précis. En 2009, le thème était : **« Les enquêtes longitudinales : de la conception à l'analyse ».**

Le Symposium 2009 a eu lieu du 27 au 30 octobre 2009 au Palais des congrès de Gatineau, au Québec; plus de 500 personnes provenant de différents pays y ont assisté. En tout, deux ateliers de travail et 69 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 du Symposium 2009 souhaitent souligner la contribution des différentes personnes — trop nombreuses pour être nommées individuellement — 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 remercient tous les participants qui ont suivi les différentes présentations.

Le comité organisateur du Symposium 2009

Logistique et opérations
Julie Trépanier, Présidente
Asma Alavi
Jean-Luc Bernier
Sylvie Gauthier
Julie Girard
Denis Lemire
Caroline Rondeau

Programme
Christian Nadeau, Président
Abdelnasser Saïdi
Claude Turmelle

MOT D’OUVERTURE

Mot d'ouverture

François Maranda[1]

Bonjour,

Au nom de Statistique Canada, il me fait plaisir de vous souhaiter, chers collègues et amis, la plus cordiale des bienvenues au Symposium 2009.

Cette année, notre symposium, qui a pour thème « Les enquêtes longitudinales : de la conception à l'analyse » revêt un caractère spécial. En effet, il s'agit du 25e Symposium international organisé par Statistique Canada sur des questions de méthodologie d'enquête, le premier ayant été tenu en 1984 sous le thème « L'analyse des données d'enquête ».

Que vous soyez des habitués de ce symposium ou que vous vous joigniez à nous pour la première fois, nous sommes heureux de vous compter parmi nous pour célébrer ce 25e anniversaire.

Entre le premier symposium et celui-ci, plusieurs thèmes, tous plus pertinents les uns que les autres, ont été abordés au fil des ans. Je vous invite à les découvrir ou à les redécouvrir en visitant l'exposition qui est présentée au Foyer tout au long de la semaine.

Avant d'aborder le thème de cette année, permettez-moi de proposer une explication quant à la longévité de ce symposium. Celui-ci a été instauré dans le but de contribuer à l'avancement de la méthodologie d'enquête en tant que discipline et de favoriser la résolution d'enjeux communs à divers organismes statistiques, et ce, tout en rassemblant des experts de divers horizons et en créant un climat de collaboration. La longévité du symposium et l'intérêt sans cesse accru qu'il suscite me portent à croire que cet objectif a été pleinement atteint.

Du point de vue de Statistique Canada, le symposium constitue encore aujourd'hui une source d'information inestimable dans le cadre des efforts déployés pour faire face aux défis courants et émergeants auxquels nous sommes confrontés en matière de méthodologie d'enquête. La pertinence et la valeur des échanges émanant du symposium sont intimement liées à votre présence, permettant ainsi, année après année, de mettre en commun des points de vue, des expériences et des expertises diversifiés pour le bénéfice de tous les participants.

En nous offrant l'occasion d'apprendre comment d'autres organismes relèvent des défis similaires aux nôtres, le symposium contribue à la fois à la résolution des enjeux qui nous préoccupent, à l'examen critique de nos travaux et au perfectionnement de notre personnel.

À en juger par votre nombre et votre provenance, il semble que les retombées positives de nos symposiums soient également reconnues sur la scène internationale. Outre les 300 participants de Statistique Canada, vous êtes environ 175 personnes provenant de 16 pays situés aux quatre coins du monde. Ensemble, vous représentez plus de 90 organisations tant du milieu universitaire que des secteurs privé et public. C'est vraiment impressionnant. Je vous remercie de votre présence.

Revenons maintenant au thème de cette année, « Les enquêtes longitudinales : de la conception à l'analyse ». Ce n'est pas la première fois que nous abordons les enquêtes longitudinales lors du symposium. En 1989, quelques

[1] François Maranda, Statistique Canada, 26-J Immeuble R.-H.-Coats, 100, promenade Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6

présentations traitaient de l'analyse de données longitudinales lors d'un symposium sous le thème de « L'analyse des données dans le temps ». Puis en 1992, le symposium « Conception et analyse des enquêtes longitudinales » s'inscrivait dans le cadre de l'entrée de Statistique Canada dans l'ère des enquêtes longitudinales. L'Enquête sur la dynamique du travail et du revenu devait en effet être lancée quelques mois plus tard alors que l'Enquête nationale sur la santé de la population et l'Enquête longitudinale nationale sur les enfants et les jeunes étaient déjà prévues pour l'année suivante. À elles trois, ces enquêtes constituaient une base solide permettant d'aborder la dynamique de la pauvreté, du travail, de la famille, de la santé et du développement des enfants.

D'autres initiatives ont par la suite vu le jour, toujours du côté de la statistique sociale, avec l'Enquête sur les jeunes en transition et l'Enquête longitudinale sur les immigrants au Canada qui ont respectivement été mises en œuvre en 2000 et 2001. Celles-ci permettaient alors de mettre l'accent sur l'étude de sous-populations vivant, ou sur le point de vivre, une période de transition déterminante dans leur vie, soit respectivement la transition entre les études et le marché du travail et l'intégration à la société canadienne suite à une récente immigration. Au cours de la même période, du côté de la statistique économique, une initiative d'enquête longitudinale voyait le jour avec l'Enquête sur le milieu de travail et les employés dont la collecte des données débutait en 1999.

La fin des années 1990 a donc vu se concrétiser la disponibilité d'ensembles de données longitudinales. Bien que l'analyse ait été abordée lors de symposiums précédents, la tenue en 1998 d'un symposium sur le thème « Analyse longitudinale pour les enquêtes complexes » a permis de faire le point sur la richesse et le potentiel analytique de même que sur l'état des connaissances dans le domaine de l'analyse de telles données. Pour Statistique Canada, ce symposium s'inscrivait également dans le cadre d'initiatives visant à favoriser la réalisation de recherche en analyse de données longitudinales.

Mais pourquoi en 2009 aborder de nouveau le thème des enquêtes longitudinales? Étant donné la diversité des expériences acquises au fil du temps, il semble que nous sommes maintenant arrivés à l'étape des bilans, et ce dans le but de toujours mieux répondre aux besoins en données. Le temps constitue ici l'élément clé. À la différence d'enquêtes purement transversales, la pertinence des données recueillies devrait s'accroître avec le temps. D'un autre côté, le phénomène d'érosion des échantillons, intimement relié au temps, contribue à réduire l'exactitude des ensembles de données longitudinales. De plus, seul le temps peut fournir des résultats d'évaluations permettant de juger de la pertinence des choix initiaux et ultérieurs, autant ceux qui ont trait au contenu qu'à la méthodologie. Il est maintenant temps d'évaluer ces expériences et de décider des orientations futures de telles enquêtes. Parmi les enjeux auxquels nous serons confrontés, notons ceux liés à la durée des enquêtes longitudinales, à la représentativité des échantillons et à l'érosion y étant associées, à l'intégration des contenus et au fardeau de réponse, à la comparabilité internationale, à l'accessibilité aux données et à la protection de la vie privée.

Au cours des trois prochains jours, nous pourrons assister à des exposés traitant de plusieurs de ces aspects, en plus de certains autres portant sur la collecte, les erreurs de mesure et de non-réponse, la pondération et l'estimation, l'utilisation de données administratives, la modélisation et l'analyse des données. J'aimerais remercier tous les présentateurs venus partagés le fruit de leurs travaux.

Hier, 140 d'entre vous avez assisté à l'un ou l'autre des deux ateliers d'une journée, le premier portant sur les leçons et innovations émanant des 15 années d'expérience de Statistique Canada en matière d'enquête longitudinale et le second sur la modélisation multiniveau des données longitudinales. J'aimerais remercier particulièrement les animateurs des ateliers d'hier pour leur importante contribution, soit Sophia Rabe-Hesketh de l'Université de Californie et Anders Skrondal de l'Institut norvégien de santé publique, de même que Michelle Simard et François Brisebois de Statistique Canada.

Finalement, avant de déclarer officiellement ouvert ce 25^{e} symposium, j'aimerais remercier le comité organisateur, les présidents et les organisateurs de sessions ainsi que tous les bénévoles qui ont rendu cette rencontre possible.

À chacune et à chacun, mes sincères remerciements.

SESSION 1 – DISCOURS PRINCIPAL

Un programme de recherche méthodologique pour les enquêtes longitudinales

Peter Lynn[1]

Résumé

Cet exposé propose un programme de recherche en méthodologie à venir sur les enjeux relatifs aux enquêtes longitudinales. Le programme porte sur les défis méthodologiques qui sont uniques aux enquêtes longitudinales et sur un examen de la recherche destinée à relever ces défis. L'accent sera mis principalement sur la recherche récente et actuelle, et une discussion se déroulera à propos des répercussions des résultats de la recherche récente, des transformations technologiques et autres innovations. Les objectifs sont de stimuler la recherche en méthodologie et de souligner la sensibilisation aux limites du savoir méthodologique actuel.

Les sujets abordés incluent le plan de sondage, les intervalles entre les vagues, le suivi des membres de l'échantillon et le maintient de la coopération, l'ajustement de la non réponse et de l'érosion du panel, l'accoutumance du panel, et la conception d'un instrument qui minimise l'erreur de mesure des changements.

[1] Peter Lynn, University of Essex, Royaume-Uni (plynn@essex.ac.uk)

SESSION 2 – COLLECTE DE DONNÉES BIOGRAPHIQUES

Recueil du Symposium 2009 de Statistique Canada
Les enquêtes longitudinales : de la conception à l'analyse

Une méthode d'évaluation par triangulation du calendrier rétrospectif de l'étude ELSA

Alice McGee, Hayley Cripps et Joanne Pascale[1]

Résumé

Utilisée dans les enquêtes depuis plusieurs décennies, la méthode du calendrier rétrospectif s'impose de plus en plus, grâce en partie aux récents travaux de recherche montrant qu'à l'égard de certains sujets, elle peut produire des données d'une plus grande qualité que les questionnaires conventionnels. Toutefois, une recherche systématique s'avère nécessaire pour éclairer deux aspects importants de cette méthode. D'une part, comment fonctionne le calendrier rétrospectif dans la pratique, notamment en ce qui concerne l'interaction intervieweur-répondant, l'utilisation de la méthode et la réceptivité dont elle fait l'objet? D'autre part, quelle est la « mécanique » des points de repère internes et externes, qui les propose comme aide-mémoire et à quel moment, et dans quelle mesure aident-ils les répondants à se rappeler les dates avec exactitude?

Cet exposé décrit une étude qui visait à déterminer ces inconnues en évaluant un calendrier rétrospectif utilisé dans le cadre de la English Longitudinal Study of Ageing (ELSA). On a enregistré un échantillon de 124 interviews qu'on a analysées au moyen de diverses méthodes d'évaluation quantitative et qualitative : un questionnaire de l'intervieweur, un questionnaire de compte rendu du répondant, un compte rendu de l'intervieweur et un codage comportemental.

À l'aide d'une méthode de triangulation, cet exposé présente les constatations tirées de ces quatre sources de données, axées sur six questions de recherche spécifiques. Il décrit la nature précise de l'interaction entre l'interviewer et le répondant, aborde la réceptivité à la méthode et l'effet de certaines caractéristiques sur le terrain, ainsi que la nature de l'utilisation des points de repère et la mesure dans laquelle ils ont servi d'aide-mémoire aux répondants.

[1] Alice McGee et Hayley Cripps, National Centre for Social Research, Royaume-Uni; Joanne Pascale, U.S. Census Bureau, États-Unis (joanne.pascale@census.gov)

L'étude des cycles de vie au moyen d'enquêtes prospectives par panel : leçons tirées de l'étude par panel auprès des familles allemandes

Josef Brüderl, Laura Castiglioni, Ulrich Krieger, Volker Ludwig et Klaus Pforr [1]

Résumé

La collecte de renseignements valides sur les événements, sur le moment où ils se produisent et sur leur alternance au cours du cycle de vie constitue un enjeu crucial pour les chercheurs. Un plan d'enquête prospective par panel permet d'améliorer la qualité des données, car le répondant n'a à se rappeler que de brèves périodes entre les cycles de panel. À la « lisière » des deux cycles de panel, toutefois, les erreurs de déclaration constituent un problème courant (effet de lisière).

Dans cet exposé, nous soutenons qu'on peut réduire considérablement l'effet de lisière en combinant les méthodes du calendrier rétrospectif et de l'interview avec rétroinformation. En nous fondant sur les données préliminaires de la toute nouvelle étude par panel auprès des familles allemandes, nous montrons comment cette approche permet de réduire l'effet de lisière.

Dans le deuxième cycle de notre étude préliminaire, nous avons administré un questionnaire papier-crayon à calendrier rétrospectif allant de l'âge de 14 ans jusqu'à la date de l'interview et couvrant l'historique des relations, de la fécondité et de la mobilité résidentielle. Dans le troisième cycle, nous avons demandé aux répondants quels événements s'étaient produits entre les interviews. À l'aide d'un plan d'enquête avec un questionnaire à deux formes, nous avons présenté soit un calendrier en blanc, soit un calendrier où figurait déjà le statut au deuxième cycle (interview avec rétroinformation). Les écarts entre les deux groupes révèlent que l'interview avec rétroinformation permet de réduire l'effet de lisière et, par conséquent, d'améliorer la qualité des données.

[1] Josef Brüderl (jbruederl@sowi.uni-mannheim.de), Laura Castiglioni, Ulrich Krieger, Volker Ludwig and Klaus Pforr, University of Mannheim, Allemagne

Recueil du Symposium 2009 de Statistique Canada
Les enquêtes longitudinales : de la conception à l'analyse

Comparaison des rapports d'enquête obtenus au moyen d'un questionnaire standard et d'un calendrier rétrospectif

Jeffrey Moore, Jason Fields, Joanne Pascale, Gary Benedetto, Martha Stinson et Anna Chan [1]

Résumé

La Survey of Income and Program Participation (SIPP) menée par le Census Bureau des États-Unis fournit des données mensuelles sur le revenu et la richesse de la population et sur sa participation aux programmes publics. À l'heure actuelle, la SIPP interviewe chaque membre de l'échantillon trois fois par année, la période de référence de chaque interview couvrant les quatre mois civils précédents. En 2006, le Census Bureau a entrepris de remanier la SIPP en menant désormais une seule interview annuelle couvrant l'année civile précédente. À cette fin, le Census Bureau entend employer des méthodes axées sur le calendrier rétrospectif. Toutefois, des études antérieures ont soulevé certaines questions quant à la qualité des données obtenues à l'aide du calendrier rétrospectif sur des sujets d'une importance capitale pour la SIPP, comme la participation aux programmes en fonction des besoins. De plus, la base de recherche ne tient pas compte de la principale modification apportée au plan d'enquête de la SIPP, soit l'intention de faire passer la période de référence de quatre à douze mois. Pour examiner les répercussions qu'auraient l'adoption du calendrier rétrospectif et la prolongation de la période de référence, le Census Bureau a procédé, au printemps 2008, à l'essai sur le terrain du calendrier rétrospectif de la SIPP. L'essai consistait essentiellement à réinterviewer à l'aide du calendrier rétrospectif les ménages du panel expiré de la SIPP de 2004. La période de référence de la réinterview était l'année civile 2007; l'échantillon était composé principalement de ménages qui avaient déjà fourni des renseignements sur l'année civile 2007 dans le cours normal de leurs trois dernières interviews dans le cadre de la SIPP. L'essai sur le terrain permet donc de comparer directement les rapports obtenus au moyen du questionnaire standard et du calendrier rétrospectif auprès des mêmes personnes, au sujet des mêmes caractéristiques et pour la même période. Cet exposé porte sur la composante principale de l'évaluation des résultats de l'essai sur le terrain : un examen de la correspondance des deux rapports – l'un obtenu d'après un questionnaire standard, l'autre d'après un instrument à calendrier rétrospectif – à l'égard de plusieurs caractéristiques essentielles visées par la SIPP et pour chaque mois de 2007.

[1] Jeffrey Moore, Jason Fields (jason.m.fields@census.gov), Joanne Pascale (joanne.pascale@census.gov), Gary Benedetto (gary.linus.benedetto@census.gov), Martha Stinson (martha.stinson@census.gov) and Anna Chan, U.S. Census Bureau, É.-U.

SESSION 3 – BIAIS D'ÉROSION ET PONDÉRATION POUR LA NON-RÉPONSE

Recueil du Symposium 2009 de Statistique Canada
Les enquêtes longitudinales : de la conception à l'analyse

Évaluation et sélection de modèles pour l'ajustement de la non-réponse due à l'attrition

Eric V. Slud et Leroy Bailey[1]

Résumé

Le présent article porte sur l'ajustement de la pondération pour tenir compte de l'attrition due à la non-réponse dans une enquête longitudinale telle que la Survey of Income and Program Participation (SIPP) réalisée aux États-Unis. Les facteurs de correction sont fondés sur un modèle, soit constant pour toutes les cellules, soit dérivé de régressions logistiques en utilisant les covariables auxiliaires. Nous évaluons le biais dans les totaux estimés des variables à la première vague (« vague 1 ») en comparant l'estimateur pondéré par les poids de sondage à la première vague et l'estimateur pondéré par les poids ajustés des totaux pour les mêmes variables qu'à la vague 1 fondés sur les répondants à des vagues ultérieures. De nouvelles mesures de la qualité sont définies pour les modèles utilisés afin de corriger l'attrition dans l'enquête longitudinale. Les mesures combinent les maxima cumulatifs des biais d'ajustement entre vagues estimés pour les variables de l'enquête, en se basant sur des sous-ensembles réordonnés aléatoirement de l'échantillon, comparativement aux totaux estimés. Nous estimons des bandes de confiance pour les mesures, puis nous appliquons les mesures pour classer la qualité d'une série de modèles de régression logistique et faire un choix parmi ceux-ci pour la correction de la non-réponse due à l'attrition dans la SIPP de 1996.

Mots-clés : cellule d'ajustement, régression logistique, pondération, classement aléatoire, ratissage, sous-domaine.

1. Introduction

Afin d'évaluer la qualité de l'ajustement pour l'attrition dans une enquête longitudinale, on souhaiterait certainement évaluer les biais des ajustements en utilisant des données externes sur la base de sondage et les variables de l'enquête si de telles données étaient disponibles. Malheureusement, on dispose rarement de données externes pour l'évaluation, de sorte que les spécialistes des enquêtes doivent habituellement évaluer des méthodes d'ajustement concurrentes et faire un choix en se basant sur des comparaisons portant sur des données internes provenant de l'enquête. Toutefois, peu de travaux méthodologiques portant sur la façon de mesurer les biais d'ajustement en s'appuyant sur les données internes d'une enquête longitudinale ont été publiés.

Des travaux théoriques antérieurs ont porté sur le comportement en grand échantillon des méthodes d'ajustement de la pondération pour la non-réponse fondées sur un modèle. Par exemple, Kim et Kim (2007) ont utilisé des méthodes de vraisemblance pondérées pour faire un choix entre divers modèles paramétriques pour la non-réponse aux enquêtes. L'un des rares articles où est évaluée explicitement l'efficacité de l'ajustement en utilisant uniquement des données provenant de l'enquête ajustée est celui d'Eltinge et Yansaneh (1997), qui discutent de plusieurs diagnostics et vérifications de la sensibilité en vue de définir des cellules d'ajustement de la pondération dans une enquête. Dufour et coll. (2001) ont étudié l'évaluation longitudinale interne de méthodes d'ajustement de la non-réponse dans une perspective de calage, en déterminant les ordres de grandeur des ajustements basés sur le calage au moyen d'une mesure qu'ils définissent en fonction des variations des poids au cours de plusieurs étapes d'une enquête longitudinale dont la pondération est ajustée.

Slud et Bailey (2006) ont étudié les estimations et les erreurs-types des différences, dans la Survey of Income and Program Participation (SIPP) des États-Unis, entre les totaux à la vague 1 de diverses variables de l'enquête transversale de 1996 et les totaux ajustés pour la non-réponse des mêmes variables de la vague 1 en utilisant les

[1] Eric V. Slud, US Census Bureau, SRD, 4600 Silver Hill Rd., Rm 5K004, Washington DC 20233-9100, and Math. Dept., Univ. of Maryland, College Park, Eric.V.Slud@census.gov; and Leroy Bailey, formerly of US Census Bureau, SRD.
Les opinions concernant les question statistiques et méthodologiques exprimées dans le présent article sont celles des auteurs et ne représentent pas forcément celles du U.S. Census Bureau.

données des réponses fournies à une vague ultérieure (4 ou 12) de la même enquête. Ils ont étudié la non-réponse ajustée au moyen de cellules d'ajustement ou au moyen de modèles de régression logistique pour les indicateurs de réponse à une vague ultérieure, et constaté que les biais relatifs et normalisés estimés variaient considérablement d'un modèle à l'autre. Comme dans Dufour et coll. (2001), de nombreux modèles concurrents pouvaient être définis, en se servant de variables différentes pour construire les cellules d'ajustement ou comme prédicteurs dans les régressions logistiques. Slud et Bailey (2006) ont noté qu'inclure `Pauvreté` comme prédicteur avait l'effet, comparable à celui du ratissage (*raking*), de rendre particulièrement faible le biais d'ajustement à la vague 1, estimé pour l'ensemble de l'échantillon, de la variable `Pauvreté`. Ils ont conjecturé que cet effet artificiel serait éliminé en considérant le biais de la vague 1 estimé dans un certain nombre de sous-domaines différents. Cela suggère que l'on crée une mesure composite en combinant les biais d'ajustement entre vagues estimés pour diverses variables de l'enquête sur divers sous-domaines.

Dans la présente étude, notre objectif était de concevoir une mesure pour faciliter la comparaison de diverses méthodes fondées sur un modèle d'ajustement pour la non-réponse due à l'attrition, et donc d'offrir un point de départ pour choisir entre les méthodes d'ajustements. Plusieurs études comparatives antérieures concernant les méthodes d'ajustement ont été réalisées – en particulier Rizzo et coll. (1994) qui ont comparé les modèles d'ajustement par régression logistique dans la SIPP en utilisant un grand nombre des variables que nous considérons plus loin, mais aucune ne semble avoir révélé un avantage manifeste pour une méthode par rapport aux autres.

Notre approche consiste à ajuster les poids en utilisant des modèles de la réponse à l'étape ultérieure, du type de cellule d'ajustement ou de régression logistique, pour calculer les estimations, d'après les données d'une vague ultérieure de l'enquête, des totaux de population et d'autres variables étudiées à la première vague. Nous comparons les biais relatifs des estimations des sous-totaux pondérés d'après les données d'une vague tardive par opposition à une vague précoce pour divers sous-domaines de population, en combinant ces estimations en deux mesures connexes de l'efficacité de l'ajustement, puis en utilisant les mesures pour faire un choix parmi les modèles d'ajustement pour la SIPP de 1996.

La présentation de l'article est la suivante. À la section 2, nous définissons les mesures et présentons leurs bornes, théoriquement valides quand le modèle d'ajustement est correct, qui peuvent être appliquées en pratique pour repérer des ajustements de la pondération inadéquats. À la section 3, nous appliquons les mesures à la comparaison d'une série de modèles à cellules d'ajustement ou de régression logistique qui ont été utilisés pour la correction de l'attrition dans les données de la SIPP de 1996, en étoffant celles étudiées dans Slud et Bailey (2006). Les totaux ajustés de la SIPP de 1996 et les mesures présentés à la section 3 sont basés sur les poids de la première étape (dans lesquels sont intégrés les ajustements de la non-réponse jusqu'à la vague 1) et les ajustements fondés sur un modèle à une vague ultérieure, mais non les totaux de contrôle de population. Ensuite, à la section 4, nous donnons les comparaisons correspondantes entre les totaux de la vague 1 basés sur les poids corrigés à la vague 1 et à une vague ultérieure rajustés par ratissage sur les totaux de la Current Population Survey, comme cela a été fait en réalité pour la SIPP. À la section 5, nous tirons des conclusions générales. La matière du présent article est présentée plus en détail dans Slud et Bailey (2009).

2. Mesures de la qualité de l'ajustement et bornes

Nous formulons le plan de sondage sous la forme de ce qu'on a appelé un *modèle de quasi-randomisation* (Oh et Scheuren 1983). Soit S l'échantillon de $n = |S|$ personnes tiré de la base de sondage U, avec les probabilités d'inclusion uniques connues ou effectivement ajustées $\{\pi_i : i \in U\}$, et répondant à la vague 1. Pour un ensemble de mesures d'une enquête transversale indicées par $k = 1,...,K$, telles que les $K = 11$ variables étudiées par Slud et Bailey (2006), nous désignons par $y_i^{(k)}$ les valeurs de la variable à la vague 1 et par $\mathbf{x}_i$ un vecteur de valeurs de variables auxiliaires pour tout $i \in U$. Soit ρ_i les indicateurs de réponse individuelle à une vague ultérieure spécifiée de la même enquête, supposés indépendants et observés pour tout $i \in S$, et soit $p_i = \mathrm{P}(\rho_i = 1 | S)$ les probabilités inconnues de réponse individuelle à cette vague ultérieure. Soit $\hat{p}_i = g(\mathbf{x}_i, \hat{\vartheta})$ les estimateurs de ces probabilités inconnues basés sur la fonction g connue dérivée d'un modèle paramétrique utilisant les données auxiliaires $\mathbf{x}_i$,

avec les estimateurs des paramètres $\hat{\vartheta}$ obtenus au moyen d'équations d'estimation (Kim et Kim, 2007). Pour tout attribut de la population $z_i, i \in U$, le total de la population de la base de sondage est $t_z = \sum_{i \in U} z_i$, et l'estimateur d'Horvitz-Thompson correspondant est $\hat{t}_z = \sum_{i \in S} z_i / \pi_i$.

Pour chaque variable étudiée $y_i^{(k)}, i \in U$, et la stratégie d'ajustement intégrée dans les probabilités de réponse estimées $\hat{p}_i$, définissons le biais de non-réponse estimé pour chaque sous-domaine de la population $D \subset U$ comme étant

$$\hat{B}_k(D) = \sum_{i \in D \cap S} \left((\rho_i / \hat{p}_i) - 1 \right) y_i^{(k)} / \pi_i \qquad (1)$$

Dans Slud et Bailey (2006) et dans des travaux antérieurs de Bailey, le domaine D correspondait à l'entièreté de U, et $\hat{B}_k(U)$ était considéré comme la différence entre un estimateur ajusté de $t_{y^{(k)}}$ basé sur les données $\left(\rho_i, \rho_i y_i^{(k)}, \mathbf{x}_i : i \in S \right)$ et l'estimateur d'Horvitz-Thompson ordinaire $\hat{t}_{y^{(k)}}$, et était considéré comme un estimateur du biais de non-réponse dû à l'attrition en raison de la méthode d'ajustement.

Quand des ajustements fondés sur un modèle sont intégrés dans la pondération de l'enquête pour des vagues ultérieures, surtout quand sont introduits dans les modèles des termes ayant trait à des items spécifiques, notre expérience (Slud et Bailey 2006) donne à penser que les termes de biais $\hat{B}_k(U)$ pour l'ensemble de la population peuvent être réduits beaucoup plus que les biais sur des domaines plus petits D. Notre but est de concevoir des mesures utiles du biais maximal sur de nombreux sous-domaines qui favorisent les ajustements fondés sur un modèle réduisant les biais sur les domaines intéressants et décourageant les modèles qui ne corrigent les biais des totaux des variables individuelles étudiées que sur l'ensemble de la population.

2.1 Biais de sous-domaine relatif

Nous proposons maintenant une mesure du biais relatif pour l'estimation des totaux de variable sur des sous-domaines. L'idée est de considérer la valeur la plus grande du biais relatif absolu $\hat{B}_k(D) \big/ \hat{t}_{y^{(k)}}$ sur une série de sous-ensembles $D \subset U$. Après reclassement des éléments de S par la permutation aléatoire $\tau = (\tau(1), \tau(2), ..., \tau(n))$, le biais absolu le plus grand dans la variable étudiée *k* sur des sous-domaines consécutivement τ -indicés de S est

$$\max \left\{ \left| \hat{B}_k \left(\{ \tau(1), ..., \tau(\alpha) \} \right) \right| : 1 \le \alpha \le n \right\}.$$

Par conséquent, pour mesurer le biais relatif global lorsque l'on estime les totaux de la variable *k* sur les sous-domaines, définissons

$$m_k = E_\tau \left(\max_{1 \le \alpha \le n} \left| \hat{B}_k \left(\{ \tau(1), ..., \tau(\alpha) \} \right) \right| \right) \Big/ \hat{t}_{y^{(k)}} \qquad (2)$$

où l'espérance est calculée, pour un échantillon fixe, sur les permutations aléatoires τ choisies de manière équiprobable parmi les *n!* permutations des éléments de S.

Dans les conditions où le biais relatif estimé $\delta^{(k)} \equiv \hat{B}_k(U) \big/ \hat{t}_{y^{(k)}}$ sur l'échantillon complet est grand, m_k et son estimateur $\hat{m}_k$ s'avèrent n'être pas très différents de $\left| \delta^{(k)} \right|$. Cependant, si $\delta^{(k)}$ est faible, alors $\hat{m}_k$ peut être beaucoup plus grand, puisque l'ajustement du modèle ne comporte peut-être pas simultanément l'ajustement pour les totaux $y_i^{(k)}$ pondérés sur des sous-ensembles types de l'échantillon triés selon τ.

La mesure m_k dépend uniquement des données d'échantillon. Bien qu'elle soit trop compliquée pour être évaluée exactement, on obtient une bonne estimation en calculant à la place la moyenne de l'espérance sur les permutations aléatoires τ en utilisant la stratégie qui suit, justifiée dans Slud et Bailey (2009). En désignant par R un nombre moyennement grand d'itérations Monte Carlo, définissons une matrice $\left(V_{ri}, 1 \le r \le R, i \in S\right)$ de variables indépendantes suivant une loi uniforme (0, 1). Alors, un estimateur Monte Carlo convergent de (2) est donné par

$$\hat{m}_k = \left(R\hat{t}_{y^{(k)}}\right)^{-1} \sum_{1 \le r \le R} \max_{0 \le x \le 1} \left| \sum_{i \in S: V_{ri} \le x} \left(\left(\rho_i / \hat{p}_i\right) - 1 \right) y_i^{(k)} / \pi_i \right| . \tag{3}$$

La qualité de l'estimation de m_k en termes de $\hat{m}_k$, et la relation entre celles-ci et $\left|\delta^{(k)}\right|$ sont examinées plus bas, à la section 2.3. Nous commençons par modifier (2) et (3) afin de permettre seulement les reclassements aléatoires des indices d'échantillon qui préservent des cellules distinctes de la population, telles que les cellules sur lesquelles les totaux de population seraient calés par ratissage, ou les sous-domaines de population qui présentent un intérêt particulier pour les utilisateurs des données.

2.2 Mesure du biais de sous-domaine sur des cellules spécifiées

La plupart des permutations aléatoires de l'échantillon détruisent entièrement toute signification des sous-domaines d'échantillon. Pourtant, l'idée qui sous-tend le ratissage ou le calage est que certains totaux de sous-domaines estimés, par exemple ceux sur les cellules d'une partition géographique-démographique spécifiée $\{A_c\}_{c=1}^{C}$ de la base de sondage, devraient être contraintes d'être égales aux totaux de cellules (les contrôles) d'un recensement courant (à jour). Donc, il est logique de mesurer les estimations du biais $\hat{B}_k\left(A_c\right)$ sur ces cellules, et de supposer à partir de maintenant qu'une partition A de U en les cellules $A_c, c = 1,..., C$ a été fixée. Par conséquent, nous modifions (2) de façon que les permutations permises maintiennent l'appartenance des éléments dans chaque cellule A_c.

Maintenant, contraignons les permutations τ dans l'espérance (2), en les désignant par σ pour refléter la nouvelle contrainte, de sorte que $\{\sigma(i): i \in A_c \cap S\} = A_c \cap S$, c.-à-d. que les unités soient reclassées dans chaque bloc d'échantillon $A_c \cap S$, mais non mélangées entre blocs. L'échantillon S peut être indicé de manière que les $n_c \equiv \left|A_c \cap S\right|$ éléments échantillonnés dans la c^{e} cellule A_c soient numérotés consécutivement de $L_{c-1}+1$ à L_c dans l'échantillon dénombré S, où $L_b = \sum_{j=1}^{b} n_j$ pour chaque $b = 1,..., C$. La préservation des cellules sous la contrainte σ signifie alors que

$$\text{pour tout } 1 \le c \le C \text{ et } i \in A_c \cap S,\ L_{c-1} < \sigma(i) \le L_c . \tag{4}$$

Les permutations aléatoires permises σ des éléments de l'échantillon sont tirées équiprobablement des $\Pi_{1 \le c \le C} n_c!$ permutations σ de $\{1,2,...,n\}$ qui satisfont (4). La mesure modifiée est définie comme étant l'espérance sur σ de la somme pondérée cumulative absolue maximale de biais au niveau de la cellule relativement à $\hat{t}_{y^{(k)}}$:

$$m_k^* \equiv E_\sigma \left(\max_{1 \le q \le n} \left| \hat{B}_k \left(\{\sigma(1),..., \sigma(q)\} \right) \right| \right) \Big/ \hat{t}_{y^{(k)}} . \tag{5}$$

Un estimateur Monte Carlo de la quantité modifiée (5) peut de nouveau être implémenté en fonction d'une matrice V_{ri} de variables aléatoires indépendantes suivant une loi uniforme (0, 1), selon la formule

$$\hat{m}_k^* \equiv \left(R\hat{t}_{y^{(k)}}\right)^{-1} \sum_{1 \le r \le R} \max_{1 \le c \le C, q \in A_c} \left| \sum_{\{i \in A_c \cap S: i \le L_{c-1} \text{ or } V_{ri} \le V_{rq}\}} \left(\frac{\rho_i}{\hat{p}_i} - 1 \right) \left(y_i^{(k)} / \pi_i \right) \right| . \tag{6}$$

2.3 Intervalles de confiance et bornes pour m_k et m_k^*

Toutes les grandeurs m_k, m_k^* sont des fonctions des données d'enquête échantillonnées, et les énoncés de probabilité formulés à cette étape concernent uniquement l'élément de chance introduit par les variables aléatoires V_{ri} utilisées pour définir (3) et (6), *conditionnellement à l'échantillon*. À la présente section, nous donnons une série utile de limites de confiance pour les différences entre les quantités m_k, m_k^* et leurs estimations $\hat{m}_k$, $\hat{m}_k^*$, ainsi que pour les différences entre ces quantités et $\left|\delta^{(k)}\right|$.

Premièrement, l'erreur-type d'échantillonnage de $\hat{m}_k$ due aux variables aléatoires V_{ri} dans (3) est

$$se\left(\hat{m}_k\right)=\left(\hat{t}_{y^{(k)}}\right)^{-1}\left[\frac{1}{R(R-1)}\sum_{1\le r\le R}\left(\max_{0\le x\le 1}\left|\sum_{i\in S}I_{[V_{ri}\le x]}\left(\frac{\rho_i}{\hat{p}_i}-1\right)y_i^{(k)}/\pi_i\right|-\hat{t}_{y^{(k)}}\hat{m}_k\right)^2\right]^{1/2}$$

et un théorème de la limite centrale s'applique si *R* est grand, de sorte qu'avec environ 99 % de confiance

$$\left|m_k-\hat{m}_k\right|\le 2.576\cdot se\left(\hat{m}_k\right). \tag{7}$$

Des énoncés de confiance empirique semblables peuvent être donnés pour borner $m_k^*-\hat{m}_k^*$.

Ensuite, il est montré dans Slud et Bailey (2009) qu'avec $\delta^{(k)}=\hat{B}_k(U)\big/\hat{t}_{y^{(k)}}$ et

$$\gamma^{(k)}=n^{-1}\sum_{i\in S}\left((\rho_i/\hat{p}_i)-1\right)^2\left(y_i^{(k)}/\pi_i\right),\ b_k=1.2286\left(n\gamma^{(k)}\right)^{1/2}\big/t_{y^{(k)}} \tag{8}$$

définis d'après l'échantillon, pour *n* grand

$$0\le m_k-\left|\delta^{(k)}\right|\le\left(\hat{t}_{y^{(k)}}\right)^{-1}E_{\mathbf{v}}\left(\max_{0\le x\le 1}\left|\sum_{i\in S}\left(I_{[V_{ri}\le x]}-x\right)\left((\rho_i/\hat{p}_i)-1\right)\left(y_i^{(k)}/\pi_i\right)\right|\right)\approx b_k. \tag{9}$$

L'assertion (9) signifie que, pour tout $\varepsilon>0$, $m_k-\left|\delta^{(k)}\right|$ est bornée supérieurement, avec une probabilité élevée quand *n* est grand, par $(1+\varepsilon)\cdot b_k$. Un argument similaire prouve que $m_k^*-\left|\delta^{(k)}\right|$ est également positive et bornée supérieurement par $(1+\varepsilon)\cdot b_k$ pour la même quantité b_k définie en (8).

Pour les items particuliers *k*, nous constatons généralement que, quand *n* est grand, pour des nombres modérés *R* (disons *R*≥100), la différence $\hat{m}_k-m_k$ (ou $\hat{m}_k^*-m_k^*$) est très faible comparativement à m_k (respectivement m_k^*). Alors, $\hat{m}_k$ doit être approximativement égale à b_k, ou inférieure, pour que les données d'échantillon soient compatibles avec un biais relatif $\delta^{(k)}$ nul. L'objectif de la présente analyse est de marquer comme étant « inadéquatement ajustées » les variables pour lesquelles l'ajustement de la non-réponse due à l'attrition fondé sur un modèle produit des valeurs de mesure $\hat{m}_k$ plus grandes que $2.576\cdot se\left(\hat{m}_k\right)+b_k$. La principale conséquence pratique est que, si l'une des trois quantités m_k, m_k^*, $\left|\delta^{(k)}\right|$ est beaucoup plus grande que la borne en (9), elles le sont alors toutes les trois, ce qui indique que la variable *k* en question a été mal ajustée (pour la vague et le modèle pris en considération). Quand les quantités m_k, m_k^*, $\left|\delta^{(k)}\right|$ sont de même taille ou plus petites que la borne b_k, la quantité m_k^* mesure la qualité de l'ajustement en tenant compte des divers sous-domaines de population, y compris toutes les cellules spécifiées A_c.

3. Valeurs des mesures d'ajustement pour la SIPP de 1996

Dans l'exemple particulier de l'enquête par panel SIPP de 1996 réalisée aux États-Unis, nous comparons les valeurs des mesures estimées individuelles $\hat{m}_k$, $\hat{m}_k^*$ ou leur moyenne sur 11 variables étudiées transversales, entre divers modèles d'ajustement afin de choisir le « meilleur » modèle. Les probabilités de réponse $\hat{p}_i$ sont estimées au moyen de modèles à cellules d'ajustement et de modèles de régression logistique décrits en détail par Slud et Bailey (2006, 2009). Les variables étudiées sont les indicateurs que le ménage de la personne reçoit des coupons alimentaires (`Foodst`) ou l'aide aux familles ayant des enfants à charge (`AFDC`), ou des indicateurs que la personne reçoit Medicaid (`Mdcd`) ou la sécurité sociale (`SocSec`), et des indicateurs que la personne est couverte par une assurance-maladie (`Heins`), vit dans la pauvreté (`Pov`), a un emploi (`Emp`), est au chômage (`UnEmp`), est inactive (`NILF`), est mariée (`MAR`) ou est divorcée (`DIV`).

Dans cet exemple de données, la non-réponse est corrigée de l'une de deux façons, soit en utilisant un ajustement par ratio des données de la SIPP basées sur 149 cellules standard définies en fonction de variables incluant l'âge, le sexe, la race, l'origine hispanique, le niveau de scolarité et la situation d'activité, ou en utilisant l'un d'une série de modèles de régression logistique B–III résumés au tableau 3-1. Les variables utilisées dans ces modèles de régression sont la race, l'origine hispanique, la situation de locataire contre propriétaire du logement, l'indicateur que la personne est la personne de référence du ménage, l'indicateur d'études collégiales, une variable à quatre catégories de type de famille utilisée pour définir les cellules de ratissage utilisées dans la SIPP, la région du recensement des États-Unis, un indicateur de propriété des actifs, ainsi que certaines des 11 variables étudiées susmentionnées de la SIPP, ou toutes. Voir Slud et Bailey (2009) pour des renseignements supplémentaires sur l'ajustement des modèles. Aux fins de comparaison, le « modèle » d'ajustement par le ratio (SIPP) avec 149 cellules donne une déviance de 76117 pour 149 degrés de liberté.

Nous évaluons maintenant les modèles en repérant les variables dont les valeurs $\hat{m}_k$, $\hat{m}_k^*$ sont grandes comparativement à b_k plus le deuxième membre de (7). Les mesures ont été calculées pour $R = 100$ répliques Monte Carlo (amplement suffisant parce que n=94444 est si grand), et les résultats pour les modèles B, D et F sont présentés au tableau 3-2. Les dernières colonnes du tableau 3-2 donnent, pour le modèle B, les bornes b_k (qui s'avèrent presque identiques, dans une vague, pour les diverses méthodes d'ajustement) respectivement pour la non-réponse aux vagues 4 et 12. Pour toutes les variables, méthodes d'ajustement et vagues, les bornes du deuxième membre de (7) sont au plus égales à 5 % des bornes correspondantes b_k.

Pour la vague 12 dans le tableau 3-2, nous avons toujours $\hat{m}_k > b_k$ sous le modèle à cellules d'ajustement et sous tous les modèles de régression logistique, à l'exception de quelques k seulement. Une exception notable est `Pov`, où, comme dans Slud et Bailey (2006), le modèle B qui comprend `Pov` comme prédicteur s'ajuste bien à la vague 4 ainsi qu'à la vague 12. De même, le modèle D qui inclut `Pov`, `Mdcd`, `Heins` et `UnEmp`, ajuste bien les totaux de ces variables selon la mesure donnée par $\hat{m}_k$. En effet, la constatation la plus frappante lorsqu'on examine les tableaux 3-2 et 3-4 est qu'inclure une variable comme prédicteur dans ces modèles de régression logistique produit généralement un très bon ajustement, qu'il soit mesuré au moyen de $\hat{m}_k$ ou $\hat{m}_k^*$. Cela est vrai même pour le modèle F comparativement au modèle D, bien que le tableau 3-1 indique que le modèle F possède effectivement une plus grande déviance que le modèle D.

Tableau 3-1
Modèles de régression logistique pour la réponse à la vague 4 ou 12 dans la SIPP 96. DDL = nombre de coefficients dans chaque modèle et Dév. = déviance pour les données d'échantillon de 94 444 enregistrements à la vague 4. AIC = Dév. + 2 DDL.

Modèle	DDL	Variables	Dév.	AIC
B	9	Wnotsp Renter College Pov Black RefPer Renter*College Black*College	76545	76563
C	14	même que B, plus Foodst Mdcd Heins UnEmp Div	76299	76327
D	14	même que B, moins Black*College plus Mdcd Heins plus UnEmp Pov*Heins Mdcd*Heins Heins*College	76242	76270
F	18	même que C, plus AFDC SocSec Emp Mar	76280	76316
I	20	College Assets hisp Region Famtyp Renter Refper Pov Mdcd Heins UnEmp Foodst SocSec Mar Renter*College	65678	65718
III	31	même que I, plus AFDC NILF Div Wnotsp Black plus 6 interactions par paire	65638	65700

Tableau 3-2
Quantités (équation) dans (3) estimées d'après les données de la SIPP96, pour l'ajustement de la non-réponse à la vague 4 ou 12, par la méthode des cellules d'ajustement (A) ou de régression logistique (modèles B, D, F) basées sur R = 100 répliques. Les deux dernières colonnes donnent les bornes dans (9), pour le modèle B.

Item	$\hat{m}^{4,A}$	$\hat{m}^{4,B}$	$\hat{m}^{4,D}$	$\hat{m}^{4,F}$	$\hat{m}^{12,A}$	$\hat{m}^{12,B}$	$\hat{m}^{12,D}$	$\hat{m}^{12,F}$	$b_{4,k}$	$b_{12,k}$
Foodst	,0052	,0186	,0076	,0039	,0442	,0130	,0110	,0093	,0056	,0123
AFDC	,0067	,0248	,0067	,0053	,1040	,0350	,0624	,0134	,0078	,0173
Mdcd	,0066	,0279	,0035	,0037	,0163	,0426	,0078	,0084	,0053	,0119
SocSec	,0191	,0116	,0125	,0027	,1118	,1068	,1066	,0073	,0041	,0086
Heins	,0085	,0065	,0013	,0012	,0197	,0133	,0027	,0028	,0019	,0040
Pov	,0187	,0033	,0032	,0032	,0372	,0091	,0074	,0085	,0047	,0097
Emp	,0016	,0017	,0015	,0014	,0082	,0122	,0161	,0034	,0020	,0041
UnEmp	,0534	,0594	,0095	,0098	,1176	,1280	,0207	,0184	,0131	,0250
NILF	,0032	,0034	,0029	,0023	,0333	,0462	,0456	,0063	,0033	,0069
MAR	,0111	,0018	,0018	,0017	,0508	,0226	,0213	,0037	,0025	,0051
DIV	,0124	,0201	,0168	,0048	,0235	,0390	,0334	,00098	,0067	,0133

Le tableau 3-3 décrit la progression de la mesure $\hat{m}_k$ moyenne (sur k = 1, ..., 11 et le chiffre de population) quand le modèle d'ajustement varie du modèle à cellules d'ajustement aux divers modèles logistiques du tableau 3-1. Les cellules spécifiées A_c utilisées dans nos calculs représentaient un système de 101 cellules définies selon le sexe, les groupes d'âge et la race, sur lesquelles est effectué, pour la SIPP, l'ajustement des poids sur les totaux de population par ratissage. Les modèles de régression logistique présentés au 3-3 sont tous meilleurs que le modèle à cellules pour l'ajustement à la vague 4 ainsi qu'à la vague 12. Puisque les modèles (sauf F) sont énumérés en ordre décroissant d'AIC, le classement des modèles selon l'AIC diffère clairement de celui obtenu selon la mesure $\hat{m}_k$ moyenne. Le modèle F semble être le mieux ajusté aux deux vagues dans le tableau 3-3, quoique les modèles I et III soient de solides concurrents et seraient privilégiés si l'on se fondait sur l'examen des déviances. La mesure $\hat{m}$ récompense le modèle F d'inclure un grand nombre de variables de la SIPP comme prédicteurs. Dans l'ensemble, ni la mesure m ni la mesure m^* ne récompense les modèles I et III à déviance nettement plus faible pour leur exactitude prédictive.

Tableau 3-3
Moyennes sur $k = 1,...,12$, où la variable 12 est le chiffre de population, des valeurs de la mesure $\hat{m}_k$ calculées au moyen des données de la SIPP de 1996 pour les modèles à cellules d'ajustement et de régression logistique

Modèle	Cell. adj.	B	C	D	F	I	III
$\hat{m}^4$	,0123	,0150	,0043	,0057	,0034	,0039	,0039
$\hat{m}^{12}$	,0474	,0389	,0248	,0281	,0078	,0137	,0083

Tableau 3-4

Valeurs estimées de la mesure $\hat{m}_k^*$ définie en (6) pour l'ajustement de la vague 4 (vague 12 uniquement à la dernière ligne) avec R =100, fondées sur les modèles à cellules d'ajustement et de régression logistique, en utilisant les données de la SIPP de 1996 avec la partition A_c en 101 cellules démographiques. Aux deux dernières lignes, la moyenne des valeurs de la mesure $\hat{m}_k^*$ est calculée sur les k variables.

Item	ModC	ModD	ModF	ModI	ModIII	Cell. adj.
AFDC	,0065	,0075	,0064	,0063	,0059	,0072
SocSec	,0115	,0127	,0038	,0048	,0045	,0192
Heins	,0021	,0022	,0020	,0027	,0026	,0086
Pov	,0050	,0052	,0051	,0049	,0049	,0191
UnEmp	,0097	,0092	,0098	,0103	,0102	,0535
MAR	,0025	,0025	,0025	,0038	,0035	,0112
DIV	,0057	,0170	,0056	,0055	,0054	,0129
Moy. vague 4	,0051	,0065	,0044	,0046	,0045	,0127
Moy. vague 12	,0273	,0304	,0110	,0163	,0113	,0492

L'examen minutieux, dans le tableau 3-4, des variables étudiées individuelles, dont plusieurs ne sont pas présentées, montre que les modèles d'ajustement paramétriques F, I et III accomplissent quelque chose que le modèle à cellules d'ajustement ne peut pas faire : ils génèrent des probabilités de réponse ayant un bon comportement sur les cellules de ratissage considérées comme des sous-domaines. Les modèles de régression logistique, surtout F et III, donnent de meilleurs résultats, variable par variable, avec un nettement plus petit nombre de paramètres que les 149 fractions de réponse des cellules d'ajustement selon lesquelles le modèle à cellules d'ajustement réparti les poids de sondage dans les cellules d'ajustement respectives. Le modèle F semble être le meilleur choix dans l'ensemble, suivi de près par le modèle III.

4. Modèles et mesures en utilisant les poids de la SIPP ajustés par ratissage

En pratique, dans le cas d'enquêtes gouvernementales telle que la SIPP, les poids sont ajustés, puis calés par ratissage de sorte que les totaux de population de certaines cellules démographiques concordent avec les totaux observés au moyen d'autres recensements et enquêtes plus précis. Pour cette raison, nous avons recalculé les mesures pour le modèle à cellules d'ajustement ainsi que les modèles présentés au tableau 4-1 en nous servant de poids ajustés qui ont été soumis à une étape finale de ratissage comme il est discuté dans Slud et Bailey (2009), en nous fondant sur des cellules définies en fonction du sexe, de l'âge, de la race, de la structure familiale et de l'origine hispanique. Pour les deux mesures, dont la moyenne est calculée sur k variables étudiées (11 variables plus le chiffre de population), nous comparons les ajustements fondés sur un modèle avec et sans ratissage au tableau 5-1. D'après les mesures m_k et m_k^*, avec les meilleurs modèles (F, I, III), cela n'a pour ainsi dire pas d'importance que l'on procède au ratissage ou non, mais ce dernier n'offre aucun avantage réel et peut même poser une certaine perte d'exactitude de l'ajustement, surtout pour les modèles F et III.

Tableau 4-1

Moyennes sur k variables de l'enquête des scores de mesure estimés $\hat{m}_k$ et $\hat{m}_k^*$ d'après les données de la SIPP de 1996, pour les poids ajustés non calés et calés par ratissage provenant des modèles de régression logistique (D, F, I, III) et du modèle à cellules d'ajustement.

Mesure	Ratissage	Vague	ModD	ModF	ModI	ModIII	Cell. adj.
m_k	Non	4	,0057	,0034	,0039	,0039	,0123
		12	,0281	,0078	,0137	,0088	,0474
	Oui	4	,0047	,0057	,0046	,0045	,0083
		12	,0163	,0120	,0166	,0113	,0181
m_k^*	Non	4	,0065	,0044	,0046	,0045	,0127
		12	,0304	,0110	,0163	,0113	,0493
	Oui	4	,0046	,0056	,0046	,0045	,0082
		12	,0164	,0121	,0165	,0113	,0182

5. Conclusion

Le présent article décrit l'élaboration de mesures de l'efficacité de l'ajustement pour la non-réponse calculé en réindexant aléatoirement l'échantillon de l'enquête et en calculant les écarts maximaux sur des sous-domaines indicés consécutifs. L'un des objectifs était de ne pas tenir compte de tout avantage qu'un modèle d'ajustement pourrait avoir en ce qui concerne l'élimination du biais de non-réponse dans l'ensemble de l'échantillon en incluant des variables de l'enquête comme prédicteurs. Cependant, appliquées aux données de la SIPP de 1996, les mesures développées avaient tendance à favoriser les modèles surajustés aux données et ceux qui comprenaient plusieurs variables de l'enquête comme prédicteurs. Alors qu'une telle stratégie d'ajustement ne pourrait pas être mise à l'essai si l'ensemble choisi de variables étudiées intéressantes est trop grand, l'intégration des variables de l'enquête comme prédicteurs de la non-réponse semble être une bonne idée dans les conditions de la SIPP avec 11 variables.

Dans le panel de la SIPP de 1996, les meilleurs modèles d'ajustement pour la non-réponse donnaient des résultats pratiquement aussi bons que les poids ajustés, qu'ils aient été calés par ratissage ou non et donc, du point de vue de nos mesures, un ratissage extensif pourrait ne pas être très utile quand un modèle d'ajustement de la non-réponse hautement efficace est utilisé.

Bibliographie

Dufour, J., Gagnon, F., Morin, Y., Renaud, M. et Särndal, C.-E. (2001), Mieux comprendre la transformation des poids à l'aide d'une mesure de changement, *Techniques d'enquête*, 27, pp. 105-116.

Eltinge, J. et Yansaneh, I. (1997). Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse, avec application à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey, *Techniques d'enquête*, 23, pp. 37-45.

Kim, J.-K. et Kim, J. (2007). Nonresponse weighting adjustment using estimated response probability, *Canadian Jour. of Statist.*, 35, pp. 501-514.

Oh, H. et Scheuren, F. (1983). Weighting adjustment for unit nonresponse, pp. 143-184 in *Incomplete Data in Sample Surveys*, vol. 2, eds. W. Madow, I. Olkin and D. Rubin. New York: Acad. Press.

Rizzo, L., Kalton, G., Brick, M. et Petroni, R. (1994). *Adjusting for panel nonresponse in the Survey of Income and Program Participation, ASA Surv. Res. Meth. Proc. paper, JSM 1994*.

Särndal, C.-E., Swensson, B., et Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer: New York.

Slud, E. et Bailey, L. (2006). Estimation of attrition biases in SIPP, *ASA Surv. Res. Meth. Proceedings paper, JSM 2006*, Seattle, WA.

Slud, E. et Bailey, L. (2009). Evaluation and selection of models for attrition nonresponse adjustment, *Techniques d'enquête*, à paraître.

Recueil du Symposium 2009 de Statistique Canada
Les enquêtes longitudinales : de la conception à l'analyse

L'ajustement des poids de non-réponse au moyen de l'imputation multiple dans la Millenium Cohort Study du Royaume-Uni

John W. McDonald et Sosthenes C. Ketende[1]

Résumé

Le présent article porte sur l'ajustement des poids de non-réponse pour la vague 3 de la Millennium Cohort Study (MCS) du Royaume-Uni. Quand le profil de non-réponse est monotone, on peut ajuster les poids de non-réponse, qui correspondent à l'inverse de la probabilité estimée de réponse fondée sur un modèle de régression logistique dans lequel sont utilisées des données provenant des vagues antérieures pour prédire la réponse à la vague courante. Quand le profil de non-réponse n'est pas monotone, certains cas présentent des données manquantes pour les vagues antérieures et cette méthode ne peut pas être appliquée facilement. Dans le cadre de la MCS, 7,5 % des familles ont participé aux vagues 1 et 3, mais non à la vague 2; il existe donc un profil non monotone de non-réponse pour 1 444 familles.

En vue d'estimer un poids de non-réponse pour la vague 3 de la MCS, nous avons recouru à l'imputation multiple en vue d'imputer pour ces 1 444 familles les valeurs manquantes à la vague 2 nécessaires pour procéder à la modélisation logistique de la réponse à la vague 3. Cette imputation reposait sur les données des vagues 1 et 3, et concernait uniquement les valeurs manquantes des variables temporellement variantes que l'on savait être des prédicteurs de la non-réponse dans la MCS. Cela s'est traduit par l'imputation multiple à la vague 3 de poids de non-réponse dont la moyenne peut être calculée pour produire un poids de non-réponse unique ou qui peuvent être utilisés individuellement pour effectuer des analyses distinctes et ensuite combiner les résultats en utilisant les règles de Rubin. Nous discutons des avantages et des inconvénients des deux approches.

Mots-clés : imputation, enquête longitudinale, non-réponse à une vague, ajustement des poids.

1. Introduction

1.1 Non-réponse à une vague dans les enquêtes longitudinales

Un problème rencontré fréquemment dans les enquêtes longitudinales, telles que les enquêtes par panel ou les études de cohortes de naissances, est la non-réponse à une vague, qui se produit lorsque des réponses sont obtenues pour certaines vagues de l'étude mais pas toutes. Little et David (1983) discutent des trois types de non-réponse à une vague, à savoir l'érosion de l'échantillon, la réentrée dans l'échantillon et l'entrée tardive dans l'échantillon. L'érosion a lieu quand une unité abandonne l'étude durant une vague et continue de ne plus participer par après; la réentrée a lieu quand une unité abandonne l'étude pendant une vague ou plus, mais recommence à participer à l'étude par après; l'entrée tardive a lieu quand une unité ne participe pas à la première vague, mais entre dans l'étude plus tard. À la vague 3 de la Millennium Cohort Study (MCS) du Royaume Uni, le profil longitudinal de réponse, qui comprend une combinaison de profils d'érosion, de réentrée et d'entrée tardive, est complexe. La situation à la vague 3 est la suivante : 2 210 familles ont cessé de participer à la vague 1 et 1 664, à la vague 2; 692 familles sont entrées tardivement dans la MCS à la vague 2 et de ces « nouvelles familles », 124 ont cessé de participer à la vague 2; enfin, 1 444 familles sont réentrées dans la MCS en participant à la vague 1, mais non à la vague 2, puis en réentrant à la vague 3.

Le profil de participation longitudinal peut être monotone, si aucun participant ne revient dans l'étude après avoir manqué une vague, on non monotone, quand certains sujets reviennent dans l'étude après avoir manqué une vague.

[1]John W. McDonald, Centre for Longitudinal Studies, Institute of Education, University of London, 20 Bedford Way, London, United Kingdom, WC1H 0AL (John.McDonald@ioe.ac.uk); Sosthenes C. Ketende, Centre for Longitudinal Studies, Institute of Education, University of London, 20 Bedford Way, London, United Kingdom, WC1H 0AL (S.Ketende@ioe.ac.uk)

Dans le présent article, nous discutons de l'ajustement des poids pour la non-réponse totale à la vague 3 de la MCS. Quand le profil de non-réponse à une vague est monotone, on peut ajuster les poids de non-réponse, qui correspondent à l'inverse de la probabilité estimée de réponse à la vague courante fondée sur un modèle de régression logistique dans lequel sont utilisées des données provenant des vagues antérieures pour prédire la réponse à la vague courante. Quand le profil n'est pas monotone, certaines unités présentent des données manquantes pour les vagues antérieures et cette méthode n'est pas facile à appliquer. Dans le cas de la MCS, le profil longitudinal de réponse est complexe et non monotone, des familles étant entrées tardivement à la vague 2 et d'autres étant réentrées à la vague 3 après ne pas avoir participé à la vague 2. Notre approche en vue d'estimer les poids de non-réponse pour la vague 3 de la MCS consistait à recourir à l'imputation multiple pour imputer à la vague 2 les valeurs manquantes requises pour la modélisation logistique de la réponse à la vague 3. Cette imputation reposait sur les données des cycles 1 et 3, et concernait uniquement les valeurs manquantes pour les variables temporellement variantes que nous savions être des prédicteurs de la non-réponse dans le cadre de la MCS. Cela a donné lieu à l'imputation multiple de poids de non-réponse dont la moyenne peut être calculée pour produire un poids de non-réponse unique ou qui peuvent être utilisés individuellement tous les 10 pour effectuer des analyses distinctes, puis combiner les résultats en utilisant les règles de Rubin. Dans le présent article, nous discutons des avantages et des inconvénients des deux approches.

1.2 Correction pour la non-réponse à une vague dans les enquêtes par panel

Kalton (1986) et Lepkowski (1989) passent en revue les méthodes de correction de la non-réponse à une vague dans les enquêtes par panel. Tous deux discutent de l'utilisation de la pondération ou de l'imputation, ainsi que des raisons pour lesquelles, en règle générale, la non-réponse totale est traitée par pondération et la non-réponse partielle, par imputation. Le choix de la stratégie utilisée dépend de l'information auxiliaire disponible pour effectuer les ajustements pour tenir compte de la non-réponse. La pondération est souvent privilégiée quand les variables auxiliaires disponibles ne sont que faiblement corrélées aux variables pour lesquelles des valeurs manquent, par exemple quand on ne possède que des renseignements sur le plan de sondage. L'imputation, quant à elle, est souvent privilégiée quand les variables auxiliaires disponibles sont fortement corrélées aux variables pour lesquelles des valeurs manquent. Dans le cas des enquêtes longitudinales, deux approches générales d'imputation peuvent être adoptées, à savoir les imputations transversales et les imputations entre vagues. Les imputations transversales sont fondées sur des données provenant d'une seule vague, tandis que les imputations entre vagues sont fondées sur des données provenant de plusieurs vagues. Par exemple, pour un panel à deux vagues avec variables répétées, on utiliserait la valeur d'une variable dans l'une des vagues pour imputer la valeur manquante de la même variable dans l'autre vague; si la valeur manque à la deuxième vague, il s'agit d'une projection et si elle manque à la première vague, il s'agit d'une rétropolation. Pour les variables qui sont répétées à chaque vague, l'imputation pour corriger la non-réponse à une vague est sans doute la meilleure solution si les réponses pour la variable répétée sont temporellement très corrélées. En général, les valeurs manquantes peuvent être extrapolées (prédites) en utilisant les valeurs de la variable aux vagues précédentes ou rétropolées en utilisant les valeurs de la variable aux vagues subséquentes, ou encore « interpolées » si la vague manquante est entourée de vagues adjacentes non manquantes. Naturellement, le scénario d'imputation entre vagues comprendrait généralement aussi l'utilisation d'autres variables, ainsi que les variables répétées d'une vague à l'autre dans le modèle d'imputation.

Comme la pondération et l'imputation ont chacune leurs points forts (avantages) et points faibles (inconvénients), tant Kalton (1986) que Lepkowski (1989) considère que la meilleure solution pourrait consister à combiner des méthodes de pondération et d'imputation. Cette stratégie mixte est celle que nous avons adoptée ici pour estimer l'ajustement des poids de non-réponse à une vague. Comme le modèle de régression logistique que nous souhaitions utiliser pour calculer la réponse à la vague 3 comprenait des variables prédictrices pour lesquelles des données manquaient, nous avons utilisé l'imputation multiple entre vagues pour traiter les cas de données partiellement manquantes pour certaines variables prédictives. Cela s'est traduit par l'imputation multiple de poids de non-réponse à la vague 3.

1.3 Ajustement des poids de non-réponse pour la Millennium Cohort Study du Royaume-Uni

La population cible de la Millennium Cohort Study du Royaume-Uni comprend les enfants nés entre le 1er septembre 2000 et le 31 août 2001 (en Angleterre et au Pays de Gales), ainsi qu'entre le 24 novembre 2000 et le 11 janvier 2002 (en Écosse et en Irlande du Nord), en vie et habitant au Royaume-Uni à l'âge de neuf mois, et admissibles à la station pour enfants à cet âge-là, puis, après neuf mois, pendant aussi longtemps qu'ils

demeurent au Royaume-Uni au moment de l'échantillonnage. Les études antérieures de cohorte de naissances en Grande-Bretagne, au Royaume-Uni lancées en 1946, en 1958 et en 1970 comprenaient l'échantillonnage de toutes les naissances en une semaine. Par contre, les naissances échantillonnées pour la MCS s'étalent sur une année civile et ont été sélectionnées selon un plan d'échantillonnage complexe. L'échantillon de la MCS est un échantillon en grappes à un degré stratifié de circonscriptions électorales. La population du Royaume-Uni a été stratifiée par pays, à savoir l'Angleterre, le pays de Gales, l'Écosse et l'Irlande du Nord. En se servant d'une mesure de la pauvreté et de l'ethnicité des enfants au niveau du quartier, en Angleterre, les quartiers ont été répartis en trois strates, à savoir les quartiers favorisés, défavorisés et ethniques, et au pays de Gales, en Écosse et en Irlande du Nord, ils ont été répartis en deux strates, à savoir les quartiers favorisés et défavorisés. Des fractions d'échantillonnage différentes ont été appliquées à chacune des neuf strates, de manière que les petits pays du Royaume-Uni soient surreprésentés comparativement à l'Angleterre, que les familles vivant dans les régions plus défavorisées soient surreprésentées en Angleterre et que les familles vivant dans les régions où la proportion de familles appartenant à une minorité ethnique était élevée soient également surreprésentées. Après avoir sélectionné les 198 quartiers, il a fallu dresser la liste de tous les enfants de neuf mois qui y vivaient. Les listes ont été produites en se servant des dossiers de prestation pour enfant versée par l'administration publique. La prestation pour enfant est une prestation universelle, payable à partir de la date de la naissance de l'enfant, et les demandes de prestation pour enfant couvrent pratiquement toute la population du Royaume-Uni, sauf les personnes qui n'y ont pas le « droit de résidence ». Toutes les familles ayant droit à la prestation pour enfant et celles dont l'enfant vivait à l'âge de neuf mois dans le quartier échantillonné ont reçu du ministère responsable de l'administration des dossiers de prestation pour enfant une lettre les invitant à participer à la MCS. Le taux de réponse à la vague 1 de la MCS était de 72 %. Pour d'autres renseignements sur le plan d'échantillonnage et la mise en œuvre de la MCS, consulter Plewis (2007a).

Certaines familles ont déménagé hors des quartiers sélectionnés dans l'échantillon et d'autres y ont emménagé. Les familles peuvent continuer de recevoir la prestation pour enfant sans avertir l'administration publique d'un changement d'adresse, en particulier si la prestation est versée directement dans un compte bancaire. Par conséquent, la liste des enfants résidant dans les quartiers sélectionnés n'était plus à jour et certains enfants admissibles ont été omis car leur existence dans le quartier sélectionné n'était pas connue durant le travail sur le terrain de la vague 1. L'échantillon de la vague 2 a été complété par 692 « nouvelles familles » qui étaient admissibles à la vague 1, mais qui avaient été omises parce que leur adresse n'était plus à jour. Dans les nouvelles familles, l'enfant faisant partie de la cohorte est entré tardivement dans la MCS, à l'âge de trois ans.

L'échantillon productif à la vague 1, c'est-à-dire celui pour lequel on disposait de données d'interview fournies par au moins le répondant principal ou son ou sa partenaire, formait l'échantillon utilisé sur le terrain à la vague 2, exception faite des membres de la cohorte que l'on savait être décédés ou avoir émigrés. Les cas de refus à la vague 1 n'ont pas été envoyés à l'organisme chargé du travail sur le terrain pour la vague 2. Le taux de réponse à la vague 2 de la MCS était de 79 %. Les corrélats de la non-réponse totale aux deux premières vagues de la MCS sont décrits dans Plewis (2007b) et dans Plewis et coll., (2008). Plewis (2007b) a constaté que, pour les vagues 1 et 2 de la MCS, les écarts entre les probabilités de réponse étaient faibles comparativement à l'inégalité des probabilités de sélection prévues dans le plan de sondage.

Plewis (2007b) décrit la méthode utilisée pour produire les poids de non-réponse à la vague 2 de la MCS. Un ensemble de prédicteurs possibles de la non-réponse à la vague 2 ont été étudiés en utilisant un modèle de régression logistique. Ces prédicteurs englobaient une gamme de variables explicatives sociodémographiques et socioéconomiques, ainsi qu'une mesure de la mobilité résidentielle dérivée de la base de données sur les adresses, qui fait partie du volet administratif des opérations d'enquête de la MCS; voir Plewis et coll., (2008) pour des renseignements plus détaillés sur cette mesure. Les nouvelles familles qui sont entrées tardivement dans la MCS ont été exclues de cette modélisation et leur poids de non-réponse à la vague 2 a été fixé à 1. Toutes les variables explicatives (autre que la mobilité résidentielle) ont été mesurées à la vague 1. À cause de la stratégie de sélection de l'échantillon, c'est-à-dire le fait que les cas de refus à la première vague n'ont pas été sélectionnés à la vague 2, les variables de la vague 1 utilisées pour la modélisation ne présentaient pas de données manquantes. Les probabilités de réponse prédites en se basant sur ce modèle de régression logistique ont été inversées pour obtenir les poids de non-réponse à la vague 2. Les poids globaux utilisés à la vague 2 sont égaux au produit des poids globaux à la vague 1 et des poids de non-réponse à la vague 2.

Pour la vague 3, toutes les familles qui n'ont pas été interviewées à la vague 2 ont été de nouveau sélectionnées pour le travail sur le terrain, excepté les cas de refus permanents et les familles dont le membre de la cohorte était décédé ou avait émigré. De nombreuses familles non interviewées à la vague 2 ont été productives à la vague 3. Par conséquent, le profil de non-réponse longitudinal était non monotone, 1 444 familles ayant produit des données aux vagues 1 et 3, mais non à la vague 2. Comme ces 1 444 cas présentent des données manquantes pour la vague 2, l'estimation des poids de non-réponse par la méthode de régression logistique ne peut pas être appliquée facilement à moins que toutes les variables explicatives du modèle aient été mesurées à la vague 1. Alors que certaines variables explicatives considérées par Plewis (2007b) comme prédictives de la réponse en passant de la vague 1 à la vague 2 étaient temporellement constantes, comme le groupe ethnique ou le membre de la cohorte, d'autres, telles que le revenu familial, le mode d'occupation du logement (propriété, location, autre) et le type de logement (maison/bungalow, autre), variaient en fonction du temps. On s'attendrait à ce que les valeurs les plus récentes, obtenues à la vague 2, soient plus prédictives de la réponse à la vague 2 que les valeurs obtenues à la vague 1, et les valeurs de ces variables explicatives variant en fonction du temps manquent à la vague 2 pour les 1 444 familles productives aux vagues 1 et 3, mais non à la vague 2. Cependant, on pourrait penser qu'il s'agit de la situation idéale pour recourir à l'imputation en vue de remplacer les valeurs manquantes à la vague 2, car nous connaissons les valeurs de ces variables pour les vagues adjacentes.

2. Imputation des données manquantes à la vague 2

L'imputation multiple a été utilisée pour remplacer les données manquantes à la vague 2 par des valeurs « plausibles ». Il convient de souligner que l'imputation ne porte pas sur la vague complète, mais uniquement sur les valeurs manquantes des variables variant temporellement qui, comme l'a montré Plewis (2007b), sont des prédicteurs de la non-réponse dans la MCS. Nous recourons à l'imputation multiple pour tenir compte de l'incertitude dans le processus d'imputation. Nous effectuons l'imputation plusieurs fois, ici dix fois, pour créer dix ensembles de données dans lesquels les valeurs manquantes ont été remplacées. Nous utilisons chacun des dix ensembles de données pour ajuster notre modèle logistique de réponse à la vague 3 en utilisant le même ensemble de variables explicatives. Chaque ensemble de données produit des estimations, habituellement différentes, des poids de non-réponse pour la vague 3 de la MCS. Donc, nous disposons, en bout de ligne, de dix estimations des poids de non-réponse. Diverses questions se posent : 1) Comment les imputations devraient-elles être effectuées? 2) Quelles sont les hypothèses qui sous-tendent la méthode d'imputation et ces hypothèses sont-elles justifiées? 3) Comment l'analyste doit-il traiter les estimations multiples des poids de non-réponse?

2.1 Méthode MICE (Multivariate Imputation by Chained Equations ou *Imputation multivariée par équations enchaînées*)

La méthode MICE (pour *Multivariate imputation by chained equations ou Imputation multivariée par équations enchaînées*) est une méthode d'imputation conçue pour remplacer des données manquantes multivariées en spécifiant des modèles entièrement conditionnels (van Buuren, 2007). L'idée fondamentale est d'imputer les valeurs manquantes des variables, une variable à la fois, en spécifiant un modèle d'imputation par variable. Pour chaque variable manquante, on peut spécifier une loi conditionnelle univariée sachant les autres variables, ce qui signifie que le modèle d'imputation est spécifié séparément pour chaque variable manquante en faisant intervenir les autres variables comme variables prédictives. La méthode consiste à procéder à des itérations sur ces lois conditionnelles en se servant de l'échantillonnage de Gibbs. À chaque étape de l'algorithme, une imputation est produite pour la variable manquante et cette valeur imputée est utilisée dans l'imputation de la variable suivante. Chaque itération passe par toutes les variables pour lesquelles des données manquent. Le processus est répété jusqu'à la convergence. Pour des précisions, consulter van Buuren (2007). La méthode MICE a été implémentée dans un certain nombre de progiciels, dont Stata, R et SPSS (Horton et Kleinman 2007, van Buuren et Groothuis-Oudshoorn, à paraître). Nous avons utilisé l'implémentation de MICE de Patrick Royston en nous servant de ses fichiers Stata ado (Royston, 2004; Royston, 2005a; Royston, 2005b; Royston, 2007).

En théorie, l'imputation multiple marche, mais en pratique, l'utilisation de MICE avec Stata 10 a posé des problèmes, car la plupart de nos variables étaient catégoriques et notre ensemble de données était grand. La présence d'interactions dans le modèle d'imputation nous a obligés à créer explicitement une variable indicatrice pour chaque catégorie de variables catégoriques comptant plus de deux catégories. Nous avons imputé les variables indicatrices lorsque les valeurs manquaient, puis nous avons imputé passivement la variable catégorique originale d'après ses

variables indicatrices imputées. Les problèmes que nous avons dû résoudre en pratique comprenaient les prédicteurs, qui prédisaient parfaitement les variables binaires de sorte que nous avons dû corriger les erreurs itérativement en choisissant une autre catégorie de référence que la première ou en regroupant les catégories. Le traitement des interactions dans les modèles d'imputation demande de la minutie (von Hippel, 2009), dans notre cas principalement à cause des tableaux quasi vides. Faute d'espace, nous ne décrirons pas plus en profondeur les modèles d'imputation utilisés, si ce n'est pour indiquer que nous avons utilisé des imputations entre vagues et essayé de tenir compte du plan d'échantillonnage dans nos modèles d'imputation en introduisant la variable de strate comme l'un des prédicteurs (Reiter et coll., 2006). Nous n'avons pas réussi à tenir compte de la mise en grappes dans nos modèles d'imputation, mais notre modèle logistique de la réponse contenait les variables de strate comme prédicteurs et tenant compte de la mise en grappes de l'échantillon de la MCS. Il convient de souligner que Stata 11 possède une nouvelle instruction mi (multiple imputation) qui résout nombre des problèmes susmentionnés que nous avons dû résoudre en utilisant Stata 10.

Notre modèle d'imputation repose sur l'hypothèse que les valeurs non observées sont de type MAR (pour (*missing at random*, aléatoirement manquantes) ce qui signifie que, sachant les données observées, le mécanisme de création des données manquantes ne dépend pas des données non observées. Dans le contexte d'une étude longitudinale dont le profil de non-réponse est monotone, l'hypothèse que les données d'enquêtes longitudinales sont de type MAR est habituellement émise pour estimer les poids de non-réponse, si bien que la probabilité de réponse à la vague t est supposée être une fonction des variables mesurées aux vagues précédentes, mais non des variables non observées mesurant les changements survenus entre les vagues t-1 et t. Plewis et coll., (2008) montrent que la mobilité résidentielle, c'est-à-dire tout changement de l'adresse postale, après la vague 1 était un prédicteur important de la non-réponse à la vague 2 dans la MCS, ce qui met en doute cette hypothèse fréquente. Dans notre modèle de non-réponse à la vague 3 de la MCS, la mobilité résidentielle entre les vagues 2 et 3 n'était pas incluse en tant que variable prédictrice.

3. Traitement des multiples poids de non-réponse imputés

3.1 Groupement des résultats en appliquant les règles de Rubin

L'imputation multiple donne lieu au remplacement de chaque valeur manquante par un ensemble de M valeurs plausibles pour produire M ensembles de données complets; dans le cas qui nous occupe, M = 10 ensembles de données. Ces 10 ensembles de données ont été utilisés individuellement pour estimer pour la vague 3 de la MCS 10 poids de non-réponse qui ont été utilisés de la façon classique pour produire des poids transversaux ou longitudinaux. Ces poids peuvent servir à produire une estimation pondérée d'une quantité d'intérêt, disons une proportion, une moyenne ou le coefficient de régression d'un modèle de régression logistique. Les résultats peuvent être combinés ou regroupés, en utilisant ce que l'on a dénommé les règles de Rubin (Rubin, 1987) pour obtenir des estimations et des erreurs-types qui tiennent compte de l'incertitude due aux valeurs imputées des données manquantes dans le modèle de non-réponse utilisé pour estimer les poids de non-réponse.

3.2 Options

Une option consiste à utiliser la moyenne des M poids de non-réponse estimés pour l'analyse. Les avantages sont les suivants : 1) il est plus simple d'utiliser le poids moyen avec un seul ensemble de données que de devoir utiliser M ensembles de données, puis de combiner les résultats en appliquant les règles de Rubin et 2) il est plus simples d'archiver un seul ensemble de données que M ensembles de données. Les inconvénients sont les suivants : 1) le poids moyen ne tient pas compte de toute l'incertitude des résultats dus aux données incomplètes, 2) il est plus compliqué de traiter M ensembles de données, tant pour l'utilisateur que pour le fournisseur et l'archiviste des données, et 3) les utilisateurs doivent comprendre l'imputation multiple et avoir la formation nécessaire pour utiliser le logiciel approprié qui permet de traiter M ensembles de données et de regrouper les résultats.

Au lieu d'inverser M probabilités de réponse estimées au moyen du modèle logistique de réponse à la vague 3 et de calculer la moyenne, une autre option consiste à calculer la moyenne des M probabilités de réponse estimées, puis de prendre son inverse. Notons que la moyen harmonique est toujours inférieure ou égale à la moyenne arithmétique, l'égalité n'étant observée que quand tous les nombres sont égaux. Autrement dit, pour chaque cas, le poids moyen est toujours supérieur ou égal à l'inverse de la probabilité moyenne, car ces deux quantités ne sont égales que quand les

M probabilités de réponse estimées sont toutes égales. Ces deux quantités varient selon la variabilité des M probabilités de réponse estimées. Du point de vue statistique, la mesure qui doit être privilégiée n'est pas vraiment évidente.

3.3 Est-il nécessaire d'utiliser de multiples poids de non-réponse imputés?

Est-il nécessaire d'utiliser de multiples poids de non-réponse imputé avec M ensembles de données, puis de regrouper les résultats? Quelles sont les différences si nous utilisons un seul ensemble de données et la moyenne des poids ou l'inverse de la moyenne des probabilités de réponse estimées pour pondérer chaque observation? Nous avons étudié ces questions en estimant une moyenne, une proportion et un coefficient de régression logistique en utilisant les trois options. Dans le cadre de la vague 3 de la MCS, nous avons estimé le poids moyen des garçons en kilogrammes, la proportion de garçons faisant de l'embonpoint ou étant obèses, et le coefficient de régression de la mobilité résidentielle dans un modèle logistique de la réponse à la vague 3. Les figures 3.3-1 à 3.3-3 représentent, de gauche à droite, une estimation groupée en utilisant les dix ensembles de données et la règle de Rubin pour regrouper les résultats, des estimations fondées sur chacun des dix ensembles de données, dénommées IM1, IM2, ..., IM10, une estimation obtenue en utilisant le poids moyen et une estimation produite en se servant de l'inverse de la moyenne des probabilités de réponse estimées. Pour le nombre limité de cas étudiés, l'application des règles de Rubin et l'utilisation du poids moyen ont produit des estimations fort semblables, l'estimation la plus faible étant dans tous les cas celle obtenue en utilisant l'inverse de la moyenne des probabilités de réponse prédites.

Figure 3.3-1
Poids moyen estimé des garçons, en kilogrammes, à la vague 3 de la Millennium Cohort Study

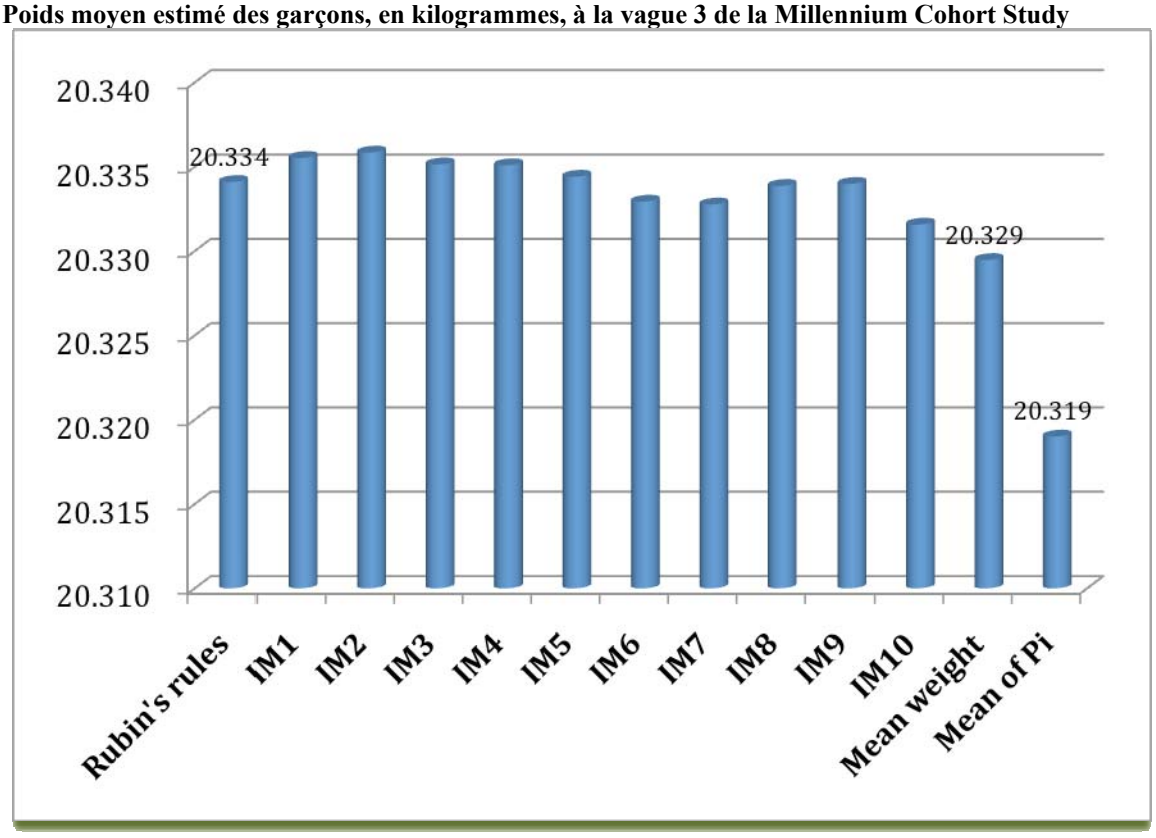


Figure 3.3-2
Proportion estimée de cas d'embonpoint ou d'obésité à la vague 3 de la Millennium Cohort Study

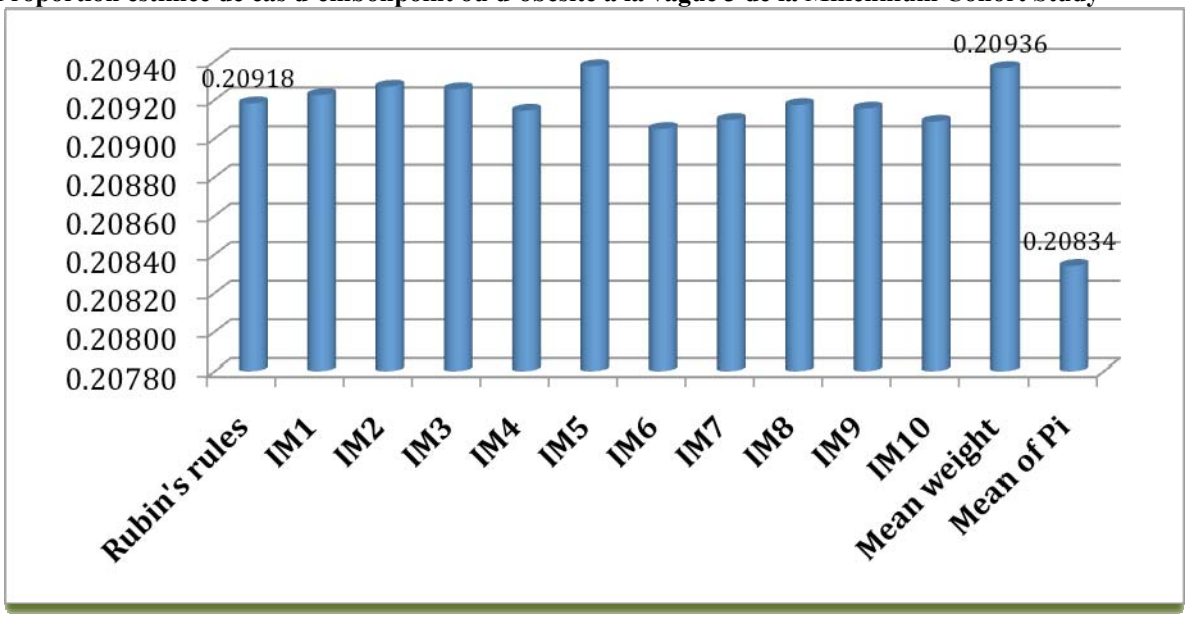
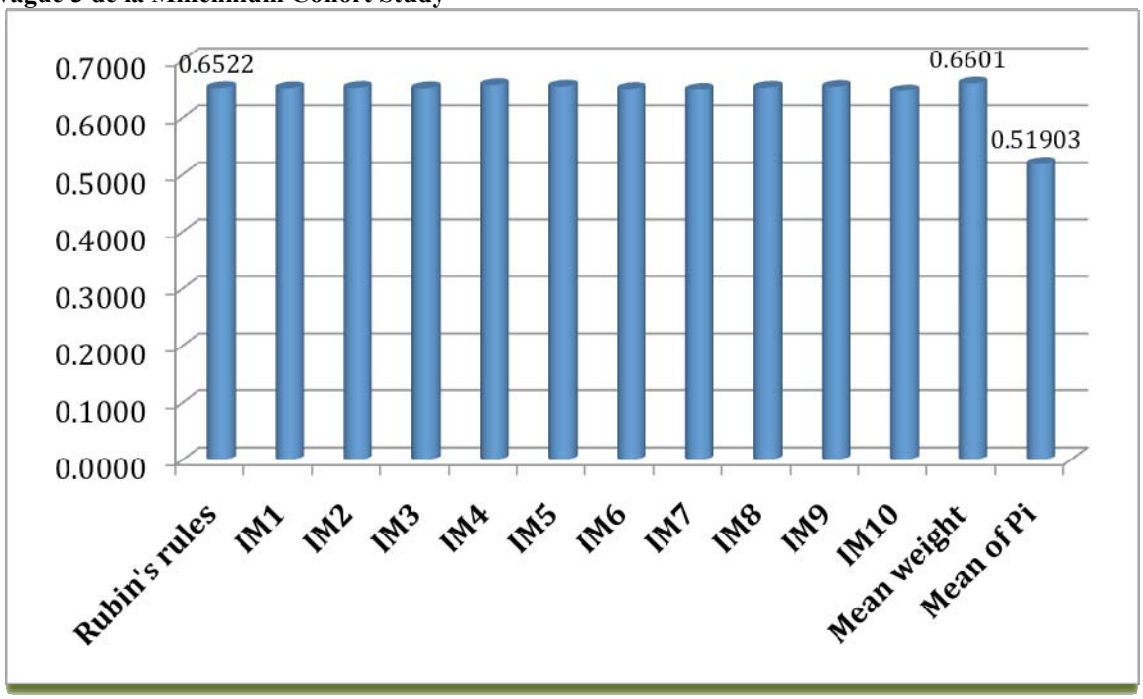


Figure 3.3-3
Coefficient de régression estimé pour la mobilité résidentielle dans le modèle logistique de la réponse à la vague 3 de la Millennium Cohort Study



Remerciements

Les auteurs remercient G. Kalton pour les références bibliographiques utiles et J. N. K. Rao pour ses commentaires.

Bibliographie

Horton, N. J. et Kleinman. K. P. (2007). Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Regression Models, *The American Statistician,* 61, pp. 79-90.

Kalton, G. (1986). Handling Wave Nonresponse in Panel Surveys, *Journal of Official Statistics*, 2, pp. 303-314.

Lepkowski, J. M. (1989). Treatment of Wave Nonresponse in Panel Surveys, in D. Kasprzyk et al. (eds*.) Panel Surveys*, New York: Wiley, pp. 348-374.

Little, R. J. A. et David, M. H. (1983). Weighting Adjustments for Non-response in Panel Surveys, Working paper, U. S. Bureau of the Census, Washington, D. C.

Plewis, I. (ed.) (2007a). Millennium Cohort Study First Survey: Technical Report on Sampling (4th ed.) , London: Centre for Longitudinal Studies.

Plewis, I. (2007b). Nonresponse in a Birth Cohort Study: The Case of the Millennium Cohort Study, *International Journal of Social Research Methodology,* 10, pp. 325-334.

Plewis, I., Ketende, S.C., Joshi, H. et Hughes G. (2008). The Contribution of Residential Mobility to Sample Loss in a Birth Cohort Study: Evidence from the First Two Waves of the UK Millennium Cohort Study", *Journal of Official Statistics* 24, pp. 365-385.

Reiter, J. P., Raghunathan, T. E. et Kinney, S. (2006). L'importance de la modélisation du plan d'échantillonnage dans l'imputation multiple pour les données manquantes, *Techniques d'enquête*, 32, pp. 161-168.

Royston, P. (2004). Multiple Imputation of Missing Values, *Stata Journal*, 4, pp. 227-241.

Royston, P. (2005a) Multiple Imputation of Missing Values: update, *Stata Journal*, 5, pp. 188-201.

Royston, P. (2005b) Multiple Imputation of Missing Values: Update of ice, *Stata Journal*, 5, pp. 527-536.

Royston, P. (2007). Multiple Imputation of Missing Values: Further Update of ice, With an Emphasis on Interval Censoring, *Stata Journal*, 7, pp. 445-464.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

van Buuren S., Boshuizen, H. C. et Knook, D. L. (1999). Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis, *Statistics in Medicine*, 18, pp. 681-694.

van Buuren, S. (2007). Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification, *Statistical Methods in Medical Research*, 16, pp. 219–242.

van Buuren, S. et Groothuis-Oudshoorn K. (à paraître). MICE: Multivariate Imputation by Chained Equations in R, *Journal of Statistical Software*.

von Hippel, P. T. (2009). How to Impute Interactions, Squares, and Other Transformed Variables, *Sociological Methodology*, 39, pp. 265–291.

Recueil du Symposium 2009 de Statistique Canada
Les enquêtes longitudinales : de la conception à l'analyse

Le biais de non-réponse dans l'Enquête longitudinale nationale sur les enfants et les jeunes

Mike Tam et Agnes Waye[1]

Résumé

Dans le présent document, nous examinons la non-réponse dans la cohorte originale de l'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ). L'ELNEJ est une enquête longitudinale qui recueille des données sur les caractéristiques du développement et du bien-être des enfants et des jeunes Canadiens au fil du temps, ainsi que sur les facteurs qui peuvent les influencer. Une cohorte originale d'enfants de 0 à 11 ans a été échantillonnée en 1994, un suivi étant assuré tous les deux ans. Compte tenu de la somme de non-réponse accumulée au fil du temps, une analyse exhaustive a été entreprise, afin de déterminer l'étendue du biais de non-réponse, ainsi que les répercussions de ce biais sur les analyses fondées sur les données de l'ELNEJ. Pour répondre à ces questions, nous tentons de cerner les déterminants possibles de la non-réponse.

Mots-clés : enquête longitudinale, analyse de la non-réponse, régression logistique.

1. Introduction

Le présent document vise à examiner certains des résultats de l'évaluation récente de la cohorte originale de l'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ) en ce qui a trait au biais de non-réponse. En 2007, les membres de la cohorte originale avaient participé à l'enquête pendant sept cycles. Au cours de cette période, les taux de réponse des répondants au cycle 1 ont diminué de façon constante, s'établissant à 64,9 % au cycle 7 (voir le tableau 3.1-1). Une évaluation du biais de non-réponse effectuée dans le cadre de l'examen méthodologique de l'enquête après dix ans, au cycle 6, a démontré la présence d'un biais dans certaines estimations découlant de la cohorte originale (Statistique Canada, 2007). Le présent document vise à fournir des renseignements concernant d'autres variables de l'ELNEJ qui pourraient être touchées par la non-réponse entre les cycles 1 et 7. Un aperçu de l'ELNEJ est fourni dans la section suivante. Il est suivi par un examen des taux de réponse et de l'évolution de l'échantillon de la cohorte originale au fil du temps. Puis, certains des résultats des diverses évaluations que nous avons effectuées sont présentés. Enfin, une brève conclusion faisant état de notre perception actuelle de la situation de non-réponse est fournie.

2. Aperçu de l'ELNEJ

L'ELNEJ est une enquête longitudinale sur les enfants et les jeunes, qui est parrainée par Ressources humaines et Développement des compétences Canada et menée par Statistique Canada. À Statistique Canada, l'enquête est gérée par la Division des enquêtes spéciales, les services méthodologiques étant assurés par la Division des méthodes d'enquêtes auprès des ménages. Parmi les objectifs principaux de l'enquête figurent les suivants : déterminer la prévalence des facteurs de risque, examiner comment ces facteurs touchent le développement et mettre l'information obtenue à la disposition des responsables des politiques. L'ELNEJ est aussi une enquête générale qui peut servir à appuyer d'autres types d'analyses. Les données de l'ELNEJ sont disponibles dans les Centres de données de recherche de Statistique Canada, un peu partout au pays. Selon certaines règles, les chercheurs universitaires et les autres chercheurs peuvent accéder aux données pour leurs recherches.

[1]Mike Tam, Statistique Canada, Immeuble R.H. Coats, 17-L, 100 Promenade Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6 (mike.tam@statcan.gc.ca); Agnes Waye, Statistique Canada, Immeuble R.H. Coats, 17-I, 100 Promenade Tunney's, Ottawa, Ontario, Canada, K1A 0T6 (agnes.waye@statcan.gc.ca)

L'enquête a commencé à l'automne de 1994 auprès d'un échantillon d'enfants de 0 à 11 ans vivant dans une des dix provinces du Canada. Cet échantillon est appelé *cohorte originale* dans l'ELNEJ. Les données de l'enquête ont par la suite été recueillies tous les deux ans, chaque période de collecte biennale correspondant à un *cycle*. À partir du cycle 2, une cohorte du développement de la petite enfance (DPE) a été ajoutée à l'ELNEJ à chaque cycle. Cette cohorte est constituée d'enfants de 0 et 1 an qui doivent être suivis de façon longitudinale pendant au moins trois autres cycles. L'ELNEJ sert aussi à l'analyse transversale. C'est donc dire qu'outre les cohortes longitudinales, un échantillon complémentaire d'enfants de 2 à 5 ans a aussi été tiré à chaque cycle pour améliorer la représentativité de la population transversale cible. À l'été 2009, huit cycles de collecte auprès de la cohorte originale des 0 à 11 ans s'étaient déroulés et les répondants étaient âgés de 14 à 25 ans. Toutefois, dans le présent document, nous mettons l'accent sur la non-réponse de la cohorte originale au cycle 7, qui a été mené à l'été 2007.

Dans l'enquête, même si l'unité d'étude est principalement l'enfant (ou les enfants) du ménage sélectionné, le répondant est l'adulte qui est la *personne la mieux renseignée* (PMR) au sujet de l'enfant concerné, tant que celui-ci n'a pas atteint l'âge de 18 ans. Toutefois, une fois que l'enfant atteint l'âge de 18 ans, il devient un *répondant à part entière*[2]. Dans le reste du présent document, on utilise le terme *répondant* de façon interchangeable dans les deux contextes.

3. Analyse de la non-réponse dans l'ELNEJ

3.1 Taux de réponse

Dans la présente section, nous examinons l'évolution de l'échantillon des membres de la cohorte originale pendant sept cycles. Le tableau 3.1-1 est constitué des taux de réponse des membres de la cohorte originale, des cycles 2 à 7 de l'enquête, par rapport au cycle 1. Au premier cycle de l'ELNEJ, on comptait 22 831 répondants. Toutefois, par suite de réductions de l'échantillon attribuables à des contraintes budgétaires, seulement 16 890 répondants du cycle 1 ont été suivis de façon longitudinale en dernier ressort.

Les taux ont été répartis entre trois groupes d'âge : 0 à 11 ans, 0 à 5 ans et 6 à 11 ans. Cette délimitation vise à permettre une observation distincte des taux de réponse des membres de la cohorte originale, dans les cas où le membre interviewé continue d'être le parent et où les membres sont maintenant des répondants à part entière. De façon plus particulière, les 0 à 5 ans étaient âgés de 12 à 17 ans au cycle 7, ce qui fait que le membre interviewé était le parent, et les 6 à 11 ans étaient âgés de 18 à 23 ans au cycle 7 et étaient par conséquent des répondants à part entière.

Tableau 3.1-1
Taux de réponse non pondérés des cycles 2 à 7 par rapport au cycle 1, selon le groupe d'âge au cycle 1

	0-11		0-5		6-11	
Cycle	Répondants	%	Répondants	%	Répondants	%
1	16 890	100,0	9 553	100,0	7 337	100,0
2	15 384	91,1	8 726	91,3	6 658	90,7
3	14 770	87,4	8 394	87,9	6 376	86,9
4	13 169	78,0	7 544	79,0	5 625	76,7
5	12 273	72,7	7 153	74,9	5 120	69,8
6	11 178	66,2	6 564	68,7	4 614	62,9
7	10 966	64,9	6 282	65,8	4 684	63,8

Les taux sont fondés sur les répondants autres que selon l'approche entonnoir, c'est-à-dire que les non-répondants du cycle précédent sont inclus dans le calcul s'ils ont répondu par la suite. Les taux de réponse cumulatifs des répondants selon l'approche entonnoir sont beaucoup plus faibles. En outre, même si les taux de réponse des deux groupes, soit les 0 à 5 ans et les 6 à 11 ans, sont faibles, la non-réponse a été plus faible dans le groupe des 0 à 5 ans au cycle 7. Comme le montre le tableau, les taux de réponse ont diminué de façon substantielle.

[2] Si la personne de 16 ou 17 ans vit de façon indépendante, elle est maintenue comme PMR.

Étant donné que nous mettons l'accent sur la non-réponse après le cycle 1, il convient de souligner que la baisse des taux de réponse dépend aussi du nombre de personnes que nous avons décidé d'inclure à nouveau dans l'enquête. Le nombre de personnes visées à chaque cycle a diminué, de même que le nombre de répondants. Cela est dû au fait que nous avons retenu des membres que nous considérions comme ayant refusé de façon catégorique ou comme ayant de trop nombreux cycles de non-réponse. Évidemment, l'intensité du refus ou les cycles consécutifs excessifs de non-réponse, sur lesquels est fondée l'admissibilité à l'enquête, peuvent eux-mêmes être liés aux caractéristiques de la personne ou du ménage. À partir du tableau, nous notons aussi une faible augmentation des taux de réponse pour le groupe des 6 à 11 ans entre les cycles 6 et 7. Cela est le résultat du retour de nombreux membres qui n'avaient pas été interviewés pendant plusieurs cycles. Dans l'ensemble, 665 membres de l'échantillon qui sont devenus des répondants à part entière (c.-à-d. âgés de 18 ans et plus) ont été réintégrés à l'enquête selon la règle du cycle 7 de ne pas tenir compte des antécédents de réponse des parents, et parmi eux, 341 ont répondu.

Tableau 3.1-2
Taux de réponse non pondérés selon certaines variables

Variable	Cycle					
	2	3	4	5	6	7
Âge effectif au cycle 7						
18	91,2	87,6	77,3	71,2	65,9	65,2
19	89,3	85,9	77,3	70,4	66,3	63,3
20	91,8	86,8	76,2	72,8	65,1	66,4
21	90,1	87,3	76,3	70,8	63,2	65,0
22	90,8	85,8	76,7	66,4	58,7	62,2
23	91,8	88,0	76,3	66,8	57,5	60,8
Sexe de l'enfant						
Masculin	91,0	87,0	77,5	71,9	64,5	63,2
Féminin	91,1	87,9	78,5	73,4	67,9	66,7
Score au chapitre du soutien social						
Faible	84,7	86,4	71,1	64,3	56,2	54,5
Moyen	89,9	85,5	75,0	69,7	62,6	61,3
Élevé	92,7	89,3	80,7	75,4	69,5	68,2

Nous avons aussi examiné le taux de réponse selon diverses variables du cycle 1, afin de déterminer comment la composition de l'échantillon pouvait changer au fil du temps. Les types de variables examinées comprenaient les variables démographiques, comme l'âge effectif de l'enfant[3], la province de résidence et le type de logement; des variables liées au marché du travail, comme le statut socioéconomique, le ratio du revenu du ménage et des seuils de faible revenu (SFR), et la situation d'activité actuelle de la PMR et du conjoint; ainsi que des variables sociales et des variables liées à la santé, comme la participation à un programme destiné à la petite enfance et le soutien social. Le tableau 3.1-2 montre les taux de réponse pour trois des variables : âge effectif au cycle 7 (18 à 23 ans seulement), sexe de l'enfant et score au chapitre du soutien social au cycle 1.

La hausse légère des taux de réponse entre les cycles 6 et 7 pour les 20 à 23 ans est due principalement au retour de non-répondants qui, au cycle 7, sont devenus des répondants à part entière. Le taux de réponse entre les cycles 6 et 7 a diminué seulement légèrement pour les 20 à 23 ans au cycle 7, qui avaient aussi participé au cycle 6. Par contre, les taux de réponse des 18 et 19 ans au cycle 7, qui avaient aussi participé au cycle 6, ont diminué dans une plus large mesure, ce qui a donné lieu à une baisse légère des taux de réponse pour l'*ensemble* des 18 et 19 ans au cycle 7.

En ce qui a trait au sexe de l'enfant, il existe une différence assez importante entre les enfants de sexe masculin et féminin. Même si elle n'est pas fournie ici, la même évaluation a été effectuée, en mettant l'accent sur la partie de la

[3] Dans l'ELNEJ, l'*âge effectif*, qui est défini comme l'année de référence moins l'année de naissance, est utilisé à la place de l'âge réel au moment de l'interview. On vise ainsi à faire en sorte que les membres demeurent dans les groupes d'âge auxquels ils ont été affectés, peu importe si la collecte se déroule avant ou après leur anniversaire.

cohorte qui a commencé à participer à l'ELNEJ entre 0 et 5 ans. Dans ce groupe d'âge, il n'y a pas beaucoup de différence dans les taux de réponse selon le sexe, en raison du degré élevé d'homogénéité du sexe des parents répondants, à savoir principalement des femmes[4]. Dans le groupe qui a commencé à participer à l'ELNEJ entre 6 et 11 ans, les taux de réponse sont plus élevés pour les enfants de sexe féminin lorsque le répondant est le jeune lui-même.

La troisième variable figurant dans le tableau 3.2-1 est le score au chapitre du soutien social obtenu au cycle 1. La variable est fondée sur un score de facteur des variables qui mesure la perception du soutien social, à savoir si de l'aide et un sentiment de sécurité sont offerts par la famille et les amis; si quelqu'un est disponible pour fournir des conseils; et si le membre fait partie d'un groupe qui partage des intérêts et des préoccupations similaires. À l'origine, la variable du score au chapitre du soutien social était fournie sur une échelle de 0 à 18, mais ici, elle est séparée en trois catégories : faible, moyen et élevé[5]. Les taux de réponse pour la catégorie faible ont diminué le plus rapidement, mais le nombre de répondants appartenant à cette catégorie est relativement peu élevé dans les faits, ce qui rend compte en partie de l'augmentation anormale et soudaine au cycle 3. Toutefois, entre les catégories moyennes et élevées, le soutien social plus grand a tendance à susciter un taux de réponse plus élevé. Cet aspect est examiné de façon plus approfondie dans le présent document.

À partir des taux de réponse pour certaines variables, il existe des indications de non-réponse différentielle en ce qui a trait à certaines variables. L'âge et le sexe de l'enfant sont des variables qui seraient habituellement prises en compte aux étapes de l'ajustement en fonction de la non-réponse et de la stratification a posteriori, ce qui fait que le biais résiduel lié à ces variables devrait être minime. Toutefois, dans le cas d'une variable comme le soutien social, cela n'est pas clair. Nous poursuivons donc par un examen étroit des différents ensembles de variables du cycle 1, afin d'avoir une idée d'où se situe peut-être le biais additionnel.

3.2 Comparaison des répondants et des non-répondants au cycle 7

Parmi les méthodes d'évaluation du biais de non-réponse possible figure la comparaison des caractéristiques des répondants et des non-répondants. Un échantillon est divisé en répondants et en non-répondants, et des estimations de divers paramètres, comme les moyennes de variables numériques ou les proportions pour les niveaux de variables catégoriques, sont calculées séparément pour chaque groupe. Voir, par exemple, Fitzgerald et coll. (1998). Les estimations pour les répondants et les estimations correspondantes pour les non-répondants sont par la suite évaluées, afin de déterminer les différences significatives. Nous mettons l'accent sur l'attrition depuis le cycle 1 de l'ELNEJ, à partir des données du fichier principal du cycle 1. En mettant l'accent sur l'attrition à partir du cycle 1, nous disposons de données pour les répondants et les non-répondants au cycle 7. Aux fins de la présente étude, les non-répondants comprennent les refus, l'absence de contact et les déménagements. Ils ne comprennent pas les décès. Pour évaluer la signification statistique des différences, nous avons calculé les erreurs-types au moyen de la méthode du bootstrap rééchelonné de Rao-Wu (Rao et Wu, 1988; Girard, 2007).

Nous mettons d'abord l'accent sur les variables sociodémographiques de base comme : la province, l'âge du parent, le nombre d'enfants dans le ménage, le niveau le plus élevé de scolarité du parent, le revenu du ménage, le mode d'occupation du logement et le sexe de l'enfant. Il s'agit de variables similaires à celles comprises dans d'autres études de la non-réponse (Fitzgerald et coll., 1998; Watson, 2003; Watson et Wooden, 2004). Étant donné que les 6 à 11 ans sont devenus des répondants à part entière, nous avons aussi envisagé des analyses distinctes pour les 0 à 11 ans, les 0 à 5 ans et les 6 à 11 ans. Comme le montre le tableau 3.2-1, il semble y avoir de nombreuses différences significatives entre les répondants et les non-répondants. Selon l'âge et le niveau de scolarité de la PMR, le nombre d'enfants dans le ménage, le mode d'occupation du logement et le sexe de l'enfant, les non-répondants sont plus susceptibles d'être jeunes, d'avoir des niveaux moins élevés de scolarité, d'appartenir à un ménage comptant un plus grand nombre d'enfants et d'être locataires. Comme le tableau décrit les 0 à 11 ans, les différences selon le sexe de l'enfant sont statistiquement significatives, tout comme dans le cas des 6 à 11 ans. Par contre, les différences selon le sexe de l'enfant ne sont pas significatives pour les 0 à 5 ans, parce que le sexe de la PMR est le facteur plus important.

---

[4] Au cycle 1, 93 % des PMR répondantes étaient de sexe féminin.
[5] Faible = 0 à 8, moyen = 9 à 13, élevé = 14 à 18.

Parmi les autres caractéristiques sociodémographiques comportant des différences importantes figurait la province de résidence, les non-répondants étant plus susceptibles de provenir de l'Ontario et moins susceptibles de provenir du Québec; le statut d'immigrant, les non-répondants étant plus susceptibles d'être des immigrants; et la situation parentale, les non-répondants étant moins susceptibles de vivre avec deux parents. Certaines des caractéristiques qui ne comportaient pas de différence significative étaient la taille du ménage et l'âge de l'enfant. De façon générale, les non-répondants semblaient provenir de ménages comptant un moins grand nombre de membres, mais aucune des différences n'était significative. Par ailleurs, lorsqu'on se limite au groupe des 0 à 5 ans, le nombre de répondants dont la taille du ménage est égale à trois est significativement plus élevé, mais lorsqu'on se limite au groupe des 6 à 11 ans, aucune des variables n'est significative. C'est donc dire que la différence de signification entre le nombre d'enfants dans le ménage et la taille du ménage peut être expliquée en partie par les répercussions différentes des variables sur les deux groupes d'âge. En ce qui a trait à l'âge de l'enfant, étant donné que les questions particulières de l'enquête dépendent de celui-ci, l'absence de différence nous a permis de conclure que les différentes questions/composantes ne sont pas liées à l'attrition.

Tableau 3.2-1
Comparaison des répondants et des non-répondants (0 à 11 ans)

Caractéristique	Répondants	Non-répondants	Différence
Âge de la PMR			
15-24	4,4	6,4	***-2,1***
25-29	15,1	18,8	***-3,8***
30-34	31,9	32,4	-0,5
35-39	29,5	27,1	2,4
40+	19,1	15,1	***4,0***
Nombre d'enfants dans le ménage			
1	21,1	16,8	***4,3***
2-4	76,1	79,4	***-3,3***
5+	2,8	3,7	-0,9
Niveau le plus élevé de scolarité			
Pas de diplôme d'études secondaires	14,0	21,2	***-7,2***
Diplôme d'études secondaires	16,8	21,4	***-4,6***
Plus qu'un diplôme d'études secondaires	29,3	26,5	***2,8***
Diplôme collégial ou universitaire	39,6	30,4	***9,2***
Revenu du ménage			
Moins de 19 000 $	4,7	7,3	***-2,6***
20 000 à 39 999 $	20,3	22,2	-1,9
40 000 à 59 999 $	27,7	23,7	***4,0***
60 000 et plus	34,5	26,0	***8,5***
Mode d'occupation du logement			
Propriétaire	75,3	63,5	***11,9***
Non propriétaire	24,7	36,5	***-11,9***
Sexe de l'enfant			
Masculin	50,3	53,0	***-2,7***
Féminin	49,7	47,0	***2,7***

Nota : Les chiffres **en gras** indiquent un niveau de signification de 5 %. Les estimations ont été calculées au moyen des poids d'enquête du cycle 1.

Lorsque l'on compare les répondants et les non-répondants, mais qu'on limite la cohorte aux 0 à 5 ans au cycle 1, nous obtenons des résultats très similaires pour l'ensemble de la cohorte, sauf pour le sexe de l'enfant, la différence n'étant plus statistiquement significative. À ce moment-là, la PMR est encore la personne qui détermine si une interview sera accordée ou non. Enfin, nous examinons les comparaisons à l'intérieur du groupe des 6 à 11 ans

seulement. De toute évidence, dans le cas de ces enfants plus âgés, le nombre de PMR de 15 à 29 ans est moins élevé. Le groupe des 30 à 34 ans comporte maintenant un nombre significativement plus élevé de non-répondants, par rapport à la cohorte des 0 à 5 ans. En outre, le nombre d'enfants dans le ménage n'a plus d'effet, mais le sexe de l'enfant en a, les répondants étant plus susceptibles d'être de sexe féminin.

3.3 Analyse logistique de la non-réponse

3.3.1 Analyse à partir de variables de base

Nous avons aussi procédé à une analyse multidimensionnelle des mêmes variables qu'à la section 3.2, afin de déterminer la signification des caractéristiques particulières, une fois contrôlées les autres variables. Nous avons élaboré des modèles de régression logistique en utilisant la situation de réponse au cycle 7 comme variable de résultat, et les variables qui viennent d'être abordées comme covariables. Cela a été fait pour chaque groupe d'âge : 0 à 11 ans, 0 à 5 ans et 6 à 11 ans. Le tableau 3.3.1-1 fournit un sommaire des estimations des coefficients de régression pour un sous-ensemble des variables. La première catégorie pour chaque variable est la catégorie de référence. Pour les 0 à 11 ans et les 0 à 5 ans, les PMR qui étaient plus âgées au cycle 1 étaient plus susceptibles de répondre par rapport aux PMR plus jeunes, tandis que pour le groupe des 6 à 11 ans, l'âge de la PMR ne comportait plus de lien significatif avec la réponse, la PMR ne répondant plus pour l'enfant. La taille plus importante du ménage représentée par le nombre d'enfants dans le ménage a tendance à réduire la probabilité de réponse. Toutefois, il n'y a pas d'effet significatif dans le cas où le ménage compte cinq enfants ou plus pour le groupe des 6 à 11 ans, peut-être parce que les enfants sont plus âgés. En ce qui a trait à la scolarité et au mode d'occupation du logement, nous voyons encore une fois une confirmation de nos résultats précédents, un niveau plus élevé de scolarité et la propriété de la résidence semblant être liés à une plus grande probabilité de répondre. Le sexe de l'enfant, lorsque celui-ci est le répondant potentiel, est un facteur important, étant donné que nous avons constaté que les enfants de sexe féminin ont tendance à répondre davantage que ceux de sexe masculin.

Tableau 3.3.1-1
Estimations des paramètres de l'ensemble de base des variables du cycle 1

Caractéristique	0-11	0-5	6-11
Âge de la PMR			
15-24	0,00	0,00	0,00
25-29	0,08	-0,03	-0,30
30-34	***0,26***	***0,31***	-0,35
35-39	***0,35***	***0,45***	-0,25
40+	***0,52***	***0,70***	-0,10
Nombre d'enfants dans le ménage			
1	0,00	0,00	0,00
2-4	***-0,41***	***-0,49***	***-0,28***
5+	***-0,49***	***-0,63***	-0,34
Niveau le plus élevé de scolarité			
Pas de diplôme d'études secondaires	0,00	0,00	0,00
Diplôme d'études secondaires	0,03	0,19	-0,10
Plus qu'un diplôme d'études secondaires	***0,38***	***0,52***	0,25
Diplôme collégial ou universitaire	***0,43***	***0,60***	***0,26***
Mode d'occupation du logement			
Propriétaire	0,00	0,00	0,00
Non propriétaire	***-0,25***	***-0,22***	***-0,28***
Sexe de l'enfant			
Masculin	0,00	0,00	0,00
Féminin	***0,12***	0,04	***0,19***

Nota : Les chiffres **en gras** indiquent un niveau de signification de 5 %. Les estimations ont été calculées au moyen des poids d'enquête du cycle 1.

3.3.2 Analyse à partir des variables de l'étude Starting Out

Dans les sections qui précèdent, on a évalué les variables démographiques de base, en vue d'avoir un aperçu général d'où se situe le biais de non-réponse possible. Toutefois, étant donné que diverses analyses sont effectuées à partir des données de l'ELNEJ, il est difficile de déterminer si les variables d'une analyse particulière ont été affectées par la non-réponse à l'ELNEJ. Il serait par conséquent utile d'évaluer les variables des analyses de fond de l'ELNEJ, afin de déterminer si leur situation actuelle comporte actuellement un potentiel de biais de non-réponse. Nous avons examiné plusieurs études de l'ELNEJ, comme celles de McIntyre (1996) et de Bohatyretz et Lipps (1999). Dans le présent document, nous examinons uniquement les résultats des variables de McIntyre (1996).

L'étude Starting Out (McIntyre, 1996) était une étude comprise dans le rapport *Grandir au Canada*. L'examen a été axé sur la santé des nouveau-nés et des tout-petits au Canada et comprenait un examen des conditions qui contribuent à une grossesse en santé et à la santé de l'enfant à la naissance. Par exemple, les variables sont liées à la santé de la mère pendant la grossesse, au comportement social de l'enfant et à l'environnement familial. Chacune des variables de l'étude s'applique à une fourchette d'âge particulière, soit les 0 à 3 ans. Dans la présente section, nous examinons les variables qui s'appliquent à chacun des trois groupes d'âge : 0 à 3 ans, 0 et 1 an, 2 et 3 ans. Nous sommes intéressés à examiner le rapport entre certaines des variables de l'étude pour les membres de la cohorte originale selon leur statut de réponse au cycle 7.

Pour commencer, nous avons examiné les variables utilisées pour dresser un profil démographique des 0 à 3 ans visés par l'étude. Parmi les variables figuraient les suivantes : sexe de l'enfant, situation du parent, revenu du ménage, âge de la mère à la naissance de l'enfant. Étant donné que ces variables sont similaires à notre ensemble de base, nous avons ajouté trois variables pour un examen plus poussé : âge de la mère à la naissance de l'enfant, statut

d'Autochtone et soutien social. Nous avons déterminé que les mères plus âgées à la naissance de leur enfant étaient moins susceptibles de faire l'objet d'une attrition. Cela est conforme à la situation des PMR plus âgées, qui sont principalement des mères et qui sont moins susceptibles de connaître une attrition. Les PMR comptant sur un soutien social plus grand étaient aussi moins susceptibles de subir une attrition, et on a déterminé que le statut d'Autochtone n'était pas lié à l'attrition. Il n'est pas étonnant de constater que l'inclusion de variables de contrôle, et plus particulièrement l'âge de la PMR, a fait en sorte que l'âge de la mère n'était pas significatif. Toutefois, le soutien social est demeuré significatif.

Le reste de la présente section représente un sommaire de certains des résultats du modèle. Un examen plus détaillé se trouve dans Tam et Waye (2010). Nous avons d'abord mis l'accent sur l'étude de la santé de la mère pendant la grossesse pour les 0 et 1 an. Les principales variables d'intérêt étaient : l'usage du tabac au cours de la grossesse, l'allaitement maternel et la dépression post-partum. Elles ont été entrées dans un modèle logistique de comportement de réponse au cycle 7. Nous avons déterminé que les mères qui avaient fumé pendant leur grossesse étaient plus susceptibles de connaître une attrition, que les mères qui avaient allaité leurs enfants étaient moins susceptibles de connaître une attrition, et que la dépression post-partum ne comportait pas de lien avec l'attrition. Toutefois, une fois que nous avons élargi le modèle pour contrôler les variables sociodémographiques, les variables clés n'étaient plus significatives. D'autres études ont démontré que l'usage du tabac pendant la grossesse était lié à des niveaux plus faibles de scolarité (Millar et Hill, 2004) et que l'allaitement maternel était lié à la fois au revenu et au niveau de scolarité (Heck et coll., 2006). Il s'agit donc de facteurs possibles de la diminution des effets à des niveaux non significatifs.

Dans une autre partie de l'étude, on a examiné les comportements sociaux et émotifs. Ces variables comprenaient : l'hyperactivité, l'agression physique, les troubles émotifs, l'interaction positive et la cohérence des pratiques parentales. En mettant l'accent uniquement sur ces variables dans une analyse logistique, nous avons déterminé que les PMR dont le style parental était l'interaction positive étaient moins susceptibles de subir une attrition plus tard, et que celles dont le style était punitif étaient plus susceptibles de subir une attrition. Toutefois, une fois les spécifications élargies, les variables n'étaient encore une fois plus significatives.

3.3.3 Analyse à partir du groupe le plus âgé (22 et 23 ans)

Dans la présente section, un modèle de régression logistique a été élaboré pour prédire la probabilité de non-réponse au cycle 7, à partir des données d'enquête du cycle 5 pour les membres de la cohorte originale qui étaient âgés de 18 et 19 ans au cycle 5 et qui avaient répondu. Le choix de ce sous-ensemble de la population a été justifié par le fait que le cycle 5 est le premier cycle pendant lequel une partie de la cohorte originale a atteint l'âge de 18 ans. Nous avons élaboré l'hypothèse que le comportement de réponse au cycle 7 dépendrait des données de l'enfant et non pas des données de la PMR lorsque l'enfant avait 18 ans et plus. Par conséquent, nous avons produit un nouveau modèle de non-réponse à partir des données disponibles au cycle 5, afin de déterminer si les caractéristiques individuelles de l'enfant seraient utiles pour prédire la non-réponse subséquente au cycle 7. En outre, ce modèle nous a permis d'utiliser des données plus récentes pour notre évaluation.

On a choisi des variables liées à l'enfant et non pas à la PMR pour ce modèle. Certaines des variables comprises dans le modèle avaient trait au bien-être psychologique, à la santé et aux compétences sociales de l'enfant. Parmi les variables additionnelles incluses dans le modèle figuraient le soutien social, la dépression, l'image de soi, les conflits parentaux, le tabagisme, la participation à un club et la scolarité. Toutefois, nos examens ont démontré que, dans un modèle qui incluait les variables de contrôle de base, seules les activités dans un club se sont révélées significatives dans le modèle logistique. Ce résultat peut être expliqué en partie par le fait que la participation à un club est synonyme d'un degré plus élevé d'interaction sociale, lequel comporte généralement un lien avec une plus grande propension à répondre[6].

[6] La signification de la variable peut aussi être le résultat d'une erreur de type 1. La valeur *p* réelle pour la catégorie significative était de 0,007.

4. Conclusion

Dans le présent document, nous décrivons certains aspects de l'étude de la non-réponse qui a été effectuée. Une analyse unidimensionnelle d'une vaste gamme de variables montre que la non-réponse différentielle existe, même si elle peut souvent être expliquée par le rapport entre la variable à l'étude et les variables sociodémographiques. L'analyse multidimensionnelle des caractéristiques sociodémographiques et des caractéristiques spécialisées fait ressortir un biais de non-réponse possible à divers degrés. Des estimations de paramètres significatives dans des modèles logistiques de réponse, lorsqu'elles sont entrées séparément, ne sont souvent plus significatives lorsque l'on contrôle les variables sociodémographiques. Ainsi, il est clair que le biais dépend de l'étude effectuée. S'il n'y a pas lieu de procéder à des inférences pour la partie de l'échantillon qui semble être biaisée en raison de l'attrition, ou si des variables de contrôle appropriées sont incluses dans l'analyse, le biais peut ne pas poser de problème. Dans l'étude de la non-réponse sur laquelle le présent document est fondé, d'autres analyses, par exemple un examen du rôle joué par l'attrition dans certaines transitions de l'ELNEJ, ainsi qu'un examen des différences possibles entre les non-répondants initiaux et les non-répondants subséquents, sont effectuées.

Remerciements

Nous aimerions exprimer nos remerciements sincères à Sarah Franklin et Joanne Moloney pour leur soutien dans le lancement de ce projet. Nous aimerions aussi faire part de notre appréciation à Dany Faucher, pour son aide et son soutien immenses tout au long du projet. Enfin, nous sommes reconnaissants à Yves Lafortune et Michel Ferland pour leurs commentaires avisés et constructifs, qui ont considérablement amélioré le présent document.

Bibliographie

Bohatyretz, S. et Lipps, G. (1999). Diversity in the Classroom: Characteristics of Elementary Students Receiving Special Education, *Education Quarterly Review,* 6(2), 7-19.

Fizgerald, J., Gottschalk, P. et Moffitt, R. (1998). An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics, *Journal of Human Resources*, 33(2), 251-298.

Girard, C. (2007). How to Avoid Getting All Tied Up Bootstrapping a Survey: A Walk-Through Featuring the National Survey of Children and Youth, document de travail, Ottawa: Statistique Canada.

Heck, K., Braveman, P., Cubbin, C., Chavez, G. et Kiely, J. (2006). Socioeconomic Status and Breastfeeding Initiation among California Mothers. *Association of Schools of Public Health*, 121 (1), 51-59.

McIntyre, L. (1996). Starting Out, *Growing Up in Canada*, Ottawa: Statistics Canada, 47-56.

Millar, W. et Hill G. (2004). Pregnancy and Smoking. *Health Reports*, 15(4), 53-56.

Rao, J.N.K. et Wu, C.F.J. (1988). Re-sampling Inference with Complex Survey Data, *Journal of the American Statistical Association*, 83(401), 231-241

Statistique Canada. (2007). The National Longitudinal Survey of Children: 10-Year Review of the Methodology, rapport non publié, Ottawa: Statistique Canada.

Tam, M. et Waye A. (2010). Non-Response Analysis of the National Longitudinal Survey of Children and Youth, rapport non publié, Ottawa: Statistique Canada.

Watson, D. (2003). Sample Attrition Between Waves 1 and 5 in the European Community Household Panel, *European Sociological Review*, 19(4), 361-378.

Watson, N. et Wooden, M. (2004). Sample Attrition in the HILDA Survey, *Australian Journal of Labour Economics*, 7(2), 293-308.

SESSION 4 – COLLECTE ET APPARIEMENT DE DONNÉES

Recueil du Symposium 2009 de Statistique Canada
Les enquêtes longitudinales : de la conception à l'analyse

Essai expérimental d'une stratégie en vue de rester en contact avec les familles entre les vagues d'une étude par panel

Katherine McGonagle, Mick Couper et Robert Schoeni[1]

Résumé

La Panel Study of Income Dynamics (PSID) est une enquête longitudinale nationale représentative de la population réalisée auprès d'environ 9 000 familles et de leurs descendants qui se déroule depuis 1968. Les familles reçoivent régulièrement par la poste une carte leur demandant de mettre à jour leurs coordonnées afin de pouvoir les suivre entre les vagues de l'enquête. L'analyse révèle que l'obtention de renseignements à jour avant la collecte des données est rentable, car elle réduit les activités de dépistage, les efforts spéciaux de conversion des refus, le nombre de personnes à interviewer et l'érosion de l'échantillon. Le présent article présente les résultats d'une expérience conçue en vue de maximiser la réponse à la demande de mise à jour des coordonnées faite aux familles étudiées.

1. Introduction

1.1 Vue d'ensemble

Faire le suivi des membres de l'échantillon entre les vagues de collecte des données aide à réduire au minimum l'érosion dans les études longitudinales. Toutes choses étant égales par ailleurs, plus l'intervalle entre les vagues de collecte des données est grand, plus il est probable que des personnes faisant partie de l'échantillon aient déménagé et plus il est difficile de retrouver ces personnes (Couper et Ofstedal, 2007; Duncan et Kalton, 1987). En 1997, la Panel Study of Income Dynamics (PSID) est passée d'une interview annuelle à une interview bisannuelle. Afin de connaître les changements résidentiels survenus entre les vagues, l'équipe de la PSID a commencé à envoyer par la poste aux familles une « carte de mise à jour des coordonnées » l'année comprise entre les vagues de collecte de données. Les familles qui mettent à jour ou confirment leur adresse et leur numéro de téléphone et renvoient la carte reçoivent un chèque postpayé de 10 $. Pour les vagues récentes, environ la moitié des familles ont répondu à cet envoi par la poste. Au cours de 2007, les familles qui avaient fourni cette information ont nécessité nettement moins d'efforts de dépistage ou de conversion des refus, et le nombre de personnes à interviewer a été réduit de moitié, ce qui souligne la rentabilité de l'envoi par la poste.

Compte tenu de ces avantages, une étude a été conçue avant les interviews sur le terrain de 2009 dans le but d'améliorer le taux de réponse à la carte de mise à jour des coordonnées envoyée par la poste. Les familles ont été assignées aléatoirement aux conditions d'expérience suivantes : 10 $ comme mesure incitative prépayée comparé à postpayée, présentation de la carte de mise à jour (habituelle au lieu de remaniée), envoi d'un bulletin concernant l'étude, et moment et fréquence de l'envoi par la poste (juillet au lieu d'octobre au lieu des deux fois). Le présent article expose les résultats en ce qui concerne les taux de réponse à la carte envoyée par la poste en fonction de ces diverses conditions d'expérience. Dans l'ensemble, le type de mesure incitative n'a aucun effet, ce qui signifie que des incitatifs postpayés sont plus rentables. La présentation habituelle donne de meilleurs résultats que la présentation remaniée. Le taux de réponse est plus élevé pour les familles qui reçoivent un deuxième envoi par la poste que pour celles qui n'en reçoivent qu'un seul. Il existe certains effets d'interaction avec le moment de l'envoi par la poste, les familles ne recevant la carte qu'en octobre ayant un taux de réponse plus élevé quand elles reçoivent aussi un incitatif prépayé et un bulletin. Nous décrivons les hypothèses concernant ces résultats et les prochaines étapes de l'analyse.

[1] Katherine McGonagle, Mick Couper et Robert Schoeni, Survey Research Center, Institute for Social Research, University of Michigan, États-unis.

1.2 Justification des conditions expérimentales

Diverses catégories de stratégies de prise de contact en vue d'améliorer la conservation du panel ont été utilisées dans des études par panel en cours (voir Couper et Ofstedal, 2007; Laurie et coll., 1999). Cependant, nous ne sommes au courant d'aucun essai expérimental ayant trait à des stratégies particulières. Bien que la question de l'efficacité de la stratégie de prise de contact n'ait pas été suffisamment étudiée, la littérature sur les méthodes d'enquête offre des suggestions quant au moyen d'améliorer les taux de réponse. Premièrement, sous certaines conditions, une mesure incitative prépayée augmente la coopération en accroissant l'importance de la mesure incitative, ainsi que le sentiment de réciprocité du répondant (Singer et coll., 1999). En outre, une présentation remaniée de la carte de mise à jour des coordonnées du répondant devrait accroître l'importance de la demande, ce qui pourrait accroître la coopération (Dillman, 2000).

De surcroît, le moment et la fréquence de la demande peuvent influer sur la coopération. Idéalement, la demande doit être faite à un moment suffisamment proche de la collecte des données à venir pour que la plupart des changements de résidence soient enregistrés, mais pas proche au point de donner l'impression que le fardeau de réponse augmente, ce qui pourrait se produire si les prises de contact sont trop nombreuses. L'interview de la PSID a lieu les années impaires, entre mars et novembre. Donc, il a été décidé de procéder à l'envoi par la poste deux fois dans l'année : à mi-chemin entre la fin de la vague précédente et le début de la suivante (c.-à-d. juillet) et à un moment aussi avancé que possible dans l'année pour mettre à jour les adresses avant que commence la production (c.-à-d. octobre). Une troisième condition d'expérience : envoi en juillet avec un nouvel envoi en octobre pour les non-répondants, a été choisie pour examiner l'effet de deux contacts au lieu d'un sur les taux de réponse.

Enfin, on ne dispose guère de données sur le nombre idéal de contacts avec les répondants. Au cours de l'horizon de deux ans d'une enquête bisannuelle, les répondants participent à une longue interview, reçoivent un bulletin concernant l'étude, une demande de mise à jour de leurs coordonnées et une lettre les avertissant de la prochaine interview. À quel moment ces contacts multiples deviennent-ils pesants? Accroissent-ils plutôt le sentiment d'identification à l'enquête? Il est probable que ces perceptions varient selon les caractéristiques des membres de l'échantillon. Donc, dans la présente étude, la manipulation de l'envoi par la poste comportait la modification des caractéristiques de la mesure incitative, de la présentation de la carte, du moment de l'envoi et du nombre de prises de contact effectuées en s'appuyant sur les données de la littérature sur les méthodes d'enquête.

2. Méthodes

Les familles de la PSID admissibles à l'interview de 2009 (n=8 512) ont été assignées aléatoirement à quatre conditions qui définissent un plan d'expérience 2 (« bulletin ») x 3 (« moment de l'envoi ») x 2 (« présentation ») x 2 (« mesure incitative ») (tableau 2-1). Pour manipuler le nombre de prises de contact avec le répondant et le fardeau de réponse, un bulletin concernant l'étude a été envoyé à la moitié des familles un an avant le début de l'interview. La deuxième condition concernait le moment de l'envoi par la poste qui, dans un tiers des cas, a été fait en juillet, dans un autre tiers, en octobre, et dans un troisième tiers, initialement en juillet avec un envoi de suivi en octobre pour les non-répondants. La troisième condition était la présentation de la carte envoyée par la poste, la moitié des familles recevant la carte habituelle en noir et blanc et l'autre moitié, une présentation remaniée (annexe I). La quatrième condition variait selon que l'incitatif de 10 $ était prépayé ou postpayé.

Tableau 2-1
Taux de réponse selon la condition expérimentale

Bulletin	Moment de l'envoi	Présentation	Incitatif	Taille de l'échantillon	Taux de réponse (%)
Oui	Juillet	Habituelle	Prépayé	330	57,9
Oui	Juillet	Habituelle	Postpayé	371	58,8
Oui	Juillet	Remaniée	Prépayé	422	53,8
Oui	Juillet	Remaniée	Postpayé	339	50,7
Oui	Juillet avec suivi en octobre	Habituelle	Prépayé	383	65,0
Oui	Juillet avec suivi en octobre	Habituelle	Postpayé	329	64,1
Oui	Juillet avec suivi en octobre	Remaniée	Prépayé	298	65,4
Oui	Juillet avec suivi en octobre	Remaniée	Postpayé	331	68,0
Oui	Octobre	Habituelle	Prépayé	395	60,8
Oui	Octobre	Habituelle	Postpayé	334	58,4
Oui	Octobre	Remaniée	Prépayé	330	57,9
Oui	Octobre	Remaniée	Postpayé	394	50,0
Total				**4256**	
Non	Juillet	Habituelle	Prépayé	366	61,5
Non	Juillet	Habituelle	Postpayé	346	63,8
Non	Juillet	Remaniée	Prépayé	351	55,6
Non	Juillet	Remaniée	Postpayé	324	54,0
Non	Juillet avec suivi en octobre	Habituelle	Prépayé	329	65,1
Non	Juillet avec suivi en octobre	Habituelle	Postpayé	386	65,8
Non	Juillet avec suivi en octobre	Remaniée	Prépayé	313	67,1
Non	Juillet avec suivi en octobre	Remaniée	Postpayé	376	67,8
Non	Octobre	Habituelle	Prépayé	325	59,4
Non	Octobre	Habituelle	Postpayé	401	54,4
Non	Octobre	Remaniée	Prépayé	369	52,6
Non	Octobre	Remaniée	Postpayé	370	47,0
Total				**4256**	

3. Résultats

3.1 Effets globaux du traitement

L'examen des taux de réponse au cours de 2009 montre que 60 % des familles ont envoyé une mise à jour ou une confirmation de leurs coordonnées. Ce taux ne variait pas selon que la famille avait reçu un bulletin ou non, ni selon le type d'incitatif. Le moment de l'envoi avait un effet significatif, le taux de réponse étant plus élevé pour l'envoi en juillet avec suivi en octobre (66 %) que pour l'envoi en juillet seulement (57 %) et pour l'envoi en octobre seulement (55 %). Enfin, le taux de réponse était significativement plus élevé pour les familles qui avaient reçu la carte dans sa présentation habituelle que pour celles qui avaient reçu la présentation remaniée (61 % contre 57 %).

La présentation habituelle a produit des résultats significativement meilleurs indépendamment de la réception d'un bulletin ou du type d'incitatif. Aucun effet différentiel du type d'incitatif ne se dégage selon la réception d'un bulletin ou la présentation de la carte envoyée par la poste.

Les effets du bulletin, de l'incitatif et de la présentation ont été examinés pour chaque condition de moment de l'envoi au moyen de modèles de régression logistique. Pour les cas d'envoi en juillet uniquement, la présentation habituelle de la carte a produit des résultats significativement meilleurs que la présentation remaniée, et ni le type d'incitatif, ni le bulletin n'avait d'effet. Pour les cas d'envoi en octobre seulement, un taux de réponse significativement plus élevé était associé à la présentation habituelle de la carte, à l'incitatif prépayé et à la réception d'un bulletin. Enfin, pour les cas d'envoi en juillet avec suivi en octobre, le bulletin, le type d'incitatif et le moment de l'envoi n'avaient aucun effet significatif.

3.2 Effets du traitement selon les variables démographiques clés

Un objectif central de l'expérience courante est de concevoir une stratégie de prise de contact avec les répondants qui maximise les chances que les coordonnées de toutes les familles participant à l'étude seront mises à jour avant les interviews de la vague suivante, le but ultime étant de réduire au minimum l'érosion du panel. Une première série d'analyses ont été effectuées afin de déterminer si les effets globaux du traitement avaient des incidences différentielles en fonction des caractéristiques démographiques clés que l'on sait être associées à l'érosion du panel, dont la situation de personnes ayant déménagé et le revenu familial. Le taux de réponse à l'envoi par la poste était plus faible pour les répondants appartenant à des familles qui avaient déménagé ou qui avaient la ferme intention de le faire que pour ceux n'ayant pas déménagé (56 % comparativement à 66 %). Comme dans l'échantillon global, les deux groupes étaient plus susceptibles de répondre à la carte sous sa présentation habituelle qu'à celle dont la présentation était remaniée, et à la carte envoyée en juillet avec suivi en octobre qu'à celle envoyée seulement en juillet ou seulement en octobre. Fait intéressant, les personnes n'ayant pas déménagé qui ont reçu l'incitatif prépayé avaient un taux de réponse significativement plus élevé que celles ayant reçu l'incitatif postpayé (68 % comparativement à 65 %, respectivement), tandis que le type d'incitatif n'avait pas d'effet sur le taux de réponse des personnes ayant déménagé. Les analyses portant sur les effets des conditions de l'expérience selon l'appartenance au quartile supérieur par opposition au quartile inférieur de revenu ont confirmé les tendances globales présentées pour l'échantillon total, le type d'incitatif et la réception d'un bulletin n'ayant aucun effet chez l'un ou l'autre groupe, et la présentation traditionnelle de la carte et l'envoi de celle-ci en juillet avec un suivi en octobre augmentant les taux de réponse chez les deux groupes.

4. Discussion

Quelles leçons pouvons-nous tirer de cette expérience? Premièrement, un envoi par la poste de suivi aux non-répondants constitue une stratégie efficace, peu coûteuse, qui pourrait, en dernière analyse, réduire les activités de dépistage coûteuses durant la collecte des données. Cette condition a produit des taux de réponse supérieurs de 9 et de 11 points de pourcentage à celui obtenu pour les deux conditions ne comportant qu'un seul envoi par la poste.

Deuxièmement, et étonnamment, la présentation habituelle de la carte a donné de meilleurs résultats. Dans les deux présentations, les dernières coordonnées connues étaient imprimées sur une carte que les répondants ont pliée, scellée à l'aide d'un autocollant attaché à la carte et envoyée par la poste. La présentation remaniée comportait aussi de longues instructions décrivant la nécessité de détacher la carte avant de la plier. Il se pourrait que les instructions supplémentaires aient donné l'impression que le renvoi de la carte était compliqué et ont découragé les répondants de le faire. Nous testerons cette hypothèse en 2010 en éliminant l'« étape du détachement de la carte » et les instructions connexes, et en modifiant uniquement la couleur et la présentation.

Troisièmement, l'incitatif prépayé n'a donné de meilleurs résultats que l'incitatif postpayé que dans le cas de l'envoi en octobre. Peut-être était-il comparativement attrayant en octobre, car l'économie américaine commençait à se défaire. Cependant, dans l'ensemble, les incitatifs prépayés n'étaient pas rentables, car ils n'augmentaient pas les taux de réponse comparativement aux incitatifs postpayés. La constatation que les familles ayant reçu le bulletin avaient un taux de réponse plus élevé en octobre que celles qui ne l'avaient pas reçu pourrait montrer l'importance qu'il y a à maintenir le contact avec les familles participant à l'étude. Les familles qui ont reçu la carte de mise à jour en octobre seulement et à qui on n'avait pas envoyé de bulletin n'avaient pas eu de contact avec l'étude pendant au moins 10 mois, et la plupart d'entre elles n'en avaient pas eu pendant 16 mois ou plus. Le taux de réponse de ces familles était inférieur de 4 % à celui des familles qui avaient reçu le bulletin. Conjugué à la constatation que les taux de réponse étaient plus élevés dans le cas d'un envoi par la poste avec suivi, ce résultat donne à penser que le

moment le plus prometteur consiste en un contact initial environ huit mois avant le début de la collecte des données, avec un suivi trois mois plus tard au besoin.

La prochaine étape de l'analyse consistera à examiner l'effet des diverses conditions de traitement sur l'effort sur le terrain et les résultats des activités de production durant la collecte des données de le PSID de 2009, y compris les taux de dépistage et de conversion des refus, le nombre de tentatives de prise de contact nécessaires pour obtenir un résultat final et l'érosion. Afin de déterminer si certains sous-groupes sont plus sensibles à certains aspects du protocole de traitement que d'autres, ce qui pourrait éventuellement mener à une approche taillée sur mesure de l'élaboration des stratégies de prise de contact, des analyses multivariées des effets du traitement en fonction de diverses conditions sociodémographiques seront effectuées, étant donné la forte corrélation de caractéristiques telles que le revenu, la situation de personnes ayant déménagé et l'âge. Cette information nous aidera à concevoir une stratégie efficace en vue de suivre les familles du panel.

Bibliographie

Couper, M.P. et Ofstedal, M.B. (2009). Keeping in Contact with Mobile Sample Members, in P. Lynn (ed.), *Methodology of Longitudinal Surveys*, New York: Wiley, pp. 183-203.

Dillman, D.A. (1987). *Mail and Internet Survey: The Tailored Design Method*, New York: Wiley.

Duncan, G.J. et Kalton, G. (2000). Issues of Design and Analysis of Surveys across Time, *International Statistical Review*, 55, pp. 97-117.

Laurie, H., Smith, R. et Scott, L. (1999). Strategies for Reducing Nonresponse in a Longitudinal Panel Survey, *Journal of Official Statistics,* 2, pp. 269-282.

Singer, E.S., Gebler, N., Raghunathan, T., Van Hoewyk, J. et McGonagle, K. (1999). The Effect of Incentives on Response Rates in Face-to-Face, Telephone, and Mixed Mode Surveys: Results of a Meta-Analysis, *Journal of Official Statistics*, 15, pp. 217-230.

ANNEXE I : CONDITIONS DE PRÉSENTATION

Présentation habituelle

Please accept this $10.00 check as a token of our appreciation!

If your name, address and phone number on the label below are correct, please check the "No Changes" box.

OR

If your name, phone and/or address on the label have changed, please make the necessary changes on the form below.

Then fold the postcard in half and seal the open edge shut with the sticker on the right.

Thank you for your help!

No Changes.
(Please check this box if the information below is correct.)

Affix label here

Please make necessary changes in name, phone, and address below:

Name: _____

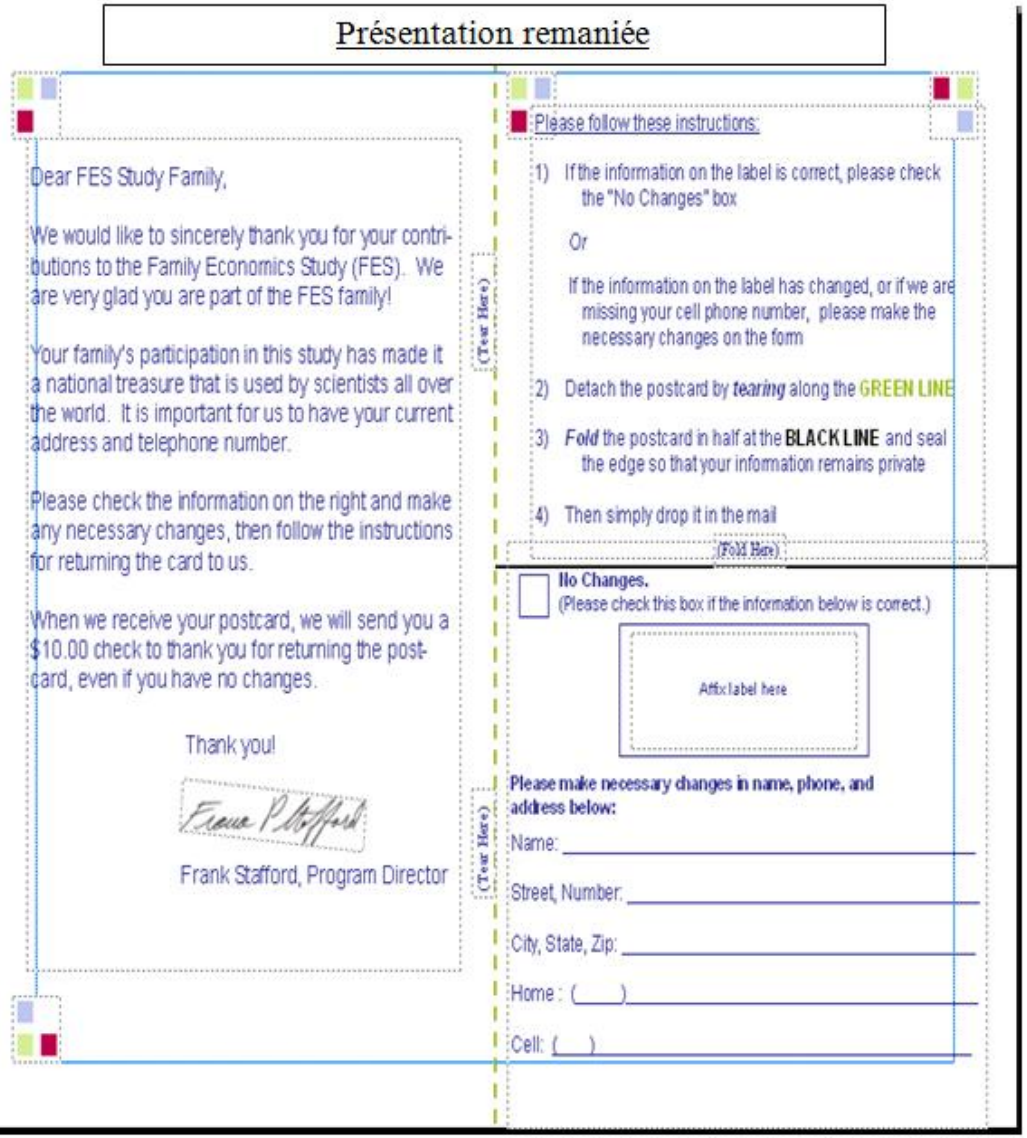
Street, Number: _____

City, State, Zip: _____

Home : ( ) _____

Cell: ( ) _____

Présentation remaniée



Les modes de collecte mixtes et multiples : l'expérience de l'enquête HILDA

Nicole Watson et Mark Wooden[1]

Résumé

Dans le présent article fondé sur l'expérience de l'enquête Household, Income and Labour Dynamics (HILDA) menée en Australie, nous cherchons à savoir si l'utilisation de différents modes de collecte des données a des conséquences involontaires. Plus précisément, nous tâchons de répondre à quatre questions clés. Premièrement, quelles caractéristiques des répondants sont les meilleurs prédicteurs du choix du mode d'enquête? Deuxièmement, le choix du mode a-t-il un effet sur la réponse, particulièrement aux vagues subséquentes? Troisièmement, les modes mixtes et multiples ont-ils un effet sur la non-réponse partielle? Enfin, peut-on prouver que les réponses diffèrent systématiquement selon le mode?

Mots-clés : mode mixte, mode multiple, non-réponse, qualité des données, enquêtes longitudinales.

1. Introduction

Il est généralement reconnu que les plans d'enquête à mode de collecte mixte peuvent améliorer les taux de réponse (p. ex., Day et coll., 1995; Voogt et Saris, 2005). Il est également bien établi que le choix du mode d'enquête peut avoir une incidence sur les réponses (p. ex., De Leeuw, 1992). Toutefois, nous en savons beaucoup moins sur l'effet de la combinaison de plusieurs modes dans les enquêtes longitudinales. Dillman (2009) s'appuie dans une large mesure sur l'expérience des enquêtes transversales pour formuler des hypothèses quant à l'effet du changement du mode d'enquête au fil du temps, mais à part cette étude, les articles de recherche publiés sur cette question sont rares.

Dans le présent article fondé sur l'expérience de l'enquête Household, Income and Labour Dynamics (HILDA) menée en Australie, nous cherchons à savoir si, dans les enquêtes à mode mixte et à mode multiple, l'interaction des différents modes peut avoir des conséquences involontaires. Plus précisément, nous tâchons de répondre à quatre questions clés. Premièrement, quelles caractéristiques des répondants sont les meilleurs prédicteurs du choix du mode d'enquête? Deuxièmement, le choix du mode a-t-il un effet, positif ou négatif, sur la réponse, particulièrement aux vagues subséquentes? Troisièmement, les modes mixtes et multiples ont-ils une incidence sur la quantité de données recueillies? Enfin, peut-on prouver que les réponses diffèrent systématiquement selon le mode?

2. L'enquête HILDA

L'enquête HILDA est une enquête-ménage par panel axée sur l'emploi, le revenu et la famille. Modélisée sur les enquêtes-ménages par panel réalisées dans d'autres pays et décrite en plus de détail dans Wooden et Watson (2007), elle a été lancée en 2001 auprès d'un grand échantillon probabiliste national de ménages australiens occupant des logements privés. L'ensemble des membres des ménages répondants à la vague 1 constitue la base du panel devant être interrogé à chaque vague subséquente (même si les interviews sont menées seulement auprès des membres des ménages âgés de 15 ans et plus), une année environ s'écoulant entre chaque vague d'interviews. Comme de nombreux autres panels de ménages, l'échantillon est élargi chaque année de manière à inclure les nouveaux membres des ménages résultant de changements dans la composition des ménages originaux. À l'instar d'autres enquêtes-ménages par panel, l'enquête HILDA recueille des données au moyen de modes multiples. Plus précisément, elle utilise à la fois des *modes multiples*, en ce que les données sont recueillies auprès de tous les répondants au moyen d'une interview sur place ainsi que d'un questionnaire papier autoadministré, et des *modes mixtes*, en ce que les répondants et les intervieweurs ont le choix entre le contact en personne ou par téléphone pour le déroulement de la composante de l'interview.

[1] Nicole Watson, Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Vic 3010, Australia (n.watson@unimelb.edu.au); Mark Wooden, Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Vic 3010, Australia (m.wooden@unimelb.edu.au).

La durée de l'interview varie, tant d'une personne à l'autre que d'une vague à l'autre, mais en général le but est de veiller à ce que les intervieweurs ne consacrent pas plus de 83 minutes en moyenne à l'interview dans le cas d'un ménage composé de deux adultes. Un questionnaire à remplir soi-même (QRSM) est fourni à toutes les personnes qui participent à une interview sur place, et il est habituellement recueilli par l'intervieweur à une date ultérieure, ou bien les répondants sont priés de retourner le formulaire rempli par la poste. Le QRSM se compose principalement de questions qui sont difficiles à administrer rapidement dans une interview en personne ou auxquelles les répondants peuvent se sentir un peu mal à l'aise de donner une réponse lors d'une interview sur place.

La conception de l'enquête HILDA a également permis de mener des interviews par téléphone plutôt qu'en personne. À la vague 1, le téléphone a été utilisé surtout en dernier recours, de sorte que la fréquence des interviews téléphoniques est faible (à peine 0,5 % de toutes les interviews ont été menées par téléphone). Aux vagues subséquentes, il a été nécessaire d'avoir davantage recours aux interviews téléphoniques, en raison du coût des visites aux membres de l'échantillon ayant déménagé dans des endroits hors de la portée « habituelle » de notre réseau d'intervieweurs. En outre, même s'il ne s'agit pas d'une option offerte explicitement, les interviews téléphoniques ont été menées auprès des membres de l'échantillon ayant indiqué une vive préférence pour ce mode de collecte des données d'enquête. La fréquence des interviews téléphoniques a donc augmenté pendant la durée du panel; à la vague 8, des interviews ont été menées par téléphone dans une proportion de 10,1 % (voir le tableau 5-1).

3. Qui est interviewé par téléphone?

En premier lieu, nous tâchons de déterminer si les types de répondants interviewés par téléphone diffèrent systématiquement de ceux interviewés sur place. Comme notre résultat est binaire, nous estimons des modèles probit qui comportent une distinction entre le mode utilisant des données regroupées des vagues 2 à 7 et celui permettant d'établir une corrélation entre les observations provenant d'une même personne. (Les données de la vague 1 sont exclues parce que, d'une part, le plan de l'étude a empêché de mener de nombreuses interviews téléphoniques à la vague 1 et, d'autre part, nous incluons les caractéristiques mesurées à la vague précédente dans notre liste de variables de contrôle.) Nous estimons deux modèles, l'un utilisant les observations des répondants à la vague précédente et l'autre, utilisant toutes les observations restantes (bien que, par souci de concision, nous présentons ici seulement les résultats du premier modèle et nos commentaires portent uniquement sur ce dernier). La dynamique est prise en compte par l'inclusion d'une variable identifiant le mode d'interview à la vague précédente. Un résumé des résultats des coefficients d'intérêt est présenté au tableau 3-1.

Nous avons posé comme hypothèse que l'interview téléphonique est privilégiée par rapport à l'interview sur place par i) les répondants qui ont changé d'adresse, particulièrement s'ils ont déménagé dans une région rurale en Australie; ii) les répondants qu'il serait plus difficile de joindre à la maison en raison de leurs heures de travail; iii) les personnes ayant une incapacité ou une maladie; et iv) les personnes qui préféreraient qu'on ne communique pas avec elles et qui se sont montrées peu disposées à participer. Nous avons trouvé de fortes preuves à l'appui de toutes les quatre hypothèses. Toutes autres choses constantes, la probabilité qu'une personne ayant déménagé soit interviewée par téléphone à la vague suivante est plus du double de celle d'une personne n'ayant pas déménagé (8,9 % par rapport à 4,3 %). Par ailleurs, que la personne ait déménagé ou non, la probabilité d'une interview téléphonique augmente avec la distance du répondant d'une grande ville. En outre, comme prévu, nous avons constaté que les personnes occupées sont significativement plus susceptibles d'être interviewées par téléphone; cependant, nous ne nous attendions pas à la probabilité relativement élevée d'interviews téléphoniques chez les personnes en chômage, dont la probabilité prévue de subir une interview téléphonique est environ la même que celle d'un employé travaillant 36 heures par semaine. Toutefois, il n'y a pas de lien entre les interviews téléphoniques et les horaires de travail irréguliers. Nous avons constaté également que les personnes ayant des problèmes de santé chroniques classés comme étant « graves » (c'est-à-dire empêchant complètement le répondant de travailler) sont beaucoup plus susceptibles (environ 1,4 fois plus) d'être interviewées par téléphone. En ce qui a trait à la quatrième l'hypothèse , les évaluations des intervieweurs de la question de savoir si un répondant s'est montré « coopératif » ou « méfiant » au sujet des objectifs de l'enquête se sont révélées être des prédicteurs puissants du choix de mode à la vague suivante. Nous avons constaté aussi que le non-retour du QRSM dans une vague, qui selon nous est un indicateur de l'engagement du répondant à l'égard de l'enquête et de l'intérêt qu'il porte à cette dernière, est un prédicteur puissant de son interview par téléphone à la vague suivante.

Tableau 3-1
Modèle probit de prédiction des résultats des interviews téléphoniques chez les répondants des vagues précédentes, vagues 2 à 7 de l'enquête HILDA – Certains coefficients et statistiques sommaires

Variable explicative	Coeffi-cient	Erreur-type robuste	Variable explicative	Coeffi-cient	Erreur-type robuste
Situation d'interview à la vague précédente			Caractéristiques des répondants		
Interview téléphonique	1,202**	,034	A déménagé	,461**	,022
QRSM non retourné	,233**	,032	Âge	-,052**	,009
Coopératif	-,134*	,068	En chômage	,109*	,051
Méfiant	,196**	,057	Heures habituelles de travail par semaine (x 10^{-2})	,301**	,001
Compréhensif	,110*	,053	Heures irrégulières	-,012	,027
Charge de travail de l'intervieweur (x 10-2)	-2,395**	,091	Maladie chronique grave	.216**	,074
Charge de travail de l'intervieweur au carré (x 10^{-4})	,917**	,038	Lieu (référence = Sydney)		
Variables nominales de vague (référence = vague 2)			Melbourne	-.273**	,038
Vague 3	,173**	,036	Brisbane	-.195**	,045
Vague 4	,240**	,034	Adélaïde	-.264**	,055
Vague 5	,212**	,034	Perth	-.285**	,055
Vague 6	,258**	,034	Autre grande ville	-.143*	,044
Vague 7	,436**	,033	Régional intérieur	.119**	,032
			Régional périphérique	.286**	,037
			Éloigné	.737**	,054
Pseudo R au carré	,229				
N	Observations = 69 548; personnes = 15 520				

* p <0,05; ** p <0,01. Sont inclus également mais non présentés les résultats des variables de contrôle pour le sexe, le pays de naissance, l'identité autochtone, l'état matrimonial, la taille du ménage, le niveau de scolarité, le revenu du ménage mis en équivalence, l'état de locataire et le secteur socioéconomique du quartier.

Il convient de souligner également que nous avons observé un rapport en forme de U entre les charges de travail des intervieweurs et la probabilité d'une interview téléphonique. Enfin, après prise en compte des effets d'une large gamme de facteurs, il demeure manifeste que la fréquence des interviews téléphoniques augmente au fil du temps, ce qui donne à penser que d'autres facteurs entrent en jeu, comme l'évolution des préférences des répondants ainsi que des intervieweurs.

4. Mode d'interview et érosion du panel

Comme nous l'avons indiqué plus haut, il est bien établi que les taux de réponse augmentent habituellement lorsqu'on offre aux membres de l'échantillon un choix de modes de collecte des données. Il n'est pas clair, cependant, si ces résultats comprennent des taux d'érosion plus élevés dans un contexte longitudinal. Comme nous l'avons souligné antérieurement (Watson et Wooden, 2009, p. 164), les données d'enquêtes longitudinales donnent à penser essentiellement que l'interview téléphonique, lorsqu'elle est utilisée dans des enquêtes menées principalement au moyen d'interviews sur place, est un prédicteur d'un taux d'érosion plus élevé (plutôt que plus faible) à la vague suivante. Ce résultat laisse supposer que l'utilisation d'autres modes d'enquête ne peut que retarder la décision de cesser de coopérer. Toutefois, il est vrai également que, dans la plupart de ces enquêtes, l'interview téléphonique était utilisée principalement en dernier recours. Par conséquent, répondre par téléphone était en soi un indicateur de manque d'intérêt à l'endroit d'une enquête ou de réticence à y participer.

En revanche, l'enquête HILDA est différente en ce qu'à l'exception de la vague 1, la collecte des collectes se fait par interview téléphonique principalement pour des raisons liées aux coûts. C'est peut-être ce qui explique le mieux pourquoi nous en sommes arrivés à des conclusions très différentes de celles de notre analyse des données de l'enquête HILDA provenant des vagues 2 à 4 (Watson et Wooden, 2009). Nous avons estimé des modèles probit indépendants prédisant les taux de réponse à l'enquête à la vague *t*, à la condition d'une prise de contact, comme fonction des caractéristiques des interviews et des répondants au temps *t-1*. Après inclusion d'un grand nombre de covariables, y compris une gamme de mesures dont nous avons

supposé qu'elles sont corrélées aux attitudes des répondants à l'égard de la participation à l'enquête, nous avons constaté que le résultat de l'interview téléphonique au temps *t-1* est associé à une plus forte probabilité de réponse au temps *t*. Toutefois, ce résultat n'est que faiblement significatif et une nouvelle estimation de ces modèles utilisant les données recueillies jusqu'à la vague 7 montre que la grandeur du coefficient de cette variable diminue fortement et devient statistiquement non significative.

Nous avons également vérifié la signification des effets d'interaction entre le mode d'enquête et d'autres covariables. La plupart des termes d'interaction ne sont pas significatifs, sauf une exception notable. La non-réponse partielle est généralement un prédicteur de l'érosion à la vague suivante, particulièrement lorsque les interviews sont menées par téléphone.

Enfin, nous avons inclus une variable explicative identifiant toutes les interviews menées au cours de la dernière étape de collecte des données. Comme prévu, cette variable explicative est négative et fortement significative, indiquant que les répondants qui sont les plus difficiles à joindre à une vague donnée sont beaucoup moins susceptibles de répondre à la vague suivante. Néanmoins, l'inclusion de cette variable a eu un effet minimal sur l'incidence estimée du mode d'enquête.

Dans l'ensemble, peu de données de l'enquête HILDA laissent supposer que le mode d'interview est associé à l'érosion aux vagues subséquentes. L'effet est positif mais petit et fort imprécis. Cela dit, nous reconnaissons qu'en l'absence d'une expérience, nous n'avons pu déterminer le contrefactuel approprié.

5. L'effet du mode d'interview sur la non-réponse partielle

Même si le mode d'enquête n'a pas de conséquences pour la réponse totale, la quantité de données d'enquête recueillies peut varier d'un mode d'enquête à l'autre, étant donné les écarts entre les taux de non-réponse partielle. Cela peut se produire dans le cas de l'enquête HILDA en raison des différences en matière de taux de retour des QRSM ou de fréquence des refus et des réponses « ne sait pas ».

5.1 Répercussions des interviews téléphoniques sur les taux de réponse au QRSM

La diminution de la probabilité de retour des QRSM est une conséquence évidente du choix des répondants de l'option d'interview téléphonique. Lorsque les interviews sont menées sur place, les QRSM sont remis directement aux personnes interrogées, et des arrangements sont pris pour que les intervieweurs ramassent les questionnaires remplis. D'ailleurs, dans de nombreux cas, l'intervieweur peut partir avec le QRSM rempli en mains le jour où il mène l'interview. Dans le cas des interviews téléphoniques, toutefois, le QRSM doit être envoyé par la poste à la personne interrogée, accompagné d'instructions lui demandant de retourner le questionnaire dûment rempli par la poste. Bien sûr, il est bien connu que les enquêtes avec envoi et retour du questionnaire par la poste sont associées à des taux de réponse beaucoup plus faibles que les enquêtes menées au moyen de contacts plus personnels, et l'enquête HILDA ne fait pas exception.

Comme le montre le tableau 5-1, les taux de réponse au QRSM baissent de façon plus ou moins constante depuis le début de l'enquête HILDA. Le tableau 5-1 donne à penser également qu'une bonne partie de cette baisse s'explique par l'augmentation au fil du temps de la fréquence des interviews téléphoniques. Chez les répondants interviewés par téléphone, le taux moyen de réponse au QRSM est de 63 % seulement, tandis qu'il est de 93 % chez les répondants interviewés sur place, bien que dans ce dernier cas la tendance reste nettement à la baisse. C'est ce que confirment les résultats de l'estimation d'un modèle probit utilisant des données regroupées qui a révèle, après prise en compte de la vague d'interviews ainsi que d'autres caractéristiques des interviews et des répondants, que le taux de réponse au QRSM des répondants interviewés par téléphone est de 23 points de pourcentage inférieur à celui des répondants autrement comparables interviewés sur place. Les résultats confirment en outre qu'indépendamment de cet effet de mode, le taux de réponse au QRSM accuse une tendance à la baisse au fil du temps. À la vague 7, on estime que le taux de réponse au QRSM a diminué de 3,3 points de pourcentage après prise en compte des différences observées entre les individus.

Tableau 5-1
Taux de réponse au QRSM (%) selon la vague, vagues 1 à 8 de l'enquête HILDA

	V1	V2	V3	V4	V5	V6	V7	V8
Taux de réponse global	93,5	93,0	92,3	91,9	89,9	90,8	89,0	87,6
Interviews sur place	93,7	93,9	93,5	93,3	91,8	92,6	91,4	90,7
Interviews téléphoniques	52,7	63,3	68,1	68,2	62,3	64,9	63,2	59,4
Interviews téléphoniques en % du total des interviews	0,5	3,0	4,6	5,6	6,5	6,6	8,5	10,1

5.2 Mode d'interview et fréquence des refus et des réponses « ne sait pas »

Au tableau 5.2, nous faisons état de la fréquence des refus et des réponses « ne sait pas » (c.-à-d. non-réponse partielle) pour certaines sous-sections de nos interviews ainsi que pour les QRSM retournés. Pour les cinq composantes d'interview, les taux de non-réponse partielle sont plus élevés dans le cas des répondants interviewés par téléphone. Le niveau général de non-réponse partielle, toutefois, est très faible, laissant supposer que la question revêt peu d'importance.

Tableau 5-2
Non-réponse partielle (%) selon le mode d'interview, vagues 2 à 7 de l'enquête HILDA

Composante du questionnaire	Interview sur place	Interview téléphonique	Différence probable = 0
Soins aux enfants/logement (QL)	0,97	1,84	,000
Emploi/population active	0,07	0,11	,000
Revenu	0,55	0,87	,000
Formation de la famille/relations familiales	0,24	0,32	,002
Modules d'interview « spéciaux »	0,79	1,26	,000
QRSM (si retourné)	1,83	0,94	,000

En revanche, les taux de non-réponse partielle sont, comme on pourrait s'y attendre, plus élevés dans le QRSM. Dans ce cas, les personnes interviewées par téléphone ont en fait de meilleurs résultats. Nous sommes d'avis que cela est simplement fonction de l'autosélection. Autrement dit, les répondants les plus susceptibles de ne pas répondre aux questions seront les mêmes que ceux qui n'ont pas retourné un QRSM, sans oublier que la fréquence de la non-réponse au QRSM est beaucoup plus élevée chez les personnes interviewées par téléphone.

6. Mode d'interview, qualité des données et réponses des répondants

Les dernières questions que nous souhaitons examiner sont celles de savoir si la qualité des réponses diffère selon le mode de réalisation de l'enquête et si les réponses aux questions sont sensibles au mode. Nous présentons tout d'abord, au tableau 6-1, un éventail restreint de mesures qui pourraient être des indicateurs de la qualité des données. Tel que prévu, les interviews téléphoniques sont nettement plus courtes que les interviews sur place, ce qui pourrait laisser supposer des réponses moins réfléchies ou bien simplement refléter la plus courte durée de l'interaction sociale dans le cas des interviews téléphoniques. Nous constatons également que la longueur de certaines réponses (p. ex., au sujet de l'emploi) est légèrement plus courte au téléphone. Toutefois, les évaluations des intervieweurs de la mesure dans laquelle les répondants comprennent la question posée donnent à penser que ce sont les répondants par téléphone qui sont les mieux informés. Toutefois, cela tient dans une large mesure aux différences en ce qui concerne les niveaux de scolarité.

Tableau 6-1
Indicateurs de la qualité des données selon le mode d'interview, vagues 2 à 7 de l'enquête HILDA

Indicateur	Interview sur place	Interview téléphonique	Différence probable = 0
Durée moyenne de l'interview QP (en minutes)	32,1	28,4	,000
% ayant une « excellente » compréhension des questions	69,2	77,2	,000
% ayant une « assez bonne », « faible » ou « très faible » compréhension des questions	4,4	4,9	,132
Longueur moyenne du texte (en caractères) des questions sur l'emploi	77,4	72,1	,000
Longueur moyenne du texte (en caractères) des questions sur l'industrie	22,7	21,7	,108

En deuxième lieu, pour un petit nombre de variables de résultat, nous avons cherché à déterminer s'il est possible de prouver que les réponses varient selon le mode d'interview. On pourrait s'attendre, par exemple, à ce que les réponses aux questions sur la satisfaction à l'égard de la vie soient entachées d'un biais dû à la désirabilité sociale, les répondants étant plus susceptibles de déclarer un niveau de satisfaction plus élevé en la présence d'un tiers. Cependant, comme le montre le tableau 6-2, ce serait plutôt le contraire. Les répondants interviewés sur place ont tendance à afficher des scores plus faibles. Ces résultats, de nouveau, s'expliquent en partie par les différences systématiques entre les deux sous-populations. Lorsque nous tenons compte

de l'effet de l'hétérogénéité individuelle (en utilisant un cadre à effets fixes), certaines de ces différences deviennent plus petites et non significatives. Toutefois, il n'en est pas de même de la satisfaction à l'égard de la situation financière, de sa propre santé et de la quantité de temps libre. Dans chacun de ces cas, les personnes interviewées par téléphone affichent des scores sensiblement plus élevés (entre 0,17 et 0,19 point de pourcentage), après prise en compte d'un petit nombre de facteurs observables (p. ex., âge, sexe, niveau de scolarité, revenu du ménage, état matrimonial, situation d'emploi, incapacité et diverses caractéristiques de l'interview) et de l'hétérogénéité non observée (effets fixes individuels). Il n'est pas évident de déterminer pourquoi les personnes interrogées par téléphone donnent plus de réponses positives et ce, seulement dans ces trois domaines.

Tableau 6-2
Scores moyens de satisfaction à l'égard de la vie (sur une échelle de 0 à 10) selon le mode d'interview, vagues 2 à 7 de l'enquête HILDA

Satisfaction à l'égard de la vie	Interview sur place	Interview téléphonique	Différence probable = 0	Coefficient (erreur-type) de l'interview téléphonique dans le modèle à effets fixes
Domicile	7,83	7,82	,794	,016 (,029)
Possibilités d'emploi	7,48	7,67	,000	,088 (,046)
Situation financière	6,50	6,66	,004	,191 (,031)
Sécurité personnelle	8,14	8,26	,001	-,047 (,025)
Sentiment de faire partie de la collectivité locale	6,71	6,88	,002	-,025 (,032)
Santé	7,60	7,77	,000	,171 (,025)
Quartier	7,90	7,95	,174	,027 (,027)
Quantité de temps libre	6,11	6,22	,062	,192 (,038)
Vie dans l'ensemble	7,89	7,88	,837	,008 (,021)

Il est plus difficile de procéder à des tests pour détecter les effets du mode en ce qui a trait aux variables de résultat plus objectives à partir de l'ensemble de données de l'enquête HILDA, dans une large mesure parce que de nombreuses variables sont corrélées au choix du mode d'interview. On pourrait soutenir, par exemple, que les biais dus à la désirabilité sociale pourraient amener les répondants (particulièrement les hommes) à exagérer le nombre d'heures travaillées chaque semaine. Cependant, comme nous l'avons déjà indiqué, nous constatons, comme prévu, que les personnes qui travaillent de longues heures sont plus susceptibles d'être interviewées par téléphone. Ainsi, il n'est pas possible de déterminer si les réponses aux questions au sujet des heures travaillées sont sujettes aux effets de mode en comparant simplement les moyennes des groupes. Toutefois, les données sur les heures travaillées sont également recueillies dans le QRSM, de sorte que nous pouvons utiliser la différence entre les données déclarées dans le QRSM et celles fournies au cours des interviews sur place pour déterminer la possibilité de biais de réponse attribuables au mode. Les simples résultats descriptifs sont fournis au tableau 6-3 et révèlent, contrairement aux attentes, que ce sont les personnes interviewées par téléphone qui sont les plus susceptibles d'exagérer le nombre d'heures travaillées chaque semaine. En rétrospective, cela pourrait s'expliquer comme fonction du conditionnement de panel. Toutefois, nous soulignons de nouveau que les différences sont de tailles relativement petites et significatives seulement chez les hommes.

Tableau 6-3
Heures habituelles de travail par semaine et mode d'interview, vagues 2 à 7 de l'enquête HILDA (personnes occupées)

	Hommes			Femmes			Personnes		
	Sur place	Tél.	P	Sur place	Tél.	P	Sur place	Tél.	P
QRSM	41,5	44,3	,000	30,5	33,0	,000	36,2	39,1	,000
Interview en personne	42,0	45,5	,000	30,9	33,6	,000	36,7	40,0	,000
Interview en personne moins QRSM	0,50	1,15	,033	0,47	0,57	,698	0,48	0,88	,051

7. Mot de la fin

Les concepteurs d'enquêtes par panel devraient apprécier les résultats présentés dans cet article, bien qu'ils soient fondés sur des analyses simples et sans doute brutes. Même si la répartition des répondants selon les différents modes d'enquête est manifestement non aléatoire, nous avons trouvé très peu de preuves que l'utilisation de modes multiples et mixtes a eu des

effets préjudiciables. Au contraire, la présence de deux modes de réalisation de l'enquête aurait pour effet d'améliorer les taux d'érosion du panel, bien que les preuves présentées ici soient extrêmement faibles. Nous avons également trouvé peu de preuves de variation marquée des réponses à différentes questions selon le mode d'enquête. Même s'ils sont loin d'être probants, nos résultats laissent supposer que les changements de mode sont peu susceptibles d'avoir des effets marqués sur l'uniformité longitudinale des données. La seule preuve d'effets négatifs évidents provient de l'interaction de modes mixtes et de modes multiples. Le fait de permettre aux répondants d'être interviewés par téléphone a manifestement eu un effet préjudiciable sur la proportion de questionnaires à remplir soi-même retournés.

Bibliographie

Day, N.A., Dunt, D.R. et Day, S. (1995). Maximizing response to surveys in health-program evaluation at minimum-cost using multiple methods, *Evaluation Review*, 19, pp. 436-450.

De Leeuw, E.D. (1992). *Data Quality in Mail, Telephone and Face-to-Face Surveys*, Amsterdam: Vrije University.

De Leeuw, E.D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys, *Journal of Official Statistic*s, 21, pp. 233-255.

Dillman, D.A. (2009). Some Consequences of Survey Mode Changes in Longitudinal Surveys, in P. Lynn (ed.), *Methodology of Longitudinal Surveys*, Chichester (UK): Wiley, pp. 127-140.

Voogt, R.J.J. et Saris, W.E. (2005). Mixed Mode Designs: Finding the Balance between Nonresponse Bias and Mode Effects, *Journal of Official Statistic*s, 21, pp. 367-387.

Watson, N. et Wooden, M. (2009). Identifying Factors Affecting Longitudinal Survey Response, in P. Lynn (ed.), *Methodology of Longitudinal Surveys*, Chichester (UK): Wiley, pp. 157-182.

Wooden, M. et Watson, N. (2009). The HILDA Survey and its Contribution to Economic and Social Research (so far), *The Economic Record*, 83, pp. 208-231.

La gestion d'un appariement inexact complexe dans les applications de codage et de couplage

Michael J. Wenzowski [1]

Résumé

Plus on s'emploie à acquérir indirectement des données de sources administratives et d'autres sources, plus il devient nécessaire d'identifier, voire de coupler ces enregistrements. Par exemple, les données recueillies à d'autres fins ne sont pas nécessairement codées comme il se doit; elles peuvent aussi s'avérer incomplètes sans l'exécution d'un couplage d'enregistrements. Ces deux activités relèvent du domaine de l'appariement inexact et probabiliste. Par exemple, les données disponibles n'ont peut-être pas été obtenues de manière à faciliter le codage selon la norme appropriée; de même, on nous présente rarement une clé unique pour effectuer un couplage déterministe. Ces conditions nous obligent à effectuer un appariement inexact ou « flou », ainsi qu'à utiliser une méthode probabiliste pour identifier les « bons » appariements.

Nous présentons les résultats d'une initiative menée récemment par Statistique Canada pour remanier nos systèmes généralisés de codage et de couplage d'enregistrements afin d'améliorer leur applicabilité pour un vaste éventail de problèmes de traitement et de domaines spécialisés. Ces systèmes sont habituellement adaptés à une utilisation donnée par l'équipe de la méthodologie chargée de créer l'application; en cours de production, ils sont couramment exécutés par le personnel des opérations d'enquête. Nous montrons comment nous avons accru l'utilité générale de ces progiciels en établissant des contrôles plus intuitifs sur la gestion de la complexité de leur traitement interne et comment nous avons simplifié leurs modèles d'installation, de montage et de traitement. Nous adoptons à cette fin le point de vue des utilisateurs de ces logiciels, qu'il s'agisse du personnel des technologies de l'information, de la méthodologie ou des opérations d'enquête.

[1] Michael J. Wenzowski, Statistique Canada, Canada (michael.wensowski@statcan.gc.ca)

La gestion des relations avec les répondants à l'Enquête nationale sur la santé de la population

Andrew MacKenzie et Natasha Zaletel [1]

Résumé

L'Enquête nationale sur la santé de la population (ENSP) est une enquête longitudinale qui recueille depuis 1994 des renseignements sur la santé de la population canadienne et des données socio-démographiques connexes. À l'automne 2009, l'ENSP aura recueilli huit cycles de données couvrant 15 années et se préparera à recueillir le neuvième cycle en 2010. Les taux de réponse à l'ENSP sont restés élevés au cours des huit premiers cycles de collecte de données mais, depuis quelques années, ils ont reculé comme dans le cas de la plupart des autres enquêtes sociales. Comme il s'agit d'une enquête longitudinale, il est essentiel de maintenir un échantillon de participants volontaires, d'autant plus qu'on perd des répondants parce qu'ils meurent ou qu'ils déménagent ou à cause d'autres formes de non-réponse. Cet exposé fait état des relations très positives entre l'ENSP et ses répondants et présente les méthodes utilisées pour favoriser et enrichir ces relations continues. Au fil des ans, l'ENSP a investi énormément dans les relations avec les répondants, notamment en constituant des groupes de discussion pour mettre à l'essai des lettres de présentation et des brochures ainsi que des cadeaux et des lettres de remerciement à l'intention des participants à l'enquête. Cet exposé aborde également l'équilibre recherché par l'ENSP lorsqu'il s'agit de décider de retirer des répondants de l'échantillon parce qu'ils sont impossibles à retracer ou qu'ils ont toujours refusé de participer.

[1] Andrew MacKenzie (andrew.mackenzie@statcan.gc.ca) et Natasha Zaletel, Statistique Canada

SESSION 5 – ANALYSE DES DONNÉES D'ENQUÊTES LONGITUDINALES

Étude par simulation d'approches par calage pour l'estimation de flux bruts

Marcel de Toledo Vieira et Gad Nathan [1]

Résumé

Les méthodes utilisées habituellement pour l'analyse de données longitudinales, comme celles s'appuyant sur l'application de modèles linéaires généralisés ou de modèles multiniveaux aux mesures répétées et sur l'utilisation d'équations d'estimation généralisées, sont fondées principalement sur un modèle. Nous examinons l'application de méthodes de calage aux fins de l'estimation de flux bruts à partir de données longitudinales. On peut ensuite procéder au calage sur les totaux connus des variables transversales ou des variables auxiliaires longitudinales et le choix de fonctions de distance appropriées donne une large gamme d'estimateurs fondés sur le plan et fondés sur le modèle, comme les estimateurs GREG. L'étude de simulation s'appuie sur des données de la British Household Panel Survey et vise à comparer l'efficacité du calage sur les variables transversales et celle du calage sur les variables auxiliaires longitudinales ainsi que l'efficacité des estimateurs habituels des flux bruts.

[1] Marcel de Toledo Vieira, Federal University of Juiz de Fora, Brésil (marcel.vieira@ufjf.edu.br); Gad Nathan, Hebrew University of Jerusalem, Israël.

Suivis impossibles et modélisation des périodes de chômage de l'EDTR à l'aide du modèle de régression à risques proportionnels de Cox

Dagmar Mariaca Hajducek et Jerry Lawless [1]

Résumé

Les enquêtes longitudinales recueillent des données sur l'éducation, l'emploi, la santé et d'autres processus de la vie. La durée des périodes vécues dans un état donné revêt souvent un grand intérêt, par exemple celle des périodes de chômage, de maladie ou de recours à l'assistance sociale. Malgré les nombreux progrès réalisés au cours des vingt dernières années dans l'analyse des processus de la vie, les données d'enquêtes longitudinales présentent toujours de nombreuses difficultés liées à des plans d'échantillonnage complexes, à la vérification intermittente des données au moyen d'interviews, à la fréquence des données manquantes et aux problèmes d'attrition de panel à cause de suivis impossibles. De plus, la durée des périodes successives vécues par une personne est habituellement corrélée, tout comme celle des périodes vécues par des personnes au sein d'une grappe. Cet exposé porte sur la modélisation et l'analyse de la durée des périodes vécues dans ces conditions. Nous abordons notamment l'utilisation de modèles marginaux de la durée ainsi que de modèles conditionnels aux antécédents d'une personne, le besoin de prendre en compte les suivis impossibles en fonction de l'état, la censure dépendante induite due à la corrélation de la durée des périodes vécues par une même personne, ainsi que les façons de compenser avec des données mal mesurées ou manquantes sur la durée ou sur d'autres variables. Ces enjeux sont illustrés par la modélisation et l'analyse des périodes sans travail d'après l'Enquête sur la dynamique du travail et du revenu de Statistique Canada.

[1] Dagmar Mariaca Hajducek (cdmariac@uwaterloo.ca) and Jerry Lawless (jlawless@math.uwaterloo.ca), University of Waterloo, Canada

Problèmes posés par la modélisation de la croissance latente au moyen de données à grande diffusion longitudinales

Laura M. Stapleton[1]

Résumé

Le présent manuscrit décrit certains problèmes que doit résoudre le spécialiste en recherche appliquée quand il analyse des données à grande diffusion longitudinales. Ce genre de données permettent de répondre à des questions de recherche ayant trait à la grandeur de la croissance au cours du temps, à la forme de la courbe de croissance et aux différences de croissance entre divers groupes; cependant, le spécialiste en recherche appliquée doit d'abord déterminer comment tenir compte des méthodes de collecte de données particulières utilisées par des tiers. Le présent article examine les décisions que doit prendre le chercheur dans le contexte de la modélisation de la croissance latente.

Mots-clés : analyse, modélisation, mesure.

1. Introduction

Le présent article passe en revue les défis auxquels fait face le spécialiste en recherche appliquée qui examine des données longitudinales provenant d'un échantillon probabiliste en recourant à la modélisation par équations structurelles (MES). Un aspect d'intérêt particulier sera la modélisation du même concept au cours du temps, afin d'examiner l'évolution de ce concept. L'article débute par une discussion des questions de recherche qui sont résolues au moyen de ces modèles et des données qui sont nécessaires pour entreprendre les analyses. Après avoir présenté la méthode générale de modélisation de la croissance latente (MCL), je cerne les questions pratiques en utilisant les données longitudinales recueillies selon des plans de sondage probabilistes. Certaines questions sont de nature théorique, tandis que d'autres sont de nature statistique et intéressent à la fois les spécialistes en recherche appliquée et ceux qui préparent les ensembles de données à grande diffusion.

2. Questions de recherche traitées par modélisation de la croissance latente

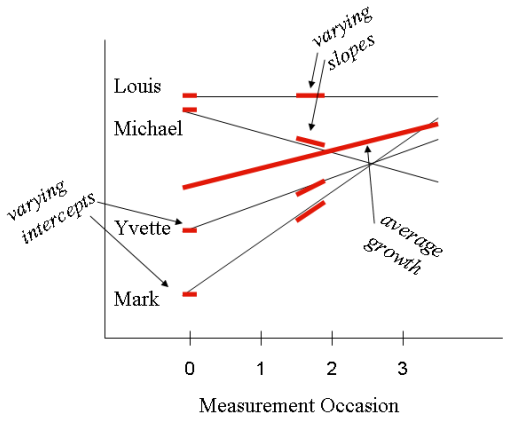
Les données longitudinales donnent la possibilité de poser des questions de recherche au sujet de l'évolution au cours du temps. Bien que l'analyse de mesures répétées existe dans le cadre de l'analyse de variance classique, les hypothèses statistiques, telles que la sphéricité et l'homogénéité de la variance de l'erreur au cours du temps, sont souvent insatisfaites lorsque l'on utilise des données empiriques. En outre, ce genre de méthodes conventionnelles sont axées sur les moyennes et n'examinent pas les différences individuelles de grandeur et de trajectoire du changement, ni les prédicteurs et les résultats de ces trajectoires individuelles. Par conséquent, la modélisation de la croissance offre une alternative aux chercheurs qui s'intéressent à l'étude de l'évolution d'un concept au cours du temps (la grandeur du changement, la forme du changement et la modération du changement). Deux approches de la modélisation de la croissance consistent à utiliser un cadre de modélisation linéaire hiérarchique (MLH) ou un cadre de modélisation de la croissance latente (MCL) (Duncan, Duncan, Strycher, Fuzhong et Alpert, 1999). Les deux approches ont de nombreuses composantes communes. Alors que la MLH offre une structure de données souples pour un petit ensemble de modèles, la MCL permet d'utiliser un plus grand ensemble de modèles, mais restreint l'analyste à un petit nombre de structures de données (Raudenbush, 2001). Les différences entre les approches de modélisation dépassent le cadre du présent article et la suite de celui-ci porte sur l'approche de la MCL.

Les questions de recherche abordées en utilisant une analyse par MCL pourraient inclure, par exemple, celle de savoir quelle est la grandeur moyenne de la croissance, au cours du temps, des compétences des élèves (voir la droite en trait gras de la figure 2-1), s'il existe ou non une variabilité des compétences de départ des élèves (voir les ordonnées à l'origine variables dans la figure 2-1) et s'il existe ou non des taux de croissance différents chez les divers élèves (voir les pentes variables

[1]Laura M. Stapleton, University of Maryland, Baltimore County, Psychology Department, 1000 Hilltop Circle, Baltimore, MD, 21230 USA

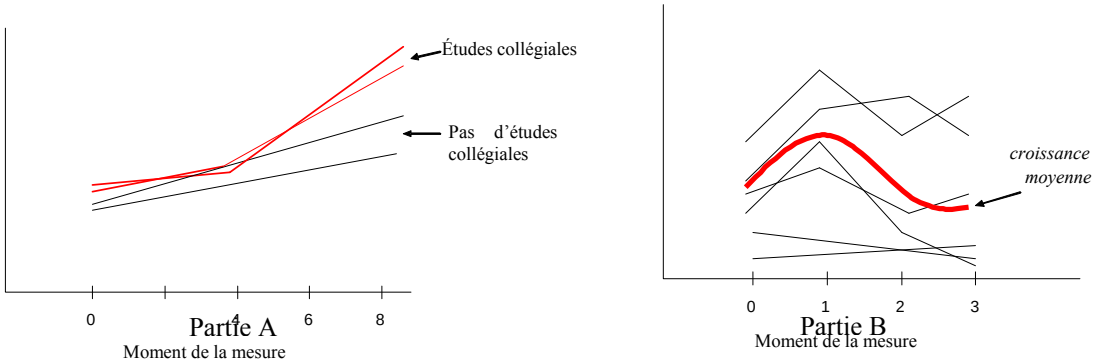
soulignées dans la figure 2-1). De surcroît, on pourrait examiner les prédicteurs et les conséquences des points de départ initiaux et des taux de croissance des élèves.

Figure 2-1
Analyse élémentaire de la croissance hypothétique



En guise d'autre exemple, la partie A de la figure 2-2 montre les niveaux de revenu mesuré chez des personnes de 18 ans, puis de 22 ans, 24 ans et 26 ans. On pourrait se demander si les courbes de croissance des personnes qui ont fait des études collégiales et de celles qui n'en n'ont pas fait sont de forme différente. Dans cet exemple, les personnes n'ayant pas fréquenté le collège ont une trajectoire de revenu linéaire, tandis que les titulaires d'un diplôme collégial ont une courbe de croissance linéaire par morceau. Enfin, comme le montre la partie B de la figure 2-2, une fonction de croissance non linéaire doit être nécessaire pour modéliser le changement au cours du temps; ici, la croissance est cubique. En outre, la forme de la courbe de croissance semble différer selon les individus; chez certains, la croissance est cubique et chez d'autres la courbe est essentiellement plane.

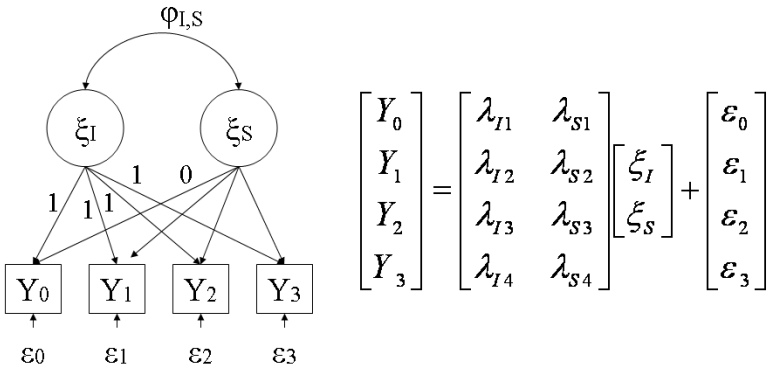
Figure 2-2
Analyses additionnelles de la croissance hypothétique



3. Détermination du modèle de croissance latente de base

Un modèle de croissance latente comportant quatre points dans le temps, qui reflète les données de croissance de base présentées à la figure 2-1) est illustré à la figure -3-1 (à noter que la structure de covariance seulement, et non la structure de la moyenne, est présentée). Habituellement, la structure des données d'un MCL pour un individu est spécifiée sous la forme $y = \lambda\xi + \varepsilon$, où **y** est un vecteur de dimension (*t* x 1) de mesures à *t* points dans le temps, λ est une matrice de dimension (*t* x 2) des saturations des facteurs ordonnée à l'origine et pente, ξ est un vecteur de dimension (2 x 1) des valeurs des ordonnées à l'origine et des pentes latentes, et ε est un vecteur de dimension (*t* x 1) des résidus particuliers à l'occasion de mesure. Ce modèle reflète l'hypothèse selon laquelle chaque observation est une fonction d'une ordonnée à l'origine (niveau) latente et d'une pente (forme) latente. On peut également émettre l'hypothèse que l'ordonnée à l'origine et la pente latentes covarient (comme l'indique la flèche à deux pointes de la figure 3-1). Dans la matrice λ, la première colonne est un vecteur de valeurs 1 pour indiquer que chaque mesure est une fonction d'une unité d'ordonnée à l'origine, et les valeurs figurant dans la deuxième colonne de la matrice λ sont définies comme des unités de pente. Si la première saturation est fixée à 0, comme le montre la figure 3-1, les saturations restantes reflètent la quantité relative de croissance ajoutée à chaque occasion de mesure supplémentaire. En supposant que les données ont été saisies à intervalles égaux (peut-être chaque année), et si les saturations restantes dans la deuxième colonne sont définies comme des entiers consécutifs (1, 2 et 3 dans l'exemple), la forme de la courbe de croissance est hypothétiquement linéaire; le concept utilisé augmente d'une unité de croissance à chaque vague. Comme la quantité du facteur de croissance latente (ξ_2) varie selon l'individu, chaque trajectoire individuelle peut avoir une pente différente (comme l'illustre la figure 2-1). Les définitions des composantes de λ dépendent du modèle hypothétique, par exemple, linéaire ou non linéaire et avec des données recueillies à des points dans le temps équidistants ou non-équidistants.

Figure 3-1
Exemple de définition d'un modèle de croissance latente pour une structure de données supposées à quatre points dans le temps



$$\begin{bmatrix} Y_0 \\ Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} \lambda_{I1} & \lambda_{S1} \\ \lambda_{I2} & \lambda_{S2} \\ \lambda_{I3} & \lambda_{S3} \\ \lambda_{I4} & \lambda_{S4} \end{bmatrix} \begin{bmatrix} \xi_I \\ \xi_S \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix}$$

L'estimation du maximum de vraisemblance peut être utilisée pour estimer les paramètres les plus susceptibles de produire les données d'échantillon observées au moyen d'une fonction d'ajustement

$$F_{ML} = \ln|\hat{\boldsymbol{\Sigma}}| + tr(\mathbf{S}\hat{\boldsymbol{\Sigma}}^{-1}) - \ln|\mathbf{S}| - p$$

où *S* représente la matrice de covariance d'échantillon, *p* est le nombre de paramètres à estimer, et $\hat{\Sigma}$ est la matrice de covariance impliquée par le modèle, définie comme étant

$$\hat{\Sigma} = \lambda\phi\lambda' + \theta$$

où ϕ est une matrice de covariance des variables latentes de dimension (2 x 2), et θ est la matrice de covariance de dimension (*t* x *t*) des résidus. Cette fonction d'ajustement est basée sur la covariance, mais peut aussi inclure le vecteur des moyennes. En outre, le maximum de vraisemblance à information complète (MVIC) donne la log-vraisemblance sommée sur l'ensemble des individus (et peut être utilisé pour traiter les données manquantes, comme il est discuté à la section suivante).

4. Questions qui se posent quand des données à grande diffusion sont utilisées pour modéliser la croissance latente

Dans cette section, je donne un aperçu général des difficultés auxquelles fait face le spécialiste en recherche appliquée quand il utilise des ensembles de données secondaires pour étudier la croissance latente. Les difficultés peuvent être groupées approximativement en deux domaines, celui de la mesure (c.-à-d. de quelle façon et quand le concept a été mesuré) et celui de l'échantillonnage (c.-à-d. comment les individus formant l'échantillon ont été identifiés). Par définition, le chercheur secondaire n'exerce aucun contrôle sur la façon et le moment où les données ont été recueillies, ni sur les sujets auprès desquels elles ont été recueillies. Il est essentiel de reconnaître explicitement ce manque de contrôle. Décider si la méthode produit une mesure adéquate des concepts d'intérêt est la première question à résoudre; le chercheur secondaire ne devrait peut-être pas entreprendre l'analyse étant donné les propriétés de la mesure. Ensuite, en supposant que la mesure est adéquate, pour décider comment généraliser les résultats en les étendant de l'échantillon à la population d'intérêt, il faudrait comprendre le plan d'échantillonnage, car les statistiques inférentielles dépendront de la méthode d'échantillonnage utilisée.

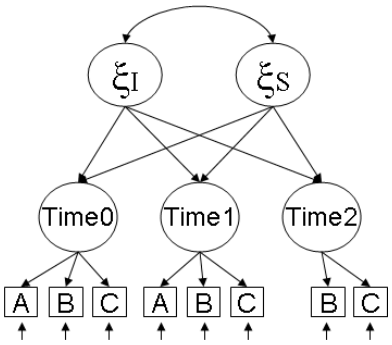
Afin de fournir un contexte à la discussion des questions que doit résoudre le chercheur secondaire, je présente ici deux ensembles de données qui ont été utilisés pour la modélisation de la croissance latente. L'Early Childhood Longitudinal Study – Kindergarten (ECLS-K) est un programme d'enquête réalisé par le National Center for Education Statistics des États-Unis. Ce programme a débuté à l'automne de 1998 avec un échantillon probabiliste d'élèves de la maternelle. Le plan d'échantillonnage comporte le tirage d'un échantillon à trois degrés de régions géographiques comme unités primaires d'échantillonnage (UPE), puis des écoles dans ces UPE comme unités secondaires d'échantillonnage (USE), puis la sélection d'élèves dans ces écoles avec divers taux de sélection, selon les caractéristiques de l'élève. Ce programme d'enquête comporte sept vagues de collecte des données, de la maternelle à la huitième année. À chaque vague, entre autres mesures, les compétences/aptitudes scolaires des enfants ont été évaluées. La National Longitudinal Study of Youth (NLSY) est réalisée par le Bureau of Labor Statistics des États-Unis et deux cohortes, celle de 1979 et celle de 1997, ont été suivies dans ce programme d'enquête. Un échantillon aréolaire a produit des ménages avec lesquels on a pris contact pour identifier les répondants âgés de 12 à 17 ans. Chaque année, ces répondants sont de nouveaux contactés et, à l'heure actuelle, avec la NLSY de 1997, les chercheurs ont à leur disposition des données couvrant une période de 11 années. Des questions ont été posées à ces adolescents concernant leur comportement, leur scolarisation, la fréquentation d'un collège, les résultats en matière d'emploi et la santé. Les ensembles de données provenant de ces deux programmes d'enquête ont souvent été utilisés pour faire la démonstration de la modélisation de la croissance en utilisant le cadre de la MES ainsi que le cadre de la MLH (p. ex., Curran et Bollen, 2001; von Hippel, Powell, Downey et Rowland, 2007). Ces démonstrations ont rarement abordé les questions liées à l'échantillonnage et peu d'entre elles ont répondu explicitement aux préoccupations concernant le processus de mesure qui est décrit plus bas.

4.2 Mesure du concept

Comme les données longitudinales ne sont pas nécessairement recueillies expressément pour la MCL, leur analyse secondaire par cette méthode peut poser plusieurs problèmes concernant la mesure du concept, dont l'uniformité avec laquelle la mesure a été faite à chaque vague, l'existence éventuelle d'effets de plafonnement ou de plancher dans la mesure et des différences de fiabilité de la mesure aux divers points dans le temps. Chacune de ces questions est discutée brièvement à la présente section.

L'une des hypothèses fondamentales de la MCL est que le même concept est mesuré à chaque occasion de mesure et, par conséquent, que la grandeur du changement dans le concept au cours du temps peut être interprétée clairement. Il se pourrait que les questionnaires soient modifiés entre deux périodes, risquant ainsi d'invalider toute comparaison au cours du temps. La définition du modèle exposée à la section 3.1 supposait la modélisation d'une seule variable manifeste au cours du temps. Pour résoudre le problème des questions qui change au fil des ans, il est possible de modéliser temporellement un concept latent, ce qui permet d'effectuer la modélisation avec des questions « légèrement » différentes à chaque période. Ce type de modélisation a été nommé modèle de courbe des facteurs (en anglais, *curve of factors model*) (McArdle et Hamagami, 2001) ou MCL de deuxième ordre (Hancock, Kuo et Lawrence, 2001). Dans ce modèle, illustré à la figure 4.2-1, les questions disponibles sont utilisées pour identifier le concept à chaque période, les valeurs des saturations respectives des facteurs étant contraintes d'être les mêmes au cours du temps (p. ex., $\lambda_{A0} = \lambda_{A1}$ et $\lambda_{B0} = \lambda_{B1} = \lambda_{B2}$). Ce modèle est assorti d'une hypothèse forte d'invariance des mesures d'une période à l'autre. Un chercheur pourrait également inclure dans ce modèle les résidus covariants de chacune des variables indicatrices manifestes respectives afin de représenter la variance hypothétique de la méthode.

Figure 4.2-1
Modèle de croissance latente à courbe des facteurs



Une deuxième préoccupation a trait à l'existence d'effets de plafonnement ou de plancher dans les évaluations. Quand un concept a été mesuré au cours du temps, il est possible que la gamme de scores qui était applicable au début de l'étude ne soit plus pertinente à un stade plus avancé de l'étude. Par exemple, dans le cas des données de l'ECLS-K, une question de recherche pourrait concerner la croissance du vocabulaire. Théoriquement, les enfants manifestent une croissance exponentielle du vocabulaire. Toutefois, les programmes d'enquête pourraient utiliser une évaluation du vocabulaire dans laquelle les enfants peuvent atteindre leur niveau le plus élevé, c'est-à-dire plafonner. Si l'on utilise ce genre d'évaluation (présentant des effets de plafonnement), la croissance paraîtra non linéaire avec un terme quadratique négatif, suggérant donc une stabilisation de la croissance du vocabulaire. L'analyste secondaire n'a d'autres recours que de reconnaître que la trajectoire ne doit pas être interprétée comme la croissance du vocabulaire, mais plutôt comme la croissance du score sur la mesure particulière

Enfin, la fiabilité d'une mesure pourrait ne pas être uniforme au cours du temps. Il est possible qu'à chaque vague successive de collecte de données, la mesure devienne plus (ou moins) fiable. Par exemple, on a constaté que les mesures autodéclarées par les enfants de moins de 12 ans sont moins fiables que celles fournies par les enfants plus âgés (de Leeuw, 2005; Rebok et coll., 2001). Donc, la variabilité des mesures pourrait être plus grande aux premières vagues de collecte des données auprès de populations d'enfants. Dans le cadre de la MCL, il est possible de composer avec cette variabilité de la fiabilité. Comme l'illustre la figure 3-1, le chercheur modélise initialement l'erreur à chaque occasion de mesure et peut spécifier que les estimations de la variance de l'erreur à chaque point dans le temps ne soient pas contraintes d'être les mêmes. La spécification de divers niveaux d'erreurs aléatoires est l'un des domaines où la MCL dans un cadre de MES est plus souple que la MLH, qui repose habituellement sur l'hypothèse d'une variance résiduelle constante aux divers points dans le temps (Rovine et Molenaar, 2001).

4.3 Moment de la collecte des données

Un ensemble général de préoccupations dont doivent tenir compte les chercheurs qui appliquent la MCL à des données secondaires concerne le moment de la collecte des données et la façon de modéliser convenablement les occasions de mesure. Quel que soit le type de modèle de croissance choisi (linéaire, linéaire par morceau, non linéaire), le chercheur doit répondre aux quatre questions qui suivent : Le choix des valeurs de saturation reflète-t-il les intervalles de temps? Les périodes de mesure concordent-elles entre les individus? Les périodes de collecte sont-elles adaptées à la courbe de croissance hypothétique? La fonction de changements appropriée est-elle la vague ou un autre concept? Ces quatre questions sont abordées à la présente section.

Dans le processus de définition du modèle de croissance latente illustré à la figure 3-1, le chercheur doit modéliser explicitement l'intervalle de temps si un modèle linéaire, linéaire par morceau ou spécifiquement non linéaire est choisi (tel qu'un modèle quadratique ou cubique) en spécifiant les valeurs de saturation des facteurs. Pour le facteur de croissance linéaire, les saturations doivent représenter l'intervalle entre les occasions de mesure et, par conséquent, il est essentiel que le chercheur comprenne cet intervalle et le modélise convenablement. Par exemple, dans le cas de l'ECLS-K, les données ont été recueillies aux périodes suivantes : automne 1998, printemps 1999, automne 1999, printemps 2000, printemps 2002, printemps 2004 et printemps 2007. Le chercheur qui suppose que la croissance est linéaire doit reconnaître que les quatre premières mesures ont eu lieu à intervalle d'environ six mois. La cinquième mesure a eu lieu après un arrêt de deux ans, la sixième, après un autre arrêt de deux ans et la septième après un arrêt de trois ans, de sorte que les saturations pourraient être égales à 0, 0,5, 1, 1,5, 3,5, 5,5 et 8,5, et les estimations seraient interprétées comme la grandeur annuelle de la croissance.

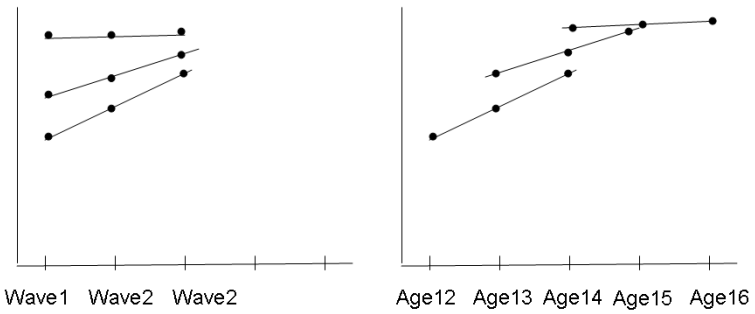
Une autre préoccupation concernant le moment de la collecte des données est celle de la concordance entre les individus. Dans le contexte de la MCL, le modèle devrait être applicable à chaque individu et donc, l'hypothèse est que chaque occasion de mesure a lieu au même moment chez tous les individus, c'est-à-dire que les mesures sont équilibrées dans le temps

(Ware ,1985). Dans le cas des données de l'ECLS-K, par contre, les mesures ne sont pas faites au même point dans le temps. Dans le cas de l'évaluation faite à l'automne à la maternelle, les évaluations pourraient avoir eu lieu n'importe quand entre septembre et décembre. De même, dans le cas de la NLSY de 1997, les chercheurs modélisent souvent les données sachant l'âge du répondant (12, 13, 14, etc.). Toutefois, au moment de la mesure, certains enfants pourraient avoir eu 12 ans et 4 mois et d'autres, 12 ans et 9 mois. McArdle et Hamagami (2001) ont résolu partiellement ce problème en utilisant des incréments d'un an et demi au lieu d'incréments d'un an. Supposer que les données sont équilibrées dans le temps quand elles ne le sont pas peut compromettre les conclusions tirées au sujet de la variabilité des trajectoires de croissance et la MLH est alors une alternative raisonnable pour la modélisation de la croissance, car elle ne comporte pas d'hypothèses d'équilibre dans le temps.

L'un des objectifs de la MCL est d'évaluer les hypothèses au sujet de la forme de la trajectoire. Cependant, les estimations du modèle dépendent du fait que les données permettent le changement à des points particuliers dans le temps. Par exemple, dans l'étude de l'évolution de l'IMC des enfants au cours du temps en utilisant les données de l'ECLS-K, von Hippel et ses collègues (2007) ont déterminé si les taux de croissance différaient durant les périodes scolaires et non scolaires (été), mais les données ponctuelles de l'ECLS-K n'étaient pas recueillies au début et à la fin de l'année scolaire; en moyenne, elles étaient recueillies à la mi-octobre et au début de mai. Les auteurs ont émis l'hypothèse d'une tendance linéaire durant les mois compris dans cette période. Comme les chercheurs secondaires n'exercent aucun contrôle sur le moment de la collecte des données, ils doivent veiller à évaluer seulement les hypothèses qui peuvent être vérifiées de manière appropriée étant donné les données dont ils disposent. Les formes ne peuvent être « vues » que si la collecte des données a eu lieu à un point d'inflexion crucial de la trajectoire.

Enfin, en plus des hypothèses au sujet de la forme des trajectoires, les chercheurs qui appliquent la MCL peuvent examiner s'il existe une variabilité entre les individus. Donc, ils devraient s'attendre dans une certaine mesure à ce que le changement puisse avoir lieu de la même façon chez les divers individus aux points dans le temps modélisés. Ce n'est pas parce que les données sont recueillies auprès d'un ensemble de personnes durant une même vague que l'on doit s'attendre à ce que toutes les personnes sont semblables à cette vague-là. Par exemple, dans le cas du plan longitudinal accéléré (ou plan à cohortes séquentielles) de la NLSY de 1997, on suit une cohorte d'une tranche d'âge. Un participant peut avoir 14 ans à la première vague, tandis qu'un autre a 12 ans à la première vague, et donc, 14 ans à la troisième. Il n'y a aucune raison de s'attendre à ce que ces individus soient les mêmes à la première vague ou qu'ils suivent des trajectoires semblables. Les données devraient être modélisées en se basant sur la fonction de changement pertinente (l'âge ou les années écoulées depuis le début de l'emploi), et non sur le moment de la collecte des données (McArdle et Hamagami, 2001). Par exemple, à la figure 4.3-1, supposons que les données ont été collectées pour les trois vagues auprès d'enfants qui avaient de 12 à 14 ans à la première vague. Le graphique de gauche serait vraisemblablement un moyen inapproprié de modéliser les données, tandis que celui de droite pourrait refléter correctement un concept de développement. De même, pour l'ECLS-K, pour examiner l'évolution de la taille au cours du temps, le semestre scolaire pourrait ne pas être pertinent et l'âge en mois devrait être une meilleure fonction du changement pour la modélisation.

Figure 4.3-1
Choix d'une référence temporelle appropriée pour la modélisation de la croissance latente



4.4 Prise en compte des données manquantes

Dans les programmes d'enquêtes longitudinales, les participants peuvent cesser de participer à une ou à plusieurs vagues, puis éventuellement revenir, ou peuvent cesser entièrement de participer à la collecte des données. Trois options pour traiter les données manquantes à une vague sont l'imputation multiple, l'estimation par la méthode du maximum de vraisemblance à information complète (MVIC) et l'utilisation de poids de panel corrigés de la non-réponse. Dans l'analyse sous imputation multiple, toutes les données disponibles pour un participant (et non simplement celles incluses dans le modèle) peuvent être utilisées pour estimer les valeurs plausibles pour les questions manquantes (habituellement on en choisit 5 ou 10). Ces tirages

aléatoires sont alors utilisés dans les analyses multiples, puis les résultats sont combinés. Cette approche suppose que les données manquent au hasard ou que le mécanisme de formation des données manquantes est « accessible » (Graham, Taylor et Cumsille, 2001). Dans l'estimation MVIC, les données disponibles sont utilisées pour trouver l'ensemble de paramètres qui maximisent la vraisemblance des données obtenues pour un individu et seuls les paramètres associés aux données disponibles sont estimés pour cette log-vraisemblance individuelle. Ensuite, les vraisemblances individuelles sont additionnées. Cette approche s'est avérée fournir des estimations robustes quand les hypothèses sont satisfaites (Enders, 2001).

Une dernière option à considérer en cas de données manquantes consiste à utiliser les poids de panel corrigés de la non-réponse qui sont fournis avec les ensembles de données. Ces poids de panel diffèrent des poids de sondage, car ils sont corrigés afin de refléter les participants qui ont cessé complètement de participer ou qui se sont arrêtés. Ces poids sont créés minutieusement afin de refléter le type de répondants qui ont décroché, en utilisant des techniques telles que l'analyse par imputation multiple et le ratissage (*raking*). Si plusieurs poids de panel sont fournis dans le fichier de données, le chercheur doit choisir celui qui est le mieux adapté à la question de recherche. Par exemple, dans le cas de l'ensemble de données de la ECLS-K, il y a eu sept périodes de collecte des données. Certains chercheurs pourraient faire la modélisation en utilisant seulement les données pour les élèves de la maternelle et de première année. D'autres pourraient vouloir utiliser les données du printemps sur les élèves de la maternelle et des 1re, 3e, 5e et 8e années, tandis que d'autres pourraient vouloir modéliser les données sur les élèves de la maternelle, de la 1re et de la 3e années. Pour chacune de ces analyses, il existe un poids de panel dont la valeur est nulle si l'individu n'a pas participé à toutes les vagues pertinentes et est non nulle autrement. Cependant, il est possible qu'un analyste souhaite ajuster un modèle aux données sur les élèves de la maternelle, et de la 1re et de la 3e années, mais veuille inclure dans l'analyse tous les participants qui ont « cessé de participer » durant la 1re année (mais qui sont finalement revenus dans l'étude à la 3e année). Dans ce cas, le chercheur ne peut pas utiliser le poids de panel et doit résoudre le problème des données manquantes par imputation multiple ou par estimation MVIC.

4.5 Probabilités de sélection inégales

Dans un contexte de MCL, ne pas tenir compte des taux différentiels de sélection donnera vraisemblablement lieu à des estimations biaisées de l'ordonnée à l'origine et de la pente, ainsi que d'autres estimations du modèle, si elles sont reliées à la probabilité de sélection. Naturellement, si les estimations pondérées diffèrent des estimations non pondérées, le chercheur doit envisager d'examiner la modération entre les strates d'échantillonnage explicites. Récemment, Asparouhov et Muthén (2005) ainsi que Rabe-Hesketh et Skrondal (2006) ont évalué une estimation du pseudo maximum de vraisemblance avec l'intégration des poids de sondage et des poids corrigés de la non-réponse dans les analyses basées sur la MES et ont constaté que l'estimation était robuste.

4.6 Estimation de la variance

Comme les données à grande diffusion ne sont habituellement pas recueillies en utilisant des échantillons aléatoires simples, des approches spéciales d'estimation des erreurs-types (appelée estimation de la variance) sont nécessaires. Dans le cas des analyses fondées sur la MES, l'interdépendance des observations affecte non seulement l'estimation de l'erreur-type, mais aussi les tests χ^2 de l'adéquation du modèle (Muthén et Satorra, 1995). Les chercheurs ont deux options pour estimer d'une manière appropriée la variance des modèles de croissance latente, à savoir les méthodes de linéarisation et les méthodes de rééchantillonnage. Pour l'option de linéarisation, les progiciels actuels utilisent une méthode d'estimation du pseudo maximum de vraisemblance telle qu'il existe un estimateur sandwich pour « rééchelonner » la matrice de covariance asymptotique des estimations, ainsi que la statistique χ^2 d'adéquation du modèle (Asparouhov et Muthén, 2005; Rabe-Hesketh et Skrondal, 2006). La plupart des logiciels de MES peuvent prendre en charge cette estimation, à condition que les indicateurs des UPE et des strates de l'échantillon soient disponibles dans l'ensemble de données. À l'heure actuelle, les méthodes de rééchantillonnage ne sont pas disponibles dans les progiciels pour la MES. Le jackknife, les répliques répétées équilibrées et le bootstrap peuvent être pris en charge et Stapleton (2008) a fourni la programmation d'une macro SAS pour exécuter ce genre d'analyse avec le logiciel Mplus et a constaté que ces méthodes sont robustes.

4.7 Observations en grappes et emboîtées

La discussion qui précède supposait que les analyses par MCL étaient entreprises au niveau individuel, que l'interprétation était limitée à l'individu et, en outre, qu'elle était applicable à tous les individus. Un autre type d'analyse, c'est-à-dire une MCL multiniveaux, permettrait de tenir compte de la nature en grappes des observations. Les questions de recherche particulières pourraient être : Les trajectoires des élèves diffèrent-elles selon l'école? Des caractéristiques particulières de l'école sont-elles associées au taux de croissance moyen des élèves dans l'école? Un aspect dont il faut se soucier dans ce genre d'analyse au moyen de données secondaires est que l'on doit disposer des poids d'échantillonnage à chaque niveau de l'analyse (Rabe-Hesketh et Skrondal, 2006). À l'heure actuelle, la plupart des ensembles de données à grande diffusion ne fournissent

que les poids d'échantillonnage au niveau de l'individu et non les poids de l'UPE ou les poids conditionnels de l'USE et de l'UFE Une autre préoccupation tient au fait que, même si l'analyse multiniveaux permet de traiter un degré de l'échantillonnage, le plan complet n'a pas été pris en compte. Par exemple, dans le cas de l'ECLS-K, si l'on entreprend une MCL multiniveaux de la croissance des élèves emboîtés dans les écoles, une partie seulement du plan d'échantillonnage a été prise en compte. Les écoles sont les USE, de sorte que la dépendance des écoles à l'intérieur des UPE demeure une préoccupation et, par conséquent, des méthodes spéciales d'estimation de la variance seraient nécessaires (Rabe-Hesketh et Skrondal, 2006).

5. Résumé

En résumé, quand on entreprend de modéliser la croissance latente au moyen de données secondaires, trois questions générales doivent être abordées. Premièrement, les questions de recherche doivent être établies à la lumière des données disponibles (en particulier, les concepts mesurés et les périodes pour lesquelles des données sont disponibles). Deuxièmement, le modèle doit être construit prudemment afin de tenir compte des caractéristiques des données (p. ex., occasions de mesure, disponibilité des données à des points dans le temps particuliers) et, si le modèle ne permet d'aborder la question de recherche, il doit être abandonné. Enfin, l'estimation des paramètres et de la variance doit tenir compte du plan d'échantillonnage, ainsi que de toute non-réponse au cours du temps.

Bibliographie

Asparouhov, T. et Muthén, B. (2005). Multivariate statistical modeling with survey data. *Proceedings of the Federal Committee on Statistical Methodology Research Conference*. http://www.fcsm.gov/05papers/Asparouhov_Muthen_IIA.pdf, retrieved September 26, 2006.

Curran, P. J. et Bollen, K. A. (2001). The best of both worlds: combining autoregressive and latent curve models. In L. M. Collins, & A. G. Sayer (Eds.) *New Methods for the Analysis of Change* (pp. 107-135). Washington, DC: American Psychological Association.

de Leeuw, E. (2005). Surveying children. In Best, S. G. (Ed.) *Polling America: An Encyclopedia of Public Opinion.* Westport, CT: Greenwood Press.

Duncan, T. E., Duncan, S. C., Strycher, L. A., Fuzhong, L. et Alpert, A. (1999). *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Applications.* Mahwah, NJ: Lawrence Erlbaum Associates.

Enders, C. K. (2006). Analyzing structural equation models with missing data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (pp. 313-342). Greenwich, CT: Information Age Publishing.

Graham, J. W., Taylor, B. J. et Cumsille, P. E. (2001). Planning missing-data designs in analysis of change. In L. M. Collins, & A. G. Sayer (Eds.) *New Methods for the Analysis of Change* (pp. 333-353). Washington, DC: American Psychological Association.

Hancock, G. R., Kuo, W.-L. et Lawrence, F. W. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling, 8,* pp. 470-489.

McArdle, J. J. et Hamagami, F. (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data. In L. M. Collins & A. G. Sayer (Eds.) *New Methods for the Analysis of Change* (pp. 137-175). Washington, DC: American Psychological Association.

Muthén, B. O. et Satorra, A. (1995). Complex sample data in structural equation modeling. In Marsden, P. V. (Ed). *Sociological Methodology*, pp. 267-316.

Rabe-Hesketh, S. et Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society, Series B,* 60, pp. 23-56.

Raudenbush, S. W. (2001). Toward a coherent framework for comparing trajectories of individual change. In L. M. Collins & A. G. Sayer (Eds.) *New Methods for the Analysis of Change* (pp. 35-64). Washington, DC: American Psychological Association.

Rovine, M. J. et Molenaar, P. C. M. (2001). A structural equations modeling approach to the general linear mixed model. In L. M. Collins & A. G. Sayer (Eds.) *New Methods for the Analysis of Change* (pp. 65-6496). Washington, DC: American Psychological Association.

Rebok, G., Riley, A., Forrest, C., Starfield, B., Green, B., Robertsoon, J. et Tambor, E. (2001). Elementary school-aged children's reports of their health: A cognitive interviewing study. *Quality of Life Research, 10,* pp. 59-70.

Stapleton, L. M. (2008). Variance estimation using replication methods in structural equation modeling with complex sample data. *Structural Equation Modeling, 15,* pp. 183-210.

von Hippel, P. T., Powell, B., Downey, D. B. et Rowland, N. J. (2007). The effect of school on overweight in childhood: Gain in body mass index during the school year and during summer vacation. *American Journal of Public Health, 97(4),* pp. 696-702.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *American Statistician, 39,* pp. 95-101.

SESSION 6 – ENJEUX MÉTHODOLOGIQUES DANS LES ENQUÊTES ÉCONOMIQUES

L'élaboration d'un prototype du programme LEED italien fondé sur des données administratives: principaux aspects méthodologiques

Colace Andrea, Congia M.Carla and Rizzi Roberta[1]

Résumé

Un premier prototype d'une base longitudinale officielle italienne de données couplées employeur-employé (LEED) a été conçu et mis en œuvre par ISTAT, principalement à partir de données administratives, et plus spécialement des déclarations de revenu soumises par les employeurs au Bureau des impôts. La LEED, qui couvre l'ensemble de la population des employeurs et des employés, est le résultat d'un processus d'intégration complexe avec d'autres sources administratives et le registre statistique italien des entreprises. Comme les entités de la LEED sont couplées longitudinalement, il est possible de suivre les travailleurs au fil du temps et de relier ce suivi à la dynamique longitudinale des entreprises, en vue de l'analyse d'aspects importants concernant le marché du travail. Dans la présente communication, les principaux aspects méthodologiques de l'élaboration de cette LEED sont abordés.

Mots-clés : données couplées employeur-employé, données administratives, données couplées longitudinales.

1. Introduction

Depuis de nombreuses années, l'Institut italien de statistique (ISTAT) exploite de façon exhaustive des données administratives à des fins statistiques. La présente communication est axée sur les progrès récents de l'utilisation de données administratives, qui vise à faire avancer la production d'ensembles de données officielles convenant à l'analyse de la dynamique du marché du travail. La création d'un ensemble de données longitudinales couplées employeur-employé (LEED), à partir principalement de fichiers administratifs, représente un grand défi pour ISTAT. Après un bref aperçu de la pertinence d'une LEED longitudinale, le contexte international et national dans lequel s'inscrit l'expérience d'ISTAT est souligné. Une description des sources administratives choisies pour l'établissement d'une LEED longitudinale officielle italienne est présentée, de même que l'infrastructure de base du premier prototype et les principaux problèmes de couplage.

2. Données couplées employeur-employé: pertinence et expériences antérieures

Les analyses de la dynamique du marché du travail qui visent à rendre compte des interactions complexes entre les employeurs et les employés au niveau micro nécessitent la disponibilité de données appropriées. La création de bases de données permettant le couplage des caractéristiques des entreprises et des travailleurs n'est pas une tâche facile, particulièrement dans un contexte longitudinal. Toutefois, la pertinence de données couplées employeur-employé pour l'analyse des politiques sur le marché du travail est largement reconnue (Abowd, 2008; Contini, 2005). Un tel ensemble de données permet aux analystes stratégiques et aux chercheurs de tenir compte simultanément des données concernant l'offre et la demande sur le marché du travail. Lorsque les entités comprises dans la LEED peuvent être couplées longitudinalement, la valeur informative de ces ensembles de données augmente énormément. Un tel ensemble de données de panel permet de suivre les mêmes employés et leur lieu de travail au fil du temps, fournit des renseignements essentiels pour examiner une vaste gamme de résultats sur le marché du travail, comme les flux de travailleurs et d'emplois, l'ancienneté dans l'emploi et le cumul d'emplois, la dynamique des salaires, la démographie et la productivité des entreprises, les effets des caractéristiques des personnes et des entreprises sur la détermination de la rémunération, la mobilité individuelle en rapport avec l'adaptation de la main-d'œuvre propre à une entreprise, etc. Ce type de base peut aussi permettre d'élargir ou d'améliorer la production de statistiques officielles et de combler des lacunes possibles au chapitre de l'information dans les études de la dynamique du marché du travail, qui existent toujours en Italie.

[1]Colace Andrea, ISTAT-Italian National Institute of Statistics, Via Tuscolana 1788, Rome, Italy, 00173 (colace@istat.it); Congia M.Carla, ISTAT- Italian National Institute of Statistics, Via Tuscolana 1788, Rome, Italy, 00173 (congia@istat.it); Rizzi Roberta, ISTAT- Italian National Institute of Statistics, Via Tuscolana 1788, Rome, Italy, 00173 (rizzi@istat.it)

Bon nombre de pays européens et d'autres pays avancés ayant une longue tradition d'analyse micro-économétrique du marché du travail ont établi des LEED adaptées à des fins statistiques. À partir des expériences internationales existantes, de nombreux types différents d'ensembles de données appariées employeur-employé ont été créés. Ils peuvent être classés selon plusieurs aspects. Tout d'abord, certains sont des ensembles de données transversaux, tandis que d'autres sont longitudinaux. La LEED transversale comprend à la fois des données sur les employeurs et les employés, mais la plupart d'entre elles concernent un point donné dans le temps. C'est pourquoi cette approche ne permet pas d'analyser réellement la dynamique du marché du travail. Par ailleurs, la LEED, qui comporte une composante longitudinale, suit les travailleurs et leurs entreprises ou lieux de travail au fil du temps et convient pour étudier les relations dynamiques entre les entreprises et les travailleurs. Le deuxième aspect qui peut être utilisé pour classer la LEED existante est sa représentativité. Certaines de ces bases de données sont fondées sur un plan d'échantillonnage axé uniquement sur les employés, tandis que d'autres représentent uniquement la population des entreprises. Il peut arriver qu'un ensemble de données appariant les employés et leurs employeurs ne soit pas conçu pour être représentatif de sections transversales et de panels de travailleurs et d'entreprises. Leurs caractéristiques de conception et leur couverture dépendent des objectifs de l'utilisateur (Abowd, 1999). Dans la présente communication, une attention spéciale est accordée aux panels représentatifs appariés de travailleurs et d'entreprises. Ils peuvent être fondés sur des enquêtes statistiques ou sur des données administratives. En général, les données d'enquête comprennent un nombre plus faible d'observations que la LEED, qui est fondée sur des données administratives, mais un nombre plus grand de covariables. La valeur de ces bases de données fondées sur des enquêtes est liée aux données détaillées qu'elles comprennent. Toutefois, de nombreux ensembles de données longitudinales appariées employeur-employé représentatifs sont fondés sur des dossiers administratifs. Parmi eux, les ensembles de données officiellement couplées qui sont élaborés à partir de données administratives deviennent une forme de LEED plus courante, les instituts statistiques nationaux étant de plus en plus sensibilisés à la nécessité de ressources de ce type et décidant d'investir dans leur production (Desai, 2008). Ces données comportent un certain nombre de faiblesses liées à l'exploitation statistique de données administratives, mais offrent la possibilité d'élaborer des ensembles de données de plus grande qualité. Ce type de base de données a été mis en œuvre dans de nombreux pays d'Europe, comme le Danemark, la Suède et la France. Parmi les expériences de premier plan à l'extérieur de l'Union européenne figurent celles de Statistics New Zealand (Padapopolous, 2004) et du U.S. Census Bureau (Abowd, 2004).

En Italie, l'enquête de l'Union européenne sur la structure des salaires représente une expérience officielle, qui est menée tous les quatre ans par ISTAT, à partir d'une base d'échantillonnage à deux degrés, qui assure uniquement la représentativité transversale des entreprises et des travailleurs. L'autre grande expérience italienne d'élaboration d'un ensemble de données couplées employeur-employé est le panel des travailleurs, qui est fondé sur des données administratives de l'Institut national de la sécurité sociale (INPS). Les données brutes, qui sont extraites directement par l'INPS de sa base de données administratives, représentent un échantillon longitudinal représentatif au niveau national des travailleurs (de 1985 à 2004), ce qui fait que la base comporte certaines lacunes au chapitre de la représentativité des entreprises. Les données sont fournies à des fins de recherche à l'Université de Turin et aux experts du Centre de recherche Laboratorio Revelli, qui ont créé le panel italien des antécédents de travail – WHIP, au ministère du Travail, pour alimenter la base de données CLAP (Campione Longitudinale degli Attivi e dei Pensionati), à l'Institut italien de développement de la formation professionnelle pour les travailleurs (ISFOL) et à certaines autres universités italiennes (Bocconi, La Sapienza).

3. Le prototype de la LEED longitudinale italienne

3.1 Sources utilisées

Pourquoi une LEED « officielle » et « fondée sur des données administratives »?

Depuis quelques années, l'Institut italien de statistique est plus sensible aux besoins des chercheurs et des analystes des politiques pour des ensembles de données appariées sur les employeurs et les employés qui sont fiables et accessibles, et a commencé à investir dans les premières activités de conception et de mise en œuvre d'un ensemble de données longitudinales couplées employeur-employé, principalement sur la base de données administratives. L'accès d'ISTAT aux microdonnées administratives est prévu dans les lois nationales. Cela signifie l'accès à une plus vaste gamme de sources de données que celles mises à la disposition des chercheurs, et permet l'élaboration d'ensembles de données de plus grande qualité, caractérisés par une homogénéité du traitement et des méthodes, une couverture plus vaste de la population et un biais moins grand. ISTAT doit en outre disposer d'un plus grand nombre de ressources pour élaborer des ensembles de données qui seront mis à la disposition des universitaires et des chercheurs.

Suivant une stratégie d'exploitation exhaustive des sources administratives à des fins statistiques depuis les années 90, ISTAT utilise des données administratives et a développé une expérience considérable du traitement et de la combinaison de ces sources. C'est dans ce contexte que s'inscrit l'utilisation des sources administratives pour l'élaboration du premier prototype officiel de LEED longitudinale. De façon plus particulière, deux expériences ont été cruciales pour la conception de la LEED qui a suivi : l'élaboration du registre des entreprises (ASIA) intégrant principalement des sources administratives et des études

de faisabilité pour l'exploitation statistique des dossiers fiscaux individuels (formulaires 770) et des dossiers individuels de la sécurité sociale (formulaires EMens).

Contenu d'information

L'élaboration de l'ASIA est le résultat de l'intégration de plusieurs sources administratives et représente le premier exemple d'un produit statistique élaboré presque exclusivement à partir de données administratives. Grâce à leur intégration, il a été possible de valider les données présentes dans les sources individuelles, en vue de corriger les erreurs, mais avant tout, de transformer les données administratives en données statistiques, en conformité avec les concepts et les définitions particulières du système statistique officiel. L'ASIA présente sur une base annuelle l'univers des entreprises industrielles et de service qui sont actives dans le pays, ainsi que leur emplacement, leur nombre d'employés, leur activité économique et leur roulement. Le registre des entreprises permet de disposer d'un cadre de référence unique pour toutes les statistiques économiques officielles. Les autres sources administratives peuvent y être couplées au moyen du numéro de code fiscal de l'entreprise, qui est disponible à la fois dans les sources fiscales et dans celles de la sécurité sociale dont il est question ci-après, ce qui leur permet d'acquérir une valeur statistique provisoire.

Le formulaire fiscal 770 comme source administrative est utilisé pour l'élaboration du prototype de la LEED et la déclaration annuelle obligatoire des impôts, des cotisations de sécurité sociale et des primes d'assurance-accidents de travail versées par les agents percepteurs (entreprises, établissements publics et privés et travailleurs autonomes). La couverture englobe les travailleurs qui, au cours de l'année précédant l'année d'imposition, ont travaillé pendant au moins une période, que ce soit comme employés ou comme travailleurs autonomes.

REGISTRE DES ENTREPRISES (ASIA)	FORMULAIRE FISCAL 770	E-MENS ANNUALISÉ
Données d'identification et données structurelles Code fiscal de l'entreprise **[n° d'identification principal de l'employeur]** Formulaire légal Secteur d'activité économique (ATECO 2002-ATECO 2007) Catégorie de roulement **Données sur l'emploi** Nombre de travailleurs Nombre d'employés Nombre de travailleurs autonomes **Données géographiques** Municipalité (description et code d'ISTAT) Province (description et code d'ISTAT) Région (description et code d'ISTAT) Macrorégion (description et code d'ISTAT)	**Données d'identification** Code fiscal de l'agent percepteur (entreprise) **[n° d'identification principal de l'employeur]** Code fiscal du travailleur **[n° d'identification du travailleur]** Date de naissance du travailleur Pays d'origine du travailleur **Données fiscales concernant le revenu** Revenu du travail assujetti aux déductions en vertu de la loi Données servant à calculer le montant d'impôt dû Nombre de jours de déductions Contributions supplémentaires à l'assurance sociale et au régime de bien-être Rémunération excédentaire **Données de la sécurité sociale concernant l'emploi et le revenu** Catégorie de bénéficiaires (employés, collaborateurs, pensionnés) Qualification professionnelle (travailleurs cols bleus, employés, etc.) (a) Durée du travail (temps plein, temps partiel) (a) Durée du contrat (ouvert, fixe, saisonnier) (a) Salaire brut Jours de rémunération (a) Semaines de rémunération (a) Mois de rémunération (détails des mois individuels) Convention collective (code) (a) Convention collective (type) (b) Niveau de régime salarial (b) Contributions versées par les travailleurs Rémunération versée aux collaborateurs Contributions payables aux collaborateurs Date de recrutement et de fin d'emploi des collaborateurs (b) Mois de rémunération versés aux collaborateurs (détails des mois individuels) (c) **Données de l'assurance concernant l'emploi** Date de recrutement et de cessation d'emploi **Données géographiques** Province où le travail est effectué (b)	**Données d'identification** Numéro du registre INPS de l'entreprise **[n° d'identification secondaire de l'employeur]** Code fiscal de l'entreprise **[n° d'identification principal de l'employeur]** Code fiscal du travailleur **[n° d'identification du travailleur]** **Données concernant l'emploi** Qualification professionnelle (travailleurs cols bleus, employés, etc.) Durée du travail (temps plein, temps partiel) Durée du contrat (ouvert, fixe, saisonnier) Type de contribution (permet d'identifier les travailleurs qui sont dans des situations particulières au chapitre des contributions) Type de travailleur (permet d'identifier les types particuliers de travailleurs) Convention collective (code) Salaire brut Jours de rémunération Semaines de rémunération Mois de rémunération (détails des mois individuels) Date de recrutement et de cessation d'emploi **Données géographiques** Municipalité où le travail est effectué principalement (code ISTAT) Province où le travail est effectué principalement (code ISTAT)

L'utilisation à des fins statistiques du formulaire fiscal 770 a commencé officiellement à ISTAT en 2006, par suite de la création d'un groupe de travail interinstitutionnel constitué de certains chercheurs de l'Institut ayant différentes expériences du traitement des sources administratives, de l'utilisation des données fiscales et de l'utilisation des variables de l'assurance sociale, etc., ainsi que de personnel spécialisé du ministère de l'Économie et des Finances et de l'Institut national de la sécurité sociale (INPS). Les activités menées par le groupe de travail visaient aussi à intégrer, à améliorer et à élargir les données officielles d'ISTAT au sujet des phénomènes économiques importants, et plus particulièrement la section de la rémunération nette et des impôts/contributions des travailleurs (Calzaroni et coll., 2008). À partir de l'année de référence 2005, le formulaire fiscal 770 a été modifié, à la suite de la modification de la législation; certaines données qu'il comprend ont donc été transférées à d'autres sources administratives. Afin de veiller à ce que les données du formulaire fiscal 770 demeurent disponibles au fil du temps, il est devenu nécessaire d'acquérir et de commencer à étudier et à traiter une nouvelle source fournie par l'Institut national de la sécurité sociale : E-Mens annualisé. Cela est le résultat du traitement effectué par l'INPS, en vue d'établir des archives comprenant des données annuelles agrégées, tirées des données des déclarations mensuelles des cotisations de sécurité sociale (formulaire E-Mens) versées par les employeurs (entreprises, établissements publics et privés) à l'INPS.

À partir du contenu original de chaque source utilisée, en vue d'élaborer le prototype de la LEED, les variables les plus pertinentes ont été sélectionnées et quelques autres variables statistiques ont été récupérées dans les données administratives. Compte tenu du nombre important de variables comprises dans la source fiscale plus particulièrement, celles présentant la plus grande pertinence pour l'élaboration de la LEED et l'analyse du marché du travail sont regroupées et résumées ci-dessus.

Chaque source est caractérisée par des variables qui peuvent être classées comme des données d'identification (dans les cas où des variables d'identification sont incluses), des données sur l'emploi et des données géographiques, le formulaire fiscal 770 et E-Mens annualisé étant aussi caractérisés par des données fiscales ou des données d'assurance concernant le revenu. Dans le tableau, la colonne concernant le formulaire fiscal 770 comprend des notes pertinentes en vue d'indiquer : les données non recueillies depuis 2005 et transférées à l'autre source (a); les données non recueillies depuis 2005 et non transférées à l'autre source (b); et les données incluses depuis 2005 (c).

3.2 Principaux aspects méthodologiques

Processus de traitement des données administratives

Un processus de traitement très complexe était nécessaire pour exploiter les données administratives utilisées. Tout d'abord, on a procédé à un examen du contenu d'information de la source et à une analyse des règles administratives, qui ont permis de déterminer les catégories générales de sujets qui doivent remplir les formulaires fiscaux ou de la sécurité sociale (univers de référence), les sujets pertinents touchés (unités d'analyse), ainsi que les catégories de revenu incluses dans le formulaire. La comparaison des définitions et des classifications comprises dans les lois fiscales, par rapport à celles des statistiques officielles, a précédé l'étape de l'intégration dans le registre des entreprises, en vue d'évaluer la couverture des univers de référence de la source, du point de vue des agents percepteurs (entreprises) et des bénéficiaires (employés).

Par la suite, il a fallu récupérer les variables statistiques des données administratives. Certaines des variables statistiques figurant dans la liste précédente sont le résultat d'opérations extrêmement laborieuses, par exemple, dans le cas des variables sur la rémunération obtenues à partir des données fiscales, qui sont déclarées de façon très détaillée dans le formulaire fiscal 770, en tenant compte de la façon dont l'impôt net est calculé en Italie, afin de déterminer la rémunération nette ainsi que la rémunération brute. Par ailleurs, certaines variables non incluses dans les sources, mais essentielles pour intégrer les archives d'entrée, ont été récupérées. C'est le cas du code fiscal de l'entreprise dans E-Mens annualisé, qui est obtenu grâce à un couplage avec le numéro du registre des entreprises de l'INPS.

Les classifications statistiques officielles ont été attribuées grâce au couplage avec le registre des entreprises ASIA, en transposant les codes administratifs, comme dans le cas des données géographiques.

On a assuré un contrôle continu des règles juridiques et administratives qui pourraient changer au fil du temps et entraîner une modification du contenu d'information des sources administratives, comme le transfert de certaines données du formulaire fiscal 770 au E-Mens annualisé.

Enfin, les formulaires multiples relatifs au même emploi ont dû être traités, et chaque source administrative a dû être validée avant son intégration.

Intégration des sources

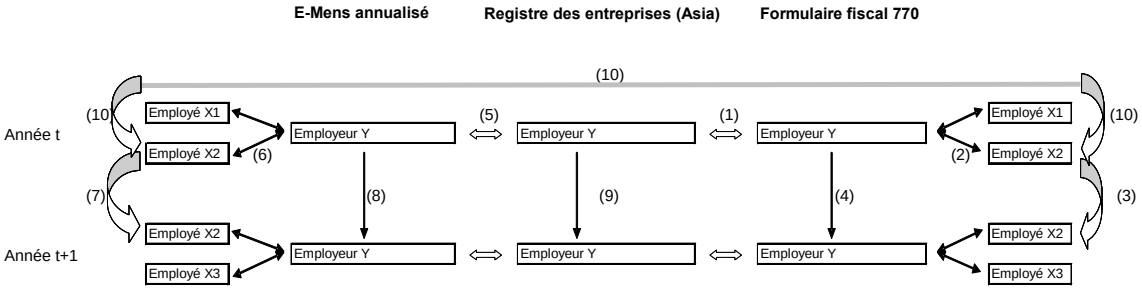
Afin d'élaborer la LEED longitudinale, il a été nécessaire de tenir compte de plusieurs aspects de l'intégration : entre les sources, entre les unités d'analyse et, enfin, en rapport avec le temps (années de référence).

1) En ce qui a trait aux sources, l'intégration a été effectuée entre celles décrites précédemment. En ce qui a trait plus particulièrement à l'intégration entre les deux sources comprenant des données sur les travailleurs (formulaire fiscal 770 et E-Mens annualisé), on a ajouté à l'opération de couplage une activité de comparaison des valeurs prises par les variables communes aux deux sources, ainsi qu'une vérification des valeurs des variables transférées par rapport aux valeurs originales.

2) En ce qui a trait aux unités d'analyse, il est nécessaire de faire une distinction entre l'ASIA et les deux autres sources. Dans les faits, l'unité d'analyse du registre des entreprises est l'entreprise, tandis que dans les deux autres systèmes d'archives, les unités d'analyse sont l'entreprise et le travailleur. La présence du code fiscal de l'entreprise dans les deux sources administratives permet de les coupler à l'ASIA et de disposer d'une base officielle des caractéristiques des entreprises auxquelles les travailleurs fournissent leur travail. Parallèlement, la disponibilité des deux codes fiscaux rend possible l'intégration des deux systèmes d'archives, ce qui permet de disposer de données d'information complètes.

3) En dernier lieu, en ce qui a trait à l'aspect temporel, on souligne que le prototype officiel italien de la LEED a trait aux années 2005 et 2006. La décision d'élaborer la LEED en fonction de la durée minimale d'une base de données longitudinales (deux ans), en dépit de l'expérience acquise sur plusieurs années de l'utilisation à des fins statistiques des sources administratives, de la gestion du registre des entreprises et de la disponibilité des données, a été dictée par des questions de compatibilité des données, par suite des modifications apportées à la législation. Le remaniement de la série chronologique de

données peut représenter un projet important en ce qui a trait au traitement des données, par suite du processus d'interprétation et de rapprochement des variables recueillies dans diverses archives administratives. Un aperçu de la complexité de l'ensemble de liens nécessaires pour gérer les sources, les unités d'analyse et les années de référence, en vue de l'élaboration du prototype de la LEED, est fourni dans le diagramme qui suit.



Dans les archives des formulaires fiscaux 770, les liens à gérer sont les suivants : le lien entre le travailleur et l'entreprise pendant une période déterminée (lien 2); le lien entre les mêmes travailleurs différentes années (lien 3), afin de suivre le cheminement professionnel de la personne; le lien entre les mêmes entreprises différentes années (lien 4), afin de suivre la dynamique des entreprises (et leur création/disparition ou fusion-acquisition subséquente).

Les mêmes liens se trouvent dans les archives d'E-Mens annualisé et comprennent : le lien entre le travailleur et l'entreprise pendant une période déterminée (lien 6); le lien entre les mêmes travailleurs différentes années (lien 7); le lien entre les mêmes entreprises différentes années (lien 8).

Les archives ASIA comprennent uniquement le lien entre les mêmes entreprises différentes années (lien 9), afin de suivre la dynamique des entreprises.

Entre les archives des formulaires fiscaux 770 et ASIA et, en parallèle, entre E-Mens annualisé et ASIA, il existe un lien entre les mêmes entreprises pour une période déterminée (lien 1 et lien 5). La combinaison des liens 1 et 5 permet de lier les trois systèmes d'archives, ainsi que de faire ressortir les entreprises qui sont communes aux trois sources, et celles qui sont communes à seulement deux des trois sources (pour les aspects liés aux questions de portée différente ou aux problèmes de sous-dénombrement de la source). Enfin, entre les archives des formulaires fiscaux 770 et celles d'E-Mens annualisé, il existe aussi un lien entre les mêmes travailleurs pour une période déterminée (lien 10).

Questions de couplage longitudinal

La gestion des liens qui précèdent n'est pas l'opération la plus simple. Plusieurs difficultés peuvent se poser, par exemple, pour le suivi des entreprises d'une année à l'autre ou d'une source à l'autre, plus particulièrement dans le cas des entreprises ayant fait l'objet d'une fusion-acquisition, ce qui pourrait mener à des distorsions de l'analyse des données démographiques de l'entreprise, du point de vue de la création et de la suppression des entreprises. Du fait que l'on ne peut pas identifier les entreprises qui subissent de tels processus – qui sont, soit dit en passant, aussi relativement répandus en Italie – on risque de déterminer certaines unités comme distinctes, alors qu'elles sont dans les faits caractérisées par une continuité statistique (à condition qu'elles soient économiquement et structurellement stables, même si elles ont des formes juridiques différentes) et vice versa, avec des effets clairement négatifs sur la détermination des flux d'emplois réels. De façon plus particulière, dans le cas des grandes entreprises, les conséquences du point de vue des flux d'emplois erronés seraient particulièrement importantes. Un autre problème moins important, mais non négligeable, est celui de la détermination correcte des codes fiscaux des travailleurs. Un code fiscal manquant ou incorrect peut en fait avoir des conséquences importantes sur le processus de reconstitution de la carrière d'une personne.

4. Observations finales

La présente communication montre que, ces dernières années, certains progrès ont été réalisés pour la production d'une LEED longitudinale pour l'analyse de la dynamique du marché du travail par l'Institut italien de statistique. L'utilisation de données administratives pour concevoir cette LEED officielle a constitué un choix naturel, compte tenu de l'expérience d'ISTAT de l'exploitation des fichiers administratifs à des fins statistiques. Beaucoup de travail a été effectué pour surmonter un certain nombre de faiblesses liées à la nature administrative des données. Cela a nécessité une analyse très approfondie du contenu

d'information des sources utilisées, ainsi que l'étude et le contrôle des changements dans les règles administratives qui ont entraîné certaines ruptures au fil du temps dans les données transférées à une autre source administrative. La récupération des variables statistiques a constitué une tâche difficile et laborieuse, nécessitant une comparaison précise des définitions et des classifications prévues dans les lois fiscales et de celles comprises dans les statistiques officielles. L'intégration des données a été effectuée grâce au couplage d'enregistrements d'unité entre les sources et au fil du temps. Le nombre de liens différents (10) donne une idée de la complexité de la procédure de couplage des enregistrements, du fait aussi des travaux supplémentaires nécessaires pour harmoniser différents identificateurs d'entreprises administratifs dans les deux sources administratives utilisées. Le couplage longitudinal des données sur les employeurs et les employés a fait l'objet d'un essai, et les premiers résultats figurent dans une autre communication des auteurs (Colace, Congia et Rizzi, 2009). Les analyses très préliminaires de la durée de l'emploi montrent le potentiel longitudinal de cette LEED, et les résultats confirment un certain nombre d'éléments de base et de faits bien connus concernant la dynamique du marché du travail. La qualité longitudinale de la LEED doit être évaluée davantage en raison de certains enjeux liés au suivi des travailleurs et des entreprises d'une année à l'autre, par suite des processus de fusion ou de scission ou de l'identification correcte des personnes cumulant des emplois, en vue d'éviter le dénombrement en double. D'autres efforts doivent être déployés pour que cette LEED officielle longitudinale soit fiable et accessible pour les chercheurs, tout en assurant la confidentialité des données. Toutefois, pour accomplir cette tâche ambitieuse, ISTAT doit investir davantage dans les ressources humaines et les TI.

Bibliographie

Abowd, J. M., Kramarz, F. et Woodcock, S. (2008). Econometric Analyses of Linked Employer–Employee Data, in Matyas, Sevestre (eds.) *The Econometrics of Panel Data*, Berlin Heidelberg: Springer-Verlag, pp. 727-760.

Abowd, J. M. et Kramarz, F. (1999). The Analysis of Labor Markets using Matched Employer-Employee Data, in O. Ashenfelter & D. Card (eds.) *Handbook of Labor Economics*, edition 1, Volume 3, Number 3, Elsevier.

Abowd, J. M., Haltiwanger, J. et Lane, J. (2004). Integrated Longitudinal Employer-Employee Data for the United States, *American Economic Review*, Volume 94, Issue 2, pp. 224-229.

Calzaroni, M., Congia, M.C., Montebugnoli, M.E., Rizzi, R. et Tronti, L. (2008). Net wages and tax/contribution wedge of employees from administrative data, paper presented at the 29th Annual Conference of the International Working Party on Labour Market Segmentation, Porto, Portugal.

Colace, A., Congia, M.C. et Rizz,i R. (2009). A Longitudinal Analysis on Italian Employees using Administrative Sources, paper presented at the 30th Annual Conference of the International Working Party on Labour Market Segmentation, Tampere, Finland.

Contini, B., et Trivellato, U. (2005). *Eppur si muove. Dinamiche e persistenze nel mercato del lavoro italiano*, Bologna: Il Mulino.

Desai, T. (2008). A Guide to Linked Employer-Employee Data Sources in the EU and Beyond: 1st Edition in final report of the Project for the European Commission DG Employment "European Labour Market Analysis using Firm-level Panel Data and linked Employer-Employee Data".

Padapopolous, T. (2004). Linked- Employer-Employee Data (LEED), in *Labour Market Statistics 2003*, Statistics New Zealand.

Les plans de sondage des enquêtes sur les prix sont-ils suffisamment robustes au vieillissement des facteurs de pondération? Étude par simulation

Zdenek Patak et Daniele Toninelli[1]

Résumé

De plus en plus d'importance a été accordée à l'étude des prix ces dernières années, et plusieurs organismes statistiques nationaux ont commencé à mettre sur pied des projets reposant sur des études longitudinales pour mesurer les mouvements des prix des produits et services. De nombreuses questions méthodologiques font l'objet de travaux de recherche intensifs en vue d'accroître la qualité de tout le processus de production des indices. L'objectif de la présente étude est de contribuer à l'amélioration de la qualité de la collecte des données en étudiant l'évolution temporelle des données d'enquête. Partant d'une population simulée, nous comparons un certain nombre de méthodes de sélection de l'échantillon utilisées fréquemment en soulignant leur efficacité relative et en évaluant comment et à quel point le changement de la mesure de taille au cours du temps affecte les estimations et le biais des résultats.

Mots clés : indice des prix, poids d'échantillonnage, étude par simulation, méthodes d'échantillonnage, biais d'indice.

1. Introduction

La publication du rapport Boskin (Boskin et coll., 1996) sur l'état des indices des prix aux États-Unis a suscité un vif débat sur les moyens d'améliorer leur qualité. L'étude de l'applicabilité des méthodes d'enquête-entreprise classiques à un domaine longtemps dominé par le jugement et le savoir des spécialistes du domaine a progressé à grands pas. Aux États-Unis, en particulier, les processus de couverture, de collecte et d'échantillonnage ont fait l'objet d'innovations importantes. Étant donné l'importance croissante du secteur des services, Statistique Canada a entrepris d'élaborer un nouvel ensemble d'indices des prix à la production pour les services dans lequel peuvent être intégrées nombre de ces innovations. Statistique Canada profite aussi de cette occasion pour étudier des questions telles que l'effet du plan d'échantillonnage sur la précision d'un indice et l'effet de l'utilisation prolongée de poids qui vieillissent sur la fiabilité. En nous appuyant sur une étude par simulation, nous comparons dans la présente étude un certain nombre de stratégies d'échantillonnage probabiliste, telles que l'échantillonnage aléatoire simple sans remise (EASSR), et plusieurs méthodes d'échantillonnage avec probabilité proportionnelle à la taille (PPT) utilisées couramment en pratique dans le contexte de l'estimation d'un indice des prix. L'étude a été entreprise en vue d'évaluer la robustesse de divers plans d'échantillonnage au vieillissement des échantillons et des poids appliqués au panier de biens et de services (un ensemble de biens et services dont les prix sont suivis au cours du temps) en ce qui concerne le biais et l'erreur quadratique moyenne des indices estimés. La population cible de l'étude par simulation a été générée en se fondant sur des paramètres calculés d'après les données de l'enquête sur les prix de l'Indice des prix à la production pour les services de gros au Canada.

1.2 Calcul de l'indice des prix

La première étape en vue d'obtenir l'indice final est le calcul de l'*indice élémentaire* au niveau de l'établissement (niveau d'agrégation le plus faible). Soit p_{hik}^{t} la marge bénéficiaire observée au temps t pour le produit k (k = 1, …, m_{hi}) de l'établissement i appartenant au groupe h, où m_{hi} est le nombre de produits observés pour l'établissement i (m_{hi} = 1, 2, 3). L'indice élémentaire $\left(P_{hi}^{t/t-1}\right)$ entre les temps t et t-1 est calculé comme étant la moyenne géométrique des ratios des marges bénéficiaires,

$$P_{hi}^{t/t-1} = \left[\prod_{k=1}^{m_{hi}} \left(\frac{p_{hik}^{t}}{p_{hik}^{t-1}}\right)\right]^{1/m_{hi}} . \qquad 1.2\text{-}1$$

[1]Zdenek Patak, Statistique Canada, 100 Promenade Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6 (Zdenek.Patak@statcan.gc.ca); Daniele Toninelli, University of Bergamo, via dei Caniana 2, Bergamo, Italy, 24127 (daniele.toninelli@unibg.it).

Cet indice est également appelé *indice de Jevons*[2]. L'*indice agrégé* au niveau de la strate (strate *h*) est une moyenne pondérée des indices élémentaires. On le calcule en se servant de la formule de Laspeyres[3],

$$\hat{I}_h^{t/t-1} = \frac{\sum_{i=1}^{n_h}\left[z_{hi}^{t-1}\cdot w_{hi}^{t-1}\cdot P_{hi}^{(t/t-1)}\right]}{\sum_{i=1}^{n_h}\left(z_{hi}^{t-1}\cdot w_{hi}^{t-1}\right)}, \qquad 1.2\text{-}2$$

où *h* est la strate, n_h est le nombre d'unités dans la strate *h*, z_{hi}^{t-1} est le *poids économique* de l'unité *i* (appartenant à la strate *h*) au temps *t*-1, et w_{hi}^{t-1} est le *poids d'échantillonnage* de l'unité *i* au temps *t*-1. De manière plus générale, pour une année *t*, le poids économique z_{hi}^{t} d'un établissement *i* appartenant au groupe d'unités *h* est calculé sous forme du ratio du revenu de l'établissement en question $\left(R_{hi}^{t}\right)$ au revenu total de tous les établissements appartenant au même groupe $\left(R_h^t\right)$: $z_{hi}^{t} = R_{hi}^{t}/R_h^t$. Le poids d'échantillonnage w_{hi}^{t} d'une unité *i* pour l'année *t* est calculé comme étant l'inverse de la probabilité de sélection, $w_{hi}^{t} = 1/\pi_{hi}^{t}$. La probabilité de sélection $\left(\pi_{hi}^{t}\right)$ de l'unité *i* est la probabilité d'être incluse dans l'échantillon sélectionné; elle varie selon la méthode de sélection de l'échantillon. Au paragraphe suivant (1.3), nous allons vérifier si les poids d'échantillonnage et les poids économiques varient en fonction du temps. Nous essaierons également de répondre à la question de savoir à quelle fréquence les poids doivent être mis à jour pour que la valeur de l'indice estimé soit aussi proche que possible de la valeur de l'indice réel.

1.3 Évolution des poids au cours du temps

Comme les poids d'échantillonnage ainsi que les poids économiques figurent dans la formule de l'indice agrégée (1.2-2), ils peuvent les uns et les autres influer sur les estimations. En fait, pour calculer un indice des prix, on doit décider du système de pondération à utiliser et cette décision peut avoir une incidence sur la qualité du résultat final. Le meilleur choix pour calculer un indice des prix pour une année particulière consisterait à utiliser les poids pour l'année en question, ce qui fournirait ce que nous appelons un « indice réel ». Malheureusement, cela n'est pas toujours possible à cause du délai habituel d'obtention des poids réels. Donc, on utilise souvent des poids qui ne sont pas actuels pour calculer un indice des prix. En outre, il est important de savoir si les poids varient au cours du temps et de quelle façon, et quel est l'effet de cette évolution sur le biais et la variance de l'indice. Au présent paragraphe, nous examinons les premiers aspects, tandis que nous discutons de l'effet de l'évolution des poids sur l'indice au paragraphe 3.1.

Une analyse des poids d'échantillonnage (w_i) et des poids économiques (z_i) pour la période allant de 2004 à 2009 a révélé qu'au cours des trois premières années, les poids d'échantillonnage sont restés relativement constants; les variations en pourcentage sont demeurées de l'ordre de 10 %. La variation d'une année à l'autre a commencé à évoluer appréciablement après quatre ans, se chiffrant à +73,6 % de 2007 à 2008 et à +46,1 % de 2008 à 2009. Si la première variation pouvait être expliquée par un changement de structure du Registre des entreprises (RE), cette raison ne peut pas être avancée pour expliquer la variation de 2008 à 2009. La variation annuelle moyenne des poids d'échantillonnage était de +15,2 %. En outre, les variations des poids d'échantillonnage dans le court terme ne semblaient pas suivre une trajectoire prévisible; la tendance était négative de 2004 à 2005 et de 2006 à 2007, mais positive dans les autres cas. Si nous considérons les variations cumulatives, un accroissement de 145,9% a eu lieu au cours des cinq années débutant en 2005, comparativement à 2004; la variation cumulative moyenne au cours de la même période était de 29,2 %. Cette variation importante des poids d'échantillonnage peut naturellement avoir un effet prononcé sur la qualité des estimations si nous utilisons des poids extrêmement dépassés.

Les poids économiques présentent un profil comparable au cours des six mêmes années (2004 à 2009), mais les sommets des variations sont plus faibles que pour les poids d'échantillonnage; la variation d'une année à l'autre maximale était de 39,8 % entre 2007 et 2008 (l'effet du passage au nouveau RE est de nouveau souligné). La variation annuelle moyenne était plus importante (20,1 %) que celle observée pour les poids d'échantillonnage.

Les données montrent que les poids d'échantillonnage varient davantage dans le long terme, tandis que les poids économiques varient davantage sur une base annuelle. Cela dit, même si la variation la plus importante (2007 à 2008) est justifiée en partie par un événement externe (qui n'explique pas entièrement la variation de 2008 à 2009), nous pouvons conclure que ces variations peuvent avoir une incidence considérable sur la précision des estimations finales. Leurs effets peuvent aussi se

---

[2] Voir le *Producer Price Index Manual* (2004) et le *Document de référence de l'Indice des pris à la consommation* (1995).

[3] Pour d'autres renseignements sur l'indice de Laspeyres, consulter le *Producer Price Index Manual* (2004), le *Document de référence de l'Indice des prix à la consommation* (1995), et Laspeyres (1884). Pour obtenir plus de renseignements au sujet de l'utilisation de l'indice de Laspeyres dans le contexte du projet de l'IPPS, voir aussi Patak et Lothian (2007).

renforcer les uns les autres, étant donné que tant les poids d'échantillonnage que les poids économiques figurent dans la formule de l'indice. Au paragraphe suivant, nous présentons la méthodologie utilisée pour évaluer l'effet de la variation des poids (c.-à-d. du vieillissement des poids) sur les estimations de l'indice, en ce qui concerne le biais ainsi que la variance.

2. Méthodologie

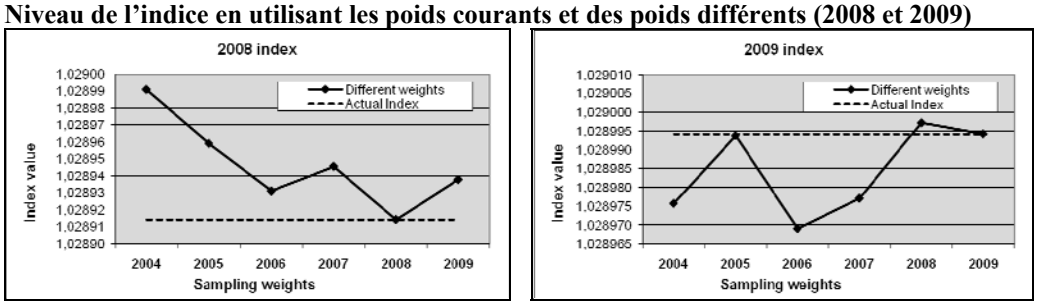
Notre étude est fondée sur une population simulée de prix relatifs. Ils ont été générés au moyen de paramètres calculés en utilisant les prix relatifs observés durant les vagues de 2006 à 2008 de l'Enquête sur le commerce de gros de Statistique Canada. Les prix relatifs simulés ont ensuite été associés à la population d'établissements de la base de sondage réelle utilisée pour sélectionner les échantillons de l'Enquête sur le commerce de gros. Pour chaque unité, les données structurelles et financières ont été extraites du RE. Après avoir simulé la population de la base de sondage, nous avons sélectionné des échantillons en utilisant trois plans de sondage, à savoir l'*échantillonnage aléatoire simple* (EAS), l'*échantillonnage avec probabilité proportionnelle à la taille* (PPT) et l'*échantillonnage de Poisson séquentiel* (EPS). Le tirage des échantillons a été effectué selon une méthode de type bootstrap, en sélectionnant un nombre maximal de 5 000 échantillons contenant chacun 1 000 unités, afin d'obtenir la convergence de la variance de l'indice. Pour chaque échantillon sélectionné, nous avons calculé diverses versions de l'indice, en faisant varier les poids d'échantillonnage (c.-à-d. en utilisant les poids d'échantillonnage pour toutes les variables du RE pour toutes les années d'intérêt). L'indice réel a été calculé pour chaque année de la population simulée (de 2004 à 2009) en utilisant les poids de la période courante; autrement dit, pour une année x (x = 2004, 2005, …, 2009), le système de poids de la même année (x) a été utilisé pour calculer l'indice. En outre, nous avons calculé d'autres *indices simulés* pour chaque année, en utilisant les poids provenant des autres années : par exemple, pour calculer un indice simulé d'une année x, nous avons utilisé les poids d'une année y (y = 2004, 2005, …, 2009, $y \neq x$). Pour chaque version de l'indice simulé, nous avons également calculé la variance et le biais obtenu en comparant l'indice simulé proprement dit à sa valeur réelle (c.-à-d. l'estimation obtenue en utilisant les poids courants; autrement dit, les poids appartenant à la même année de l'indice). Donc, ce biais résulte de l'utilisation de poids non à jour. De cette façon, pour l'indice de chaque année, nous pouvons calculer à la fois le biais des estimations et la variance que nous obtenons en utilisant les poids vieillissants, ce qui nous donne l'évaluation de la précision des estimations. De surcroît, nous pouvons comparer les résultats obtenus pour divers scénarios de sélection de l'échantillon et mesurer l'effet relatif sur le biais de l'indice, ainsi que leur robustesse. La partie suivante de l'article (paragraphe 3.1) portera, en particulier, sur l'étude de l'effet qu'ont les poids d'échantillonnage (notre facteur expérimental) sur les estimations de l'indice. Au paragraphe 3.2, nous discuterons de certains résultats (en ce qui concerne le biais et la variabilité) produits par les méthodes d'échantillonnage mises en œuvre. Vient ensuite l'évaluation comparative de ces méthodes (paragraphe 3.3).

3. Certains résultats

3.1 Niveau et variance de l'indice

Au présent paragraphe, nous évaluons l'effet des poids d'échantillonnage sur le niveau des indices des prix. En utilisant l'entièreté de l'ensemble de données simulé servant de base de sondage, nous avons calculé, pour chaque année (de 2004 à 2009), l'indice réel et les indices simulés. Le graphique 3.1-1/2 donne les résultats pour l'indice de 2008 (3.1-1, à gauche) et pour l'indice de 2009 (3.1-2, à droite). Dans les deux graphiques, la ligne discontinue représente l'indice réel (c.-à-d. l'indice calculé en utilisant les poids d'échantillonnage courants), tandis que la ligne noire montre les estimations des indices simulés (calculés en utilisant des poids d'échantillonnage non courants).

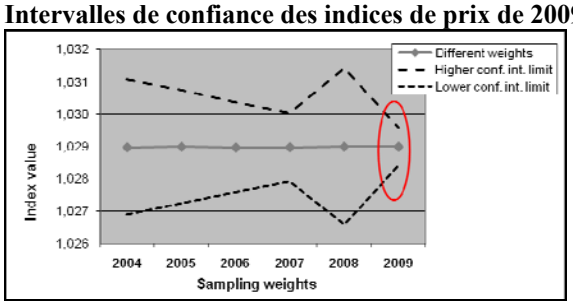
Graphique 3.1-1/2
Niveau de l'indice en utilisant les poids courants et des poids différents (2008 et 2009)



La situation la plus commune (observée en 2005, 2006 et 2007) est semblable à celle présentée pour l'indice de 2008 (graphique 3.1-1) : plus la date des poids utilisés s'écarte de l'année courante, plus le biais de l'indice simulé augmente,

comparativement à la valeur réelle (indice de 2008 calculé avec les poids de 2008). Cependant, parfois (p. ex. en 2004 et en 2009), nous pouvons également obtenir une situation semblable à celle soulignée dans le graphique de 2009 (3.1-2) : même si nous utilisons des poids vieillis (c.-à-d. des poids de 2005), nous pouvons obtenir des estimations s'approchant de la valeur de l'indice réel. Le graphique 3.1-3 montre les intervalles de confiance construits autour des valeurs d'indice de 2009, obtenus en utilisant différents poids d'échantillonnage. La situation dépeinte est commune à la majorité des cas (sauf 2008) : quand nous utilisons les poids courants, l'indice calculé est plus précis (autrement dit, l'intervalle de confiance est plus étroit).

Graphique 3.1-3
Intervalles de confiance des indices de prix de 2009

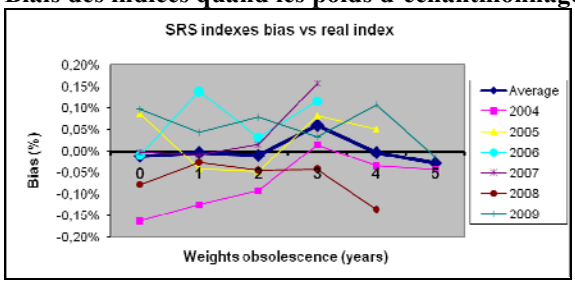


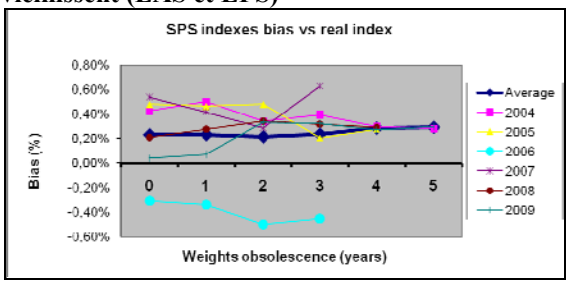
À la lumière de tout ce qui précède, il est possible de conclure que nous n'obtenons pas toujours des estimations extrêmement biaisées quand nous n'utilisons pas les poids courants, mais ces estimations sont habituellement moins précises que celles obtenues en utilisant les poids mis à jour. Maintenant que nous avons analysé ce qui se passe quand l'échantillon (et les poids d'échantillonnage) deviennent dépassés, au paragraphe 3.2 qui suit, nous étudions l'effet de la méthode d'échantillonnage sur la précision de l'indice.

3.2 Biais et écart-type des estimations

Les résultats présentés ici ont été obtenus en exécutant un nombre maximal de 500 itérations bootstrap pour sélectionner des échantillons de 1 000 unités à partir de la base de sondage simulée. Les unités ont été sélectionnées selon les méthodes suivantes : échantillonnage aléatoire simple (EAS), échantillonnage avec probabilité proportionnelle à la taille (PPT) et échantillonnage de Poisson séquentielle (EPS). Étant donné que les résultats obtenus par la méthode PPT sont semblables à ceux obtenus par la méthode EPS, seule cette dernière sera discutée dans le présent article. Les graphiques 3.2-1/2 donnent le biais obtenu dans le cas de l'EAS (graphique 3.2-1, à gauche) et de l'EPS (graphique 3.2-2, à droite). En particulier, nous montrons le comportement de l'indice à mesure que vieillissent les poids d'échantillonnage : la distance (de 0 à 5 ans) des poids utilisés par rapport à l'année sur laquelle porte l'analyse est représentée sur l'axe des *x* (donc, « 0 » signifie que nous utilisons les poids courants).

Graphique 3.2-1/2
Biais des indices quand les poids d'échantillonnage vieillissent (EAS et EPS)





Dans le graphique 3.2-1, la ligne épaisse représente la trajectoire moyenne de tous les indices à mesure que les poids d'échantillonnage deviennent dépassés : le biais (et sa variance) semble augmenter à mesure que les poids utilisés s'éloignent de l'année courante. Cependant, nous constatons que la situation moyenne reflète la compensation de trajectoires entièrement différentes représentées par les courbes plus fines (se rapportant chacune à un indice distinct). En fait, si nous examinons les détails, nous voyons que pour certains indices (p. ex., 2007), quand nous utilisons des poids plus anciens, nous obtenons un biais plus important; par contre, pour d'autres (p. ex., 2004), le biais diminue à mesure que les poids utilisés deviennent plus anciens. Si nous observons l'indice de 2008, nous constatons que les plus mauvais résultats sont obtenus en utilisant des poids qui datent de 4 ans, mais nous voyons aussi que le biais de l'indice est réduit si nous utilisons des poids qui datent de 1, 2 ou 3 ans, au lieu des poids courants. Au contraire, la valeur la plus précise pour l'indice de 2009 est obtenue en utilisant des poids vieux de 5 ans. Le graphique 3.2-2 montre les résultats obtenus par échantillonnage de Poisson séquentiel. À un niveau moyen

(courbe épaisse), le biais est plus important que celui obtenu par échantillonnage aléatoire simple (il convient de noter les échelles différentes des deux graphiques). Cela tient au fait que les unités sont sélectionnées avec une probabilité proportionnelle à la taille et que le système de poids donne plus d'importance aux unités les plus grandes qui sont habituellement caractérisées par des valeurs plus élevées de l'indice (Toninelli, 2008). Une constatation commune aux deux méthodes d'échantillonnage est que nous obtenons une trajectoire moyenne plus biaisée quand nous utilisons des poids plus anciens. De nouveau, cette courbe moyenne résulte de la neutralisation des trajectoires entièrement différentes des divers indices. Pour certains d'entre eux (p. ex., 2006, 2008 et 2009), le biais est extrêmement réduit en utilisant les poids courants, tandis que pour d'autres (p. ex., 2004 et 2005), le biais augmente quand on utilise d'anciens poids; d'autres indices (comme celui de 2007) ont une trajectoire plus irrégulière. Ces résultats montrent clairement qu'en utilisant différents plans d'échantillonnage, nous obtenons des trajectoires très irrégulières des indices qui n'ont souvent pas grand-chose en commun les unes avec les autres. Pour cette raison, il est impossible de définir un effet général précis du vieillissement des facteurs de pondération sur les estimations de l'indice. Cela tient vraisemblablement au fait que chaque indice prend en considération un système évolutif de poids qui se rapporte à divers secteurs économiques; chacun d'eux peut influencer, de manière plus ou moins prononcée, la structure générale et la distribution des poids, d'où les estimations de l'indice et leur biais. Au paragraphe 3.3, nous comparons les méthodes d'échantillonnage en examinant le biais et l'écart-type des estimations.

3.3 Comparaison des plans d'échantillonnage

Dans l'analyse qui suit, nous résumons, par année, l'effet de l'utilisation des poids courants plutôt que des poids plus anciens. Selon une méthode de sélection de type bootstrap, nous avons tiré 5 000 échantillons de 1 000 unités chacun. L'évaluation des estimations, pour chaque année considérée, porte sur le biais et l'écart-type qui nous donnent une idée de la précision de ces estimations. Pour obtenir le biais, nous comparons l'estimation simulée calculée de l'indice (moyenne des estimations obtenue pour les 5 000 échantillons sélectionnés) à l'indice réel correspondant. L'écart-type est calculé pour chaque indice obtenu au moyen des 5 000 échantillons sélectionnés par la méthode bootstrap. La moyenne de ces valeurs d'échantillon est ensuite comparée à l'écart-type de l'indice réel. En nous fondant sur ces critères, nous avons évalué, pour chaque plan d'échantillonnage, la réduction en pourcentage du biais et de l'écart-type obtenue en utilisant les poids courants (tableau 3.3-1/2). Dans les tableaux, les cases (année/plan d'échantillonnage) où les poids mis à jour donnent de meilleurs résultats sont mises en relief en gris. La réduction en pourcentage du biais en utilisant les poids mis à jour est présentée dans le tableau 3.3-1 (à gauche). L'EPS donne de meilleurs résultats que les autres méthodes d'échantillonnage, quand on utilise des poids d'échantillonnage mis à jour, dans deux cas sur six (2005 : -27,3 %; 2008 : -39,7 %); dans un troisième cas (2007 : -0,8 %), en utilisant les poids à jour plutôt que non mis à jour, nous obtenons à peu près le même niveau de biais, tandis que dans les autres cas (2004, 2006 et 2009), nous obtenons un biais plus important (plus de +47 %) en utilisant les poids mis à jour; les constatations sont les mêmes en ce qui concerne le niveau moyen (moyenne de six ans : +28,4 %). L'échantillonnage PPT et l'EPS donnent de meilleurs résultats en utilisant les poids d'échantillonnage mis à jour pour les mêmes années (2006, 2008 et 2009), c'est-à-dire dans trois cas sur six. En outre, en moyenne, le biais est réduit, respectivement de 10,7 % et de 6,2 %, comparativement au biais obtenu en utilisant des poids non mis à jour.

Tableau 3.3-1/2
Estimation avec des poids à jour c. non mis à jour (réduction en pourcentage du biais et de l'écart-type)

	BIAIS		
Indice (année)	EAS	PPT	EPS
2004	56,4	5,8	17,3
2005	-39,7	17,7	27,8
2006	47,5	-43,1	-29,1
2007	-0,8	29,5	30,9
2008	-27,3	-32,6	-28,6
2009	76,3	-65,3	-83,2
Moyenne	*28,4*	*-10,7*	*-6,2*

	ÉCART-TYPE		
Indice (année)	EAS	PPT	EPS
2004	0,2	-10,8	-15,7
2005	-3,0	-25,1	-25,9
2006	-0,5	-12,3	-17,9
2007	5,1	-14,8	-16,1
2008	-8,0	-29,3	-44,0
2009	1,5	43,8	49,1
Moyenne	*0,1*	*-12,6*	*-14,2*

La réduction en pourcentage de l'écart-type (comparativement à celui de l'indice réel) en utilisant les poids d'échantillonnage mis à jour plutôt que les poids non mis à jour est présentée au tableau 3.3-2 (à droite). Dans le cas de l'EAS, nous n'observons habituellement pas de résultats considérablement meilleurs quand nous utilisons les poids mis à jour : en moyenne, nous obtenons une variabilité qui est presque la même. En fait, la différence entre l'écart-type obtenu avec des poids mis à jour plutôt que non mis à jour est de 0,1 % seulement. De toute façon, pour les six indices considérés (2004 à 2009), nous n'obtenons un écart-type légèrement plus faible en utilisant les poids mis à jour que dans trois cas sur six (2005, 2006 et 2008), tandis que pour les autres années, l'écart-type est plus faible si nous utilisons les poids non mis à jour. Par ailleurs, dans le cas de la sélection avec probabilité proportionnelle à la taille (PPT et EPS), nous obtenons de meilleurs résultats dans la plupart des

cas (cinq années sur six) si nous utilisons les poids mis à jour. De surcroît, le niveau de l'écart-type est fortement réduit (pour passer d'environ -10 % à -44 %); sur les six années, l'échantillonnage PPT donne lieu à une réduction moyenne de -12,6 % et l'EPS donne d'encore meilleurs résultats, produisant une réduction moyenne de -14,2 %. Pour 2009 seulement, l'écart-type augmente considérablement si l'on utilise les poids mis à jour plutôt que non mis à jour (PPT : +43,8 %; EPS : +49,1 %).

3.3 Remarques finales

L'étude de la variabilité des poids au cours du temps nous a permis de constater que les poids d'échantillonnage varient davantage dans le long terme, tandis que les poids économiques varient davantage sur une base annuelle. En outre, la surestimation ou la sous-estimation de l'indice réel et ses variances diffèrent considérablement d'année en année : il est impossible, en moyenne, de dégager une tendance claire de l'évolution des poids. Une étude supplémentaire devra être réalisée aussitôt que les données du nouveau RE seront disponibles. En outre, une étude plus approfondie, tenant compte des différents secteurs économiques, aiderait peut-être à observer les trajectoires qui pourraient ne pas être décelées quand elles sont étudiées globalement. Cela nous permettra non seulement de savoir si nous pouvons définir une évolution plus régulière de certains secteurs économiques au cours du temps, mais aussi à prévoir leur évolution future. Afin de déterminer à quelle fréquence les poids devraient être mis à jour, nous avons comparé les valeurs de l'indice réel (calculées avec les poids courants) et les valeurs des indices simulés (calculées avec des points non mis à jour). Nous avons également étudié la trajectoire des indices à mesure que les poids deviennent dépassés. En examinant les résultats pour l'EAS, nous avons constaté que, même si, au niveau individuel, les trajectoires des indices diffèrent vraiment les unes des autres, en moyenne, le biais de l'indice (et sa variance) est habituellement réduit quand on utilise des poids qui ne datent pas de plus de deux ans. Pour les méthodes d'échantillonnage avec probabilité proportionnelle à la taille, nous avons habituellement obtenu un plus grand biais (vu l'importance plus grande accordée aux indices des unités les plus grandes), qui augmente aussi légèrement quand les poids deviennent plus anciens. Toutefois, de nouveau, il s'agit de la moyenne de trajectoires entièrement différentes. Ensuite, nous avons comparé certaines méthodes d'échantillonnage en ce qui concerne la réduction en pourcentage du biais et de l'écart-type obtenue en utilisant des poids mis à jour (courants). Comparativement à l'échantillonnage aléatoire simple, les méthodes d'échantillonnage avec probabilité proportionnelle à la taille donnent de meilleurs résultats (dans trois cas sur six et en moyenne). En outre, quand nous utilisons les poids mis à jour, l'écart-type de l'indice est réduit également (dans cinq cas sur six et en moyenne), ce qui signifie que l'indice est plus précis. En revanche, l'échantillonnage aléatoire simple ne produit habituellement pas de meilleurs résultats si l'on utilise les poids courants : il donne de meilleurs résultats de manière « plus aléatoire ». Donc, nous devrions toujours utiliser les poids mis à jour. Toutefois, le principal problème quand on choisit un système de pondération est que les poids courants ne sont habituellement pas disponibles pendant un certain temps. La question est donc de savoir à quelle fréquence les poids devraient être mis à jour. Aussi souvent que possible, évidemment. Idéalement, ils devraient l'être annuellement (ou aussitôt que de nouvelles données sont disponibles); en fait, les poids d'échantillonnage datant de plus de deux ans ne sont pas recommandés (voir les graphiques 3.2-1/2). Si les poids sont mis à jour moins fréquemment, on aura peut-être parfois de la chance et on utilisera un système de poids qui correspondent aux poids courants, parce que les poids ne suivent pas une tendance linéaire. Toutefois, il est impossible de déterminer quel est le système de poids le plus semblable à ceux d'une année particulière (c.-à-d. l'année courante) si nous ne connaissons pas les poids correspondants. En outre, si les poids ne sont pas mis à jour, la variance de l'indice est plus forte, ce qui réduit sa précision. Nous avons concentré la deuxième partie de notre étude sur l'évolution des poids d'échantillonnage, de sorte que d'autres études seront nécessaires pour mesurer l'effet des poids économiques sur les estimations des indices, ainsi que l'effet combiné des poids d'échantillonnage et des poids économiques en tenant compte, tout compte fait, des divers secteurs économiques.

Bibliographie

Boskin, M.J., Dulberger, E., Gordon, R., Griliches, Z. et Jorgenson, D. (1996). Toward a More Accurate Measure of the Cost of Living. *Final Report to the Senate Finance Committee* from the Advisory Commission To Study The Consumer Price Index. December 4. http://www.ssa.gov/history/reports/boskinrpt.html.

Consumer Price Index reference paper (1995). Statistique Canada, Price Division. Ottawa, Canada: Minister of Industry.

Laspeyres, E. (1884). Hamburger Warenprise 1851-1860 un die californisch-australicher soldent dukung seit 1848, Ein beitrag zur dehre von des Geldentwertung, in *Jarbüucher fur Nationalöokonomie.*

Patak, Z. and Lothian, J. (2007). *Enhancing the Quality of Price Indexes - A Sampling Perspective*. Ottawa, Canada: Statistique Canada.

Producer Price Index Manual - Theory and Practice (2004). Washington, D.C.: International Monetary Fund (http://www.imf.org).

Toninelli, D. (2008). Survey Techniques: an application to prices data for the computation of price indexes, unpublished PhD thesis, Bergamo, Italy: University of Bergamo.

L'ajout d'une composante longitudinale au programme des données fiscales agricoles de Statistique Canada

Terri Blanchard et Peter Xiao[1]

Résumé

Le Programme des données fiscales des entreprises agricoles (PDF) de Statistique Canada est conçu avant tout pour produire des estimations transversales des variables financières. Les données administratives proviennent des déclarations de revenus envoyées à l'Agence du revenu du Canada (ARC) par les entreprises agricoles non constituées en société, constituées en société et collectives, par voie électronique ou sur support papier. Une composante longitudinale a été ajoutée au PDF à compter de l'année d'imposition 2001. L'objectif de la nouvelle composante est de suivre l'évolution des entreprises agricoles individuelles au cours du temps. Le panel de l'année d'imposition 2001 a fait l'objet d'un suivi annuel. Des panels annuels de cohortes ont été créés à compter de l'année d'imposition 2006 et font également l'objet d'un suivi annuel.

Le présent article décrit la composante longitudinale du PDF, la stratégie d'imputation conçue spécialement pour les unités longitudinales et la méthode de pondération.

Mots-clés : données longitudinales, dossiers administratifs, données fiscales.

1. Introduction

Le Programme des données fiscales des entreprises agricoles (PDF) de Statistique Canada produit des estimations transversales de variables financières, telles que les revenus et dépenses d'exploitation moyens, le bénéfice net d'exploitation et le revenu hors ferme des exploitants agricoles, des exploitations agricoles et des familles d'agriculteurs. Le PDF obtient auprès de l'Agence canadienne du revenu (ARC) des données fiscales provenant de trois sources, à savoir les exploitants agricoles non constitués en société (qui produisent une déclaration T1), les entreprises agricoles constituées en société (qui produisent une déclaration T2) et les exploitations agricoles collectives (qui produisent une déclaration T3). À partir de l'année d'imposition 2001, une composante longitudinale a été ajoutée au PDF en vue de créer une base de données de recherche pour suivre l'évolution des entreprises agricoles individuelles au cours du temps. Par exemple, la base de données pourrait être utilisée pour déterminer si des résultats financiers solides ou médiocres sont une caractéristique transitoire ou permanente propre à certains types d'entreprises agricoles. La cohorte de 2001 est suivie annuellement et il existe à l'heure actuelle une base de données longitudinales contenant des données recueillies pendant six années pour cette cohorte. Deux cohortes supplémentaires ont été créées à partir des années d'imposition 2006 et 2007, respectivement.

Le présent article porte sur la composante longitudinale du PDF, en particulier la création de la base de données longitudinales pour la cohorte de 2001. À la section 2, nous donnons un aperçu du plan d'échantillonnage. À la section 3, nous nous penchons sur la stratégie d'imputation élaborée pour traiter la non-réponse au sein de la cohorte longitudinale et à la section 4, nous décrivons le calcul des poids utilisés pour l'estimation.

2. Plan d'échantillonnage

Le PDF des entreprises agricoles produit annuellement des estimations transversales fondées sur des données fiscales et, par conséquent, le plan d'échantillonnage et les systèmes de traitement des données sont bien établis. Au moment du lancement de la composante longitudinale en 2001, il a été décidé de maximiser le chevauchement entre la cohorte longitudinale et l'échantillon transversal afin d'utiliser dans la mesure du possible l'infrastructure existante pour l'échantillonnage et le traitement des données. Pour la première année de la cohorte longitudinale, l'échantillon est en fait le même que l'échantillon transversal pour cette année-là. Par conséquent, le plan d'échantillonnage aux sections qui suivent est celui utilisé pour la première année de l'échantillon longitudinal ainsi que pour les échantillons transversaux annuels.

[1]Terri Blanchard, Statistique Canada, 100 Promenade Tunney's Pasture, Ottawa, Canada, K1A 0T6, terri.blanchard@statcan.gc.ca,; Peter Xiao, Statistique Canada, 100 Promenade Tunney's Pasture, Ottawa, Canada, K1A 0T6, peter.xiao@statcan.gc.ca

Bien que l'on tire des échantillons distincts et que l'on recueille des éléments de données différents pour les déclarants T1, T2 et T3, les données pour les trois types de déclarants sont consolidées en un ensemble commun de variables pour la production des estimations. Le PDF produit des estimations qui s'appuient sur les données combinées des déclarants T1, T2 et T3.

Les entreprises du secteur agricole n'utilisent pas les mêmes variables que les entreprises des autres secteurs commerciaux pour déclarer leurs données fiscales. À Statistique Canada, la Division de l'agriculture est entièrement responsable de la vérification, de l'imputation et de l'analyse en profondeur des données utilisées pour produire les estimations concernant le secteur agricole. Les données fiscales provenant de l'ARC utilisées par les autres secteurs de programme de Statistique Canada sont entièrement épurées et traitées par la Division des données fiscales.

Un recensement des quelques 300 déclarants de l'univers T3 est réalisé chaque année. La suite de l'article porte uniquement sur les déclarants T1 et T2.

2.1 Plan d'échantillonnage pour la partie T1

Pour la partie T1 de l'échantillon longitudinal, la population cible englobe toutes les personnes ayant déclaré un revenu agricole provenant d'un travail autonome (déclarants T1) la première année de la cohorte. Les données sont fournies par l'ARC qui les extrait des déclarations de revenus produites par les particuliers. L'ARC reçoit les données des déclarants T1 qui déclarent des revenus agricoles sous trois formes, à savoir les formules imprimées envoyées par la poste à l'ARC, les formules transmises par voie électronique et les formules des programmes Agri-stabilité/Agri-investissement[2]. Les formules imprimées envoyées à l'ARC comprennent les formules remplies à la main et celles remplies en se servant d'un logiciel de calcul d'impôt qui contiennent un code à barres à la première page. Les formules imprimées porteuses d'un code à barres (déclarants avec code à barres) sont traités par le PDF comme des déclarants par voie électronique. L'ARC envoie à Statistique Canada les données provenant de toutes les formules T1 transmises électroniquement, de toutes les formules à codes à barres et de toutes les formules Agri-stabilité/Agri-investissement et d'un échantillon des formules imprimées ne portant pas de codes à barres. La sélection de l'échantillon de tous les types de formules et la saisie des données des formules imprimées est faite par l'ARC conformément aux spécifications fournies par le PDF.

Un échantillon des déclarants par voie électronique et des déclarants du programme Agri-stabilité/Agri-investissement est sélectionné même si le PDF reçoit les données pour l'univers complet de ces catégories de déclarants. L'échantillonnage a pour objectif de réduire le coût et le temps requis pour vérifier, corriger et imputer les données erronées ou incomplètes de toutes les déclarations par voie électronique et du programme Agri-statibilité/Agri-investissement en utilisant le système existant. Afin de maximiser leur utilisation, les données des déclarations par voie électronique et du programme Agri-statibilité/Agri-investissement sont soumises à une série de vérifications automatisées afin d'évaluer leur qualité. Les unités qui sont acceptées à toutes les vérifications et qui, par conséquent, ne nécessitent aucune intervention manuelle sont considérées comme épurées. Les unités épurées qui ne sont pas sélectionnées aléatoirement sont ajoutées à l'échantillon transversal. Ces unités supplémentaires ne font pas partie de la cohorte longitudinale.

La population T1 est répartie en strates correspondant aux dix provinces et aux trois territoires du Canada, puis, au plus, en huit strates de taille basées sur le revenu agricole brut (RAB) par la méthode de la courbe cumulative de la racine carrée des fréquences. Les taux d'échantillonnage par strate sont déterminés en répartissant une taille d'échantillon prédéterminée entre les provinces et territoires selon une méthode de répartition proportionnelle à la racine carrée, puis entre les strates de taille dans chaque province ou territoire par la méthode de répartition de Neyman. Le Système généralisé d'échantillonnage (SGECH) de Statistique Canada est utilisé pour effectuer la stratification et la répartition. Pour plus de renseignements sur le SGECH, voir Statistique Canada (2006). Un recensement est effectué à Terre-Neuve-et-Labrador et dans les trois territoires étant donné les petites tailles de population. Un échantillonnage est effectué dans les neuf autres provinces.

L'échantillon est sélectionné selon la méthode d'échantillonnage de Poisson. Un numéro pseudo-aléatoire, fondé sur le numéro d'assurance sociale, est attribué à chaque déclaration de revenus T1 admissible en se servant d'une fonction de hachage qui produit une valeur comprise entre 0 et 999. Le taux d'échantillonnage dans chaque strate est également converti en un nombre compris entre 0 et 999 dénommé nombre limite supérieur de hachage. Une unité est sélectionnée dans l'échantillon si son numéro pseudo-aléatoire est inférieur ou égal au nombre limite supérieur de hachage pour la strate à laquelle elle appartient. Cette méthode est utilisée afin que l'ARC puisse saisir les données des déclarations imprimées sélectionnées au moment où elle les reçoit et que le PDF puisse traiter les données provenant des unités échantillonnées à mesure qu'elles sont transmises par

[2] Le programme Agri-stabilité/Agri-investissement était appelé anciennement Programme canadien de stabilisation du revenu agricole (PCSRA) et Compte de stabilisation du revenu net (CSRN).

l'ARC. Si l'on attendait d'avoir reçu les déclarations pour l'univers complet avant de procéder à l'échantillonnage et au traitement des données, les estimations ne seraient pas publiées dans les délais voulus.

2.2 Plan d'échantillonnage pour la partie T2

Pour la partie T2 de l'échantillon longitudinal, la population cible comprend toutes les entreprises agricoles constituées en société (déclarants T2). Les données obtenues auprès de l'ARC proviennent des déclarations de revenus que les sociétés transmettent à l'Agence. L'ARC fournit à Statistique Canada des données pour l'univers T2 complet. Un échantillon de déclarants T2 est sélectionné, parce que les données reçues ne sont pas toujours complètes et nécessitent parfois des corrections manuelles importantes qui sont à la fois longues et coûteuses.

La population T2 est stratifiée en fonction des dix provinces et des trois territoires du Canada, puis en fonction de 11 catégories d'exploitation agricole. Ces 11 catégories sont les suivantes : graines oléagineuses et céréales, pommes de terre, autres légumes, fruits et noix, serres et pépinières, autres récoltes, bovins de boucherie, bovins laitiers, porcins, volaille et production d'œufs, et autre élevage d'animaux. Chaque strate géographie-catégorie d'exploitation agricole est ensuite stratifiée en, au plus, quatre strates de taille en fonction du chiffre des ventes selon la méthode de la courbe cumulative de la racine carrée des fréquences. Les taux d'échantillonnage sont déterminés en attribuant d'abord une taille d'échantillon prédéterminée aux strates géographie-catégorie d'exploitation agricole selon la méthode de répartition proportionnelle à la racine carrée. Ensuite, dans chacune des strates géographie-catégorie d'exploitation agricole, l'échantillon est réparti entre les strates de taille selon la méthode de répartition Neyman. Le même système généralisé d'échantillonnage (SGECH) que celui utilisé pour les déclarants T1 est utilisé pour la stratification et la répartition de l'échantillon. Un recensement est réalisé dans les quatre provinces atlantiques et dans les trois territoires à cause des petites tailles de population. Dans les six autres provinces, on procède à un échantillonnage par la méthode de Poisson, comme pour les déclarants T1. Le numéro pseudo-aléatoire produit pour chaque unité en utilisant une fonction de hachage est basé sur le numéro d'entreprise de l'unité.

2.3 Suivi des unités de la cohorte longitudinale

Comme nous l'avons mentionné aux sections précédentes, pour la première année d'une cohorte, l'échantillon longitudinal est constitué de toutes les unités sélectionnées aléatoirement de la composante transversale de la même année. Une fois que ces unités sont identifiées, elles sont suivies annuellement pendant dix ans. Afin d'être certain que Statistique Canada reçoive les données de l'ARC pour les unités longitudinales d'une cohorte au cours des années subséquentes, les unités longitudinales sont spécifiées de façon à ce qu'elles soient incluses dans les fichiers des échantillons transversaux, même si elles ne sont pas sélectionnées aléatoirement. La fonction de hachage utilisée pour calculer les numéros pseudo-aléatoires des unités pour l'échantillonnage maximise le chevauchement entre les unités sélectionnées aléatoirement d'une année à la suivante. Cette stabilité de l'échantillon réduit au minimum le nombre d'unités longitudinales qui sont ajoutées aux fichiers des échantillons transversaux.

La cohorte de 2001 contient 26 086 déclarants T1, 11 674 déclarants T2 et 287 déclarants T3. La cohorte de 2006 contient 20 345 déclarants T1, 7 523 déclarants T2 et environ 300 déclarants T3. La cohorte de 2007 contient 19 856 déclarants T1, 7 803 déclarants T2 et 300 déclarants T3. La diminution de la taille de l'échantillon de déclarants T2 pour les années d'imposition 2006 et suivantes est due à une réduction de la taille des échantillons transversaux à compter de l'année d'imposition 2006.

3. Traitement de la non-réponse

Comme dans les enquêtes, les cohortes longitudinales présentent des cas de non-réponse partielle et totale. La non-réponse partielle et la non-réponse totale sont traitées de manière différente, comme le décrivent les sections qui suivent.

3.1 Non-réponse partielle

La non-réponse partielle peut être due au fait que les déclarants qui produisent les déclarations de revenus ne sont pas obligés de remplir tous les champs figurant sur la formule de déclaration de l'ARC. Par exemple, certains déclarants peuvent fournir uniquement des renseignements génériques, tels que des totaux, au lieu de donner des renseignements détaillés sur les composantes de ces totaux. La non-réponse partielle est traitée par imputation déterministe, imputation historique, imputation par donneur ou imputation manuelle. L'imputation par donneur s'appuie sur la méthode du plus proche voisin et est utilisée principalement pour attribuer des valeurs totales à une ventilation plus détaillée quand certains détails ne sont pas fournis. La ventilation détaillée de l'enregistrement donneur est appliquée à l'enregistrement receveur. Toutes les imputations partielles

pour les unités de la cohorte longitudinale font partie du système de traitement utilisé pour produire les estimations transversales.

3.2 Non-réponse totale

La non-réponse totale peut se produire pour diverses raisons, par exemple quand une unité ne produit pas de déclaration de revenus pour une année particulière ou que l'unité a produit une déclaration de revenus, mais que celle-ci a été reçue après que le PDF ait achevé le traitement des données. Les unités qui ne produisent pas de déclaration de revenus pour une année particulière ne figurent pas dans le fichier de l'univers de l'enquête et sont codées comme des unités non agricoles. Dans la base de données longitudinales, les données sont imputées pour les unités qui figurent dans le fichier de l'univers de l'enquête, mais dont les données sont reçues après le traitement par le PDF. Pour l'estimation transversale, les unités qui produisent une déclaration tardive sont traitées par repondération.

La majorité des unités manquantes sont imputées en recourant à l'imputation par donneur selon la méthode du plus proche voisin. Le module d'imputation massive de Banff, qui est le système généralisé de vérification et d'imputation de Statistique Canada, est utilisé pour l'imputation par donneur. Pour plus de renseignements sur Banff, voir Statistique Canada (2007).

Les sections qui suivent portent sur la stratégie d'imputation appliquée aux données de la cohorte de 2001 pour les années d'imposition 2001 à 2005 pour lesquelles l'imputation a été effectuée simultanément et pour l'année d'imposition 2006, pour laquelle l'imputation a été effectuée ultérieurement.

3.2.1 Stratégie d'imputation par donneur pour la non-réponse totale de 2001 à 2005

Les données de la cohorte longitudinale de 2001 pour les années d'imposition 2001 à 2005 ont été imputées de façon groupée parce que l'imputation a été effectuée seulement après avoir reçu les données pour l'année d'imposition 2005. La stratégie d'imputation a été élaborée de manière à profiter du fait que l'on disposait de données pour une période de cinq ans, ce qui a permis d'utiliser la courbe de réponse pour les cinq années afin de créer des classes d'imputation. En outre, le nombre de variables d'appariement pouvant être utilisées pour recherche un bon donneur était grand.

La stratégie d'imputation consistait à utiliser l'année la plus récente pour laquelle le receveur avait fourni des données avant la ou les années pour lesquelles des données manquaient afin de repérer un donneur dont les données pourraient être utilisées pour l'imputation pour la ou les années manquantes. Le même donneur a été utilisé pour toutes les années manquantes consécutives, afin que les données imputées soient cohérentes. Les données de l'enregistrement donneur ont été corrigées en se basant sur une variable commune avant d'être copiée dans l'enregistrement receveur afin d'aligner les données du donneur avec les données existantes du receveur.

La première étape du processus d'imputation consistait à grouper les enregistrements receveurs et les enregistrements donneurs selon leur courbe de réponse de 2001 à 2005. Par exemple, les enregistrements donneurs contenant des données pour 2001, 2002 et 2003, mais non pour 2004 et 2005 ont été regroupés avec des donneurs possibles contenant des données pour 2003, 2004 et 2005. Les données pour 2003 contenues dans les enregistrements receveurs et donneurs ont été utilisées comme variables d'appariement pour trouver un donneur. Dix groupes ont été créés ainsi selon les courbes de réponse observées pour 2001 à 2005. Voir Blanchard (2008) pour plus de renseignements sur les groupes selon la courbe de réponse.

Dans chaque groupe caractérisé par une courbe de réponse, nous avons créé des classes d'imputation. Ces dernières sont des groupes d'enregistrements homogènes dans lesquels des donneurs seront trouvés pour les receveurs. La première exécution de l'imputation par donneur a été effectuée en définissant les classes d'imputation selon la province et la catégorie d'exploitation agricole. Les mêmes catégories d'exploitation agricole ont été utilisées pour toutes les provinces. Pour la deuxième exécution de l'imputation par donneur, les classes d'imputation ont été définies selon la catégorie d'exploitation agricole seulement.

Les variables d'appariement utilisées pour trouver un donneur étaient le revenu total d'exploitation, les dépenses totales d'exploitation et le revenu selon la catégorie d'exploitation agricole pour l'année la plus récente pour laquelle le receveur avait fourni des données, ainsi que le revenu agricole brut (RAB) (déclarants T1) ou le chiffre des ventes (déclarants T2) pour la ou les années manquantes. Par exemple, si un receveur avait fourni des données en 2003, mais que des données manquaient pour 2004 et 2005, le revenu total d'exploitation, les dépenses totales d'exploitation et le revenu selon la catégorie d'exploitation agricole déclarés pour 2003, ainsi que le RAB ou le chiffre des ventes tiré des fichiers de l'univers de l'enquête pour 2004 et 2005 ont été utilisés comme variables d'appariement.

Chaque classe d'imputation devait contenir au moins dix donneurs et le pourcentage minimal de donneurs par classe d'imputation était de 30 %. Les enregistrements imputés durant la première tentative d'imputation par donneur pour les

données longitudinales n'ont pas été inclus dans les réserves de donneurs pour les exécutions subséquentes de l'imputation par donneur.

Après avoir trouvé un donneur, un facteur de correction correspondant au ratio des revenus totaux d'exploitation du receveur et du donneur a été appliqué aux données de l'enregistrement donneur avant de les copier dans l'enregistrement receveur, afin qu'elles concordent avec les données existantes de l'enregistrement receveur.

3.2.2 Stratégie d'imputation par donneur pour 2006

Les données manquantes pour l'année d'imposition 2006 ont été imputées après avoir achevé l'imputation pour les années 2001 à 2005. La stratégie d'imputation était fort semblable à celle suivie pour les années 2001 à 2005, mais il existait certaines différences.

L'année la plus récente pour laquelle l'enregistrement donneur contenait des données non imputées antérieurement à 2006 a été utilisée pour trouver un donneur dont les données pourraient être utilisées pour l'imputation pour 2006. L'année la plus récente utilisée pour trouver un donneur était l'année d'imposition 2002, qui était antérieure de quatre ans à l'année pour laquelle des données manquaient (2006). Cette approche était en harmonie avec la stratégie utilisée pour imputer les données pour les années d'imposition 2001 à 2005. Les cas pour lesquels 2001 était la seule année pouvant être utilisée pour trouver un donneur ont été examinés manuellement afin de déterminer s'ils devaient être codés comme des entreprises non agricoles.

Le regroupement des receveurs et des donneurs selon la courbe de réponse a produit des groupes légèrement différents de ceux créés pour l'imputation pour 2001 à 2006. Seulement cinq groupes selon la courbe de réponse ont été créés en se fondant sur l'existence de données non imputées pour 2001 à 2005. Ainsi, les enregistrements receveurs contenant des données non imputées pour 2004, des données imputées pour 2005 et aucune donnée pour 2006 ont été groupés avec les donneurs possibles s'ils contenaient des données non imputées pour 2004 et pour 2006. Voir Blanchard (2008) pour des renseignements plus détaillés sur les groupes selon la courbe de réponse utilisés pour l'imputation des données pour l'année d'imposition 2006. Dans les cas où les données de 2005 de l'enregistrement receveur étaient imputées, le donneur choisi en 2005 n'a pas été automatiquement utilisé pour imputer les données pour 2006. Cela signifie que la règle d'utilisation du même enregistrement donneur pour les années manquantes consécutives ne s'applique qu'aux années 2001 à 2005.

Les variables d'appariement utilisées pour trouver un donneur étaient les mêmes que celles employées pour imputer les données pour les années d'imposition 2001 à 2005. Après avoir trouvé un donneur, une correction correspondant au ratio des revenus totaux d'exploitation a été appliquée aux données de 2006 de l'enregistrement donneur avant qu'elles soient copiées dans l'enregistrement receveur afin d'harmoniser les données du donneur avec celle du receveur.

3.2.3 Imputation historique

Un petit nombre d'unités pour lesquelles on ne pouvait utiliser l'imputation par donneur ont fait l'objet d'une imputation historique au moyen d'une tendance. Certains enregistrements nécessitant une imputation n'ont pu être imputés par donneur parce que la classe d'imputation ne contenait pas suffisamment de donneurs ou parce que les données de l'enregistrement receveur pour les variables d'appariement employées pour trouver un donneur n'étaient pas valides. Dans ces cas, les données utilisées pour imputer une unité manquante étaient celles pour l'année la plus récente pour laquelle l'unité en question contenait des données non imputées. Il s'agissait habituellement de la même année que celle qui aurait été utilisée pour trouver un donneur pour l'imputation par donneur. Les données historiques ont été corrigées en se basant sur le revenu agricole brut, dans le cas des déclarants T1, ou sur le chiffre des ventes, dans le cas des déclarants T2, pour la ou les années manquantes.

L'imputation par donneur a été exécutée avant l'imputation historique afin de tenir compte des changements de catégorie d'exploitation agricole d'une unité d'une année à l'autre. Chaque année, la catégorie d'exploitation agricole est attribuée en se basant sur les produits dont les revenus représentent plus de 50 % du revenu total pour l'année courante. Autrement dit, une unité peut changer de catégorie d'exploitation agricole d'une année à l'autre si la majorité de ses revenus proviennent de différents produits d'une année à la suivante.

3.2.4 Taux d'imputation

En général, les taux d'imputation sont assez faibles; ils sont habituellement inférieurs à 2 % pour les déclarants T1 la plupart des années et inférieurs à 6 % pour les déclarants T2 la plupart des années. Les taux d'imputation pour les déclarants T1 et T2 pour la cohorte de 2001 sont présentés au tableau 3.2.4-1. Toute anomalie dans les taux d'imputation est due à des problèmes de traitement des données ou à des changements survenus durant l'année en question uniquement. Ces problèmes ont été résolus et l'on s'attend à une diminution des taux d'imputation quand ceux-ci sont élevés. Le tableau montre aussi que le

pourcentage d'unités codées comme des entreprises non agricoles augmente, parce qu'elles ne figurent pas dans le fichier de l'univers de l'enquête. Il convient de souligner que les déclarants T1 pourraient ne pas apparaître dans ce fichier parce qu'ils se sont constitués en société et qu'ils font désormais partie de l'univers T2. Comme il est impossible de dépister ces cas, ils sont codés comme des entreprises non agricoles dans l'univers T1.

Tableau 3.2.4-1
Taux d'imputation pour la cohorte de 2001

	Déclarants T1 (%)			Déclarants T2 (%)		
Année d'imposition	Entreprises agricoles	Entreprises non agricoles	Taux d'imputation	Entreprises agricoles	Entreprises non agricoles	Taux d'imputation
2001	100,0	…	…	100,0	…	…
2002	94,2	3,8	1,9	96,6	2,0	1,3
2003	86,0	7,5	6,6	91,3	3,3	5,4
2004	87,9	10,2	1,9	88,6	6,5	5,0
2005	84,8	13,5	1,7	81,1	5,9	13,0
2006	81,3	17,3	1,4	81,8	6,9	11,3

Nota : … n'ayant pas lieu de figurer

4. Pondération longitudinale

Les poids longitudinaux appliqués à la cohorte longitudinale de l'année d'imposition 2001 ont pour but de s'assurer que les données sont représentatives de la population durant l'année d'imposition 2001. Les populations longitudinales T1 et T2 ont été créées après le traitement des données obtenues pour l'année d'imposition 2001. Ensuite, les tailles de population et d'échantillon ont été déterminées pour chaque strate. Le poids d'estimation longitudinal est égal au quotient de la taille de la population divisée par la taille de l'échantillon. Il convient de souligner que l'échantillon longitudinal ne contient que les unités sélectionnées aléatoirement et n'englobe pas les unités ayant produit une déclaration électronique épurée ni les unités ayant produit une déclaration Agri-stabilité/Agri-investissement épurée qui ne sont pas sélectionnées aléatoirement. Pour chacune des dix années pour laquelle on possède des données pour la cohorte de 2001, les poids d'estimation demeurent les mêmes, parce que la cohorte représente les populations T1 et T2 de l'année d'imposition initiale. Pour les cohortes de 2006 et de 2007, les poids d'estimation peuvent être calculés de la même façon et demeureront constants pour les dix années subséquentes.

5. Conclusion

La composante longitudinale du Programme de données fiscales des exploitations agricoles fournit aux analystes et aux chercheurs une foule de renseignements sur les changements qui surviennent dans le secteur de l'agriculture au cours du temps. Lorsque la cohorte de 2001 achèvera, on disposera de données couvrant dix années pour l'analyse.

À mesure qu'augmente le nombre de cohortes suivies, le nombre d'unités longitudinales traitées dans le système de traitement transversal augmente également. Afin de réduire le fardeau de la composante transversale, la validation et l'imputation des données pour les unités longitudinales se feront de manière plus automatisée, séparément du traitement des unités transversales. L'emploi d'un système de traitement des données plus automatisé pourrait également permettre de créer un fichier longitudinal contenant des données pour l'univers complet des exploitants agricoles non constitués en société et des entreprises agricoles constituées en société plutôt qu'un échantillon seulement de ces unités.

L'élaboration d'un processus de validation et d'imputation des données automatisé, les améliorations apportées à la stratégie d'imputation et la création d'un fichier longitudinal contenant des données pour l'univers complet seront les sujets des travaux à venir concernant la composante longitudinale.

Bibliographie

Blanchard, T. (2008), Imputation Strategy for the 2001 Cohort of the Agriculture Tax Data Program Longitudinal Data Project, document interne, Statistique Canada.

Statistique Canada (2007), BANFF User's Guide (for v2.01 and 2.02), document interne, Statistique Canada.

Statistique Canada (2006), Generalized Sampling System Version 2.3 User's Guide, document interne, Statistique Canada.

L'entreprenariat : enquêtes longitudinales auprès de populations émergentes et difficiles à retracer

E.J. Reedy[1]

Résumé

Au cours des dix dernières années, la Ewing Marion Kauffman Foundation (la Fondation) a été l'un des principaux bailleurs de fonds de la création d'ensembles de données de recherche longitudinales sur l'entreprenariat. La mesure de l'entreprenariat porte sur les populations des ménages et des établissements, un domaine d'étude des plus intéressants sur le plan méthodologique. Le présent article porte plus particulièrement sur ce que la Fondation a appris de la Kauffman Firm Survey (KFS) et de la Panel Study on Entrepreneurial Dynamics (PSED), enquêtes longitudinales menées auprès de jeunes et de nouvelles entreprises.

Mots-clés : entreprenariat, nouvelle entreprise, enquête-établissements, enquête-ménages, longitudinale.

1. Introduction

1.1 Mesure de l'entreprenariat

Ewing Kauffman était le type même de l'entrepreneur, qui s'est lancé en affaires avec peu de ressources et en 40 ans a transformé son entreprise en une société ayant un chiffre d'affaires de plusieurs milliards de dollars. Son esprit d'entreprise et sa réussite en affaires lui ont procuré des avantages ainsi qu'à sa famille, à ses associés et à sa collectivité. Il était d'avis que l'activité philanthropique pouvait et devait promouvoir l'entreprenariat afin d'améliorer la vie des gens. La Ewing Marion Kauffman Foundation est entrée dans l'ère moderne en 1992 en fournissant une formation aux entrepreneurs existants ainsi que des programmes d'études en entreprenariat aux étudiants. Vers la fin des années 1990, nous avons commencé à réorienter une partie de notre financement vers les projets de collecte de données ainsi que d'autres projets de recherche, tendance qui s'est accélérée en 2009 sous la direction de notre président actuel. Aujourd'hui, la Kauffman Foundation soutient de nombreux projets originaux de collecte de données et des efforts visant à améliorer l'infrastructure statistique à l'échelle mondiale.

Mais pourquoi les recherches et les mesures sont-elles importantes pour la promotion de l'entreprenariat? Promouvoir des politiques positives destinées à encourager l'entreprenariat demeure pour nous une priorité essentielle. Cependant, les politiques judicieuses sont fondées sur des recherches de qualité et, par le passé, l'entreprenariat comme domaine de recherche a été dans une large mesure relégué aux établissements de qualité inférieure. Nous avons reconnu la nécessité de modifier l'infrastructure des données afin d'améliorer les recherches. À compter de 2003, nous avons fortement augmenté nos dépenses au chapitre des données et élargi notre collaboration avec les communautés statistiques nationales et internationales sur la meilleure façon de mesurer l'entreprenariat d'une manière uniforme[2].

Pour mesurer l'entreprenariat, il faut d'abord définir ce concept jusqu'ici resté nébuleux. Un examen récent des ouvrages publiés révèle de nombreuses mesures communes : entreprises en démarrage, âge/survie de l'entreprise, travail autonome, financement de capital de risque, universitaires occupant un deuxième emploi, taille de l'entreprise, entreprises de haute technologie et caractéristiques du propriétaire, pour n'en nommer que quelques-unes (Reedy et Frazelle, 2009). Aux fins de nos

[1]E.J. Reedy, Ewing Marion Kauffman Foundation, 4801 Rockhill Road, Kansas City, MO, USA, 64110, ereedy@kauffman.org and www.kauffman.org/datamaven.

[2] Le présent article est trop court pour nous permettre d'exposer notre stratégie en détail. Le lecteur intéressé peut obtenir plus de renseignements dans notre site Web et, plus particulièrement, en consultant la série de publications annuelles intitulée *Kauffman Thoughtbooks* ou à l'adresse www.kauffman.org/research.

mesures, plutôt que d'en arriver à la définition correcte du terme entreprenariat, nous nous penchons sur les différentes parties du concept et nous tâchons de mieux mesurer chacune[3].

1.2 Panel Study on Entrepreneurial Dynamics (PSED)

Nous avons entrepris notre étude des données de panel dans le cadre du programme de recherche PSED qui visait à mieux faire comprendre sur le plan scientifique comment les gens démarrent une entreprise. La PSED fournit des données valides et fiables sur le processus de création d'entreprises fondées sur des échantillons représentatifs à l'échelle nationale de nouveaux entrepreneurs qui créent de nouvelles entreprises. La PSED I a débuté par une présélection en 1998-2000 pour choisir une cohorte de 830 entreprises auprès desquelles ont eu lieu les trois interviews de suivi. La PSED II a débuté par une présélection en 2005-2006 afin de repérer et de suivre 1 214 nouveaux entrepreneurs, et bénéficie à l'heure actuelle d'un financement pour réaliser six vagues. L'information obtenue comprend des données sur les nouveaux entrepreneurs, les activités entreprises durant le processus de démarrage et les caractéristiques des efforts de démarrage qui aboutissent à la création de nouvelles entreprises.

1.3 Kauffman Firm Survey (KFS)

La Kauffman Foundation a créé la KFS, qui est la plus grande enquête longitudinale auprès des nouvelles entreprises au monde, en sélectionnant une cohorte de 4 928 entreprises qui sont entrées en exploitation en 2004. La cohorte est suivie annuellement et des questions sont posées sur les antécédents des fondateurs, les sources et les montants de financement, les stratégies et les innovations des entreprises, ainsi que les résultats tels que les ventes, les bénéfices et la survie. La KFS en est à la cinquième période de collecte des données et huit périodes sont prévues.

2. Obstacles surmontés et leçons tirées

2.1 Construction de l'échantillon

Dans le cas de la plupart des enquêtes portant sur l'entreprenariat, il est difficile de trouver une base de sondage représentative de qualité élevée. Même les institutions publiques qui mènent pareilles recherches n'ont pas accès habituellement aux données fiscales et autres données leur permettant de créer une base de sondage à cause d'obstacles juridiques ou de problèmes d'actualité. Les établissements privés se heurtent à encore plus de difficultés. Pour la PSED, la population cible était celle des nouveaux entrepreneurs, soit une « personne » ou une équipe de personnes, de sorte que la base de sondage choisie était constituée de ménages. Toutefois, étant donné l'incidence relativement faible de nouveaux entrepreneurs actifs[4], à savoir des personnes qui, selon leur comportement récent, se lancent en affaires, qui s'attendent à être propriétaires d'une entreprise et dans le cas desquelles il n'y a pas de preuve d'une entreprise en exploitation, la présélection des ménages afin d'identifier les personnes qui se lancent en affaires a entraîné le dépistage de plus de 62 000 ménages dans la PSED I et de près de 32 000 ménages dans la PSED II[5]. Pour réduire les coûts, on a demandé à une entreprise de recherche par sondages du secteur privé de présélectionner un échantillon représentatif d'adultes afin d'identifier ceux qui procèdent activement à la création d'une entreprise. La même procédure a été utilisée dans la plupart des itérations internationales de la PSED[6]. Les personnes satisfaisant aux critères ont été invitées à subir une interview plus longue; environ 87 % des nouveaux entrepreneurs retenus à la présélection ont accepté d'y participer. Quelque 60 % de ces nouveaux entrepreneurs consentants ont participé à l'interview téléphonique initiale d'une heure pour le premier cycle (Reynolds et Curtin, 2008). Reynolds et Curtin se sont longuement penchés sur les modifications à apporter au protocole de présélection qui, en rétrospective, a peut-être éliminé du panel les cas dont on a déterminé qu'ils n'étaient pas de nouveaux entrepreneurs procédant activement à la création de nouvelles entreprises; à leur avis, toutefois, une interview initiale de 20 minutes aurait été nécessaire pour éliminer 83 cas de la PSED I et 277 cas de la PSED II (Reynolds et Curtin, 2008). D'ailleurs, pour la PSED, il serait presque impossible d'utiliser les données gouvernementales existantes pour construire une base de sondage, à moins d'ajouter des questions à une enquête mensuelle existante sur la population, comme la Current Population Survey aux États-Unis (Reynolds, 2008).

---

[3] De nombreuses améliorations ont été apportées aux définitions. L'Organisation de coopération et de développement économiques a pris l'initiative d'un processus pluriannuel d'harmonisation des définitions des bureaux de statistique nationaux, mais ce processus a porté surtout sur les données administratives des administrations publiques et les statistiques dérivées.

[4] Il s'agit d'environ 6 % de la population adulte aux États-Unis (Reynolds et Curtin, 2008).

[5] L'échantillon considérablement plus grand de la PSED I témoigne d'une procédure conçue pour augmenter le nombre de femmes et de minorités dans la cohorte des nouveaux entrepreneurs (Reynolds et Curtin, 2008).

[6] http://www.kauffman.org/Blogs/DataMaven/August-2009/Overview-of-International-Studies-on-Nascent-Entre.aspx

La Kauffman Firm Survey a été conçue comme une enquête longitudinale auprès des nouvelles entreprises (et non de personnes) et, par conséquent, nous avons déterminé qu'il était prudent d'utiliser une source commerciale de listes de nouvelles entreprises qui pourraient faire l'objet d'un processus de présélection plus poussé. Dun et Bradstreet (D&B) a la plus grande base de données comprenant les noms et adresses d'entreprises de tout organisme du secteur privé aux États-Unis. Même si, à l'instar de nombreuses autres personnes, nous ne sommes pas tout à fait convaincus de la qualité des données brutes, D&B était résolue d'être la seule base de données représentative à l'échelon national à partir de laquelle un échantillon pourrait être tiré. Toutes les données dans l'ensemble de données de recherche final ont été fournies ou confirmées par les participants à la suite d'une présélection de grande portée visant à déterminer que 2004 était l'année de la création de l'entreprise (Ballou, Barton, DesRoches, Potter, Zhao, Santos et Sebastian, 2007).

2.2 Financement et administration

La conception initiale du financement du programme PSED était tout à fait novatrice, ce qui a peut-être nui à la durabilité du financement à plus long terme. Le programme était initialement autofinancé, sans important soutien philanthropique ou gouvernemental, au moyen de la création de l'Entrepreneurial Research Consortium composé de 120 chercheurs de 34 établissements de recherche qui ont contribué financièrement à la création des données, chacun participant dans une mesure limitée à l'élaboration du questionnaire. En 1999, lorsque le Survey Research Laboratory de l'Université du Wisconsin, le fournisseur chargé de recueillir les données, a fermé ses portes de façon inattendue, la Kauffman Foundation est devenue l'un des importants bailleurs de fonds de la PSED. Le transfert des données de panel à un nouveau fournisseur, l'Institute for Social Research de l'Université du Michigan, s'est révélé un processus coûteux, et le consortium ne disposait pas de fonds lui permettant de poursuivre la collecte (Curtin et Reynolds, 2007). Avec la PSED II, la Kauffman Foundation s'est engagée initialement à financer le projet à l'Université du Michigan pour trois ans. Après trois ans, les chercheurs concernés ont réussi à obtenir le soutien de la Small Business Administration et de la National Science Foundation (NSF) pour une quatrième année de collecte de données. Le maintien du panel a permis aux chercheurs principaux d'obtenir de la NSF deux années supplémentaires de financement, portant la durée du panel à six ans en tout.

Dans le cas de la Kauffman Firm Survey, la Fondation a continué d'assurer l'administration et la gestion du projet. Bien que nous ayons reçu de l'extérieur un important encadrement et de nombreuses idées et suggestions pratiques touchant la collecte (Ballou et coll., 2007), la KFS n'aurait pas été réalisée sans d'importants appuis à l'interne. Nous avons appris de l'expérience de la PSED qu'il est difficile d'assurer la bonne entente des chercheurs tout au long du processus, même si nous avons, nous aussi, tâché de procéder par concertation pendant trop longtemps. Nous avons décidé d'assumer la responsabilité entière du projet, l'un de nos plus grands, en le baptisant Projet Kauffman et en nous en chargeant nous-mêmes. La firme Mathematica Policy Research a été sélectionnée au moyen d'un processus d'appels d'offres concurrentiel pour mener l'enquête. Scott Shane de la Case Western Reserve University a été le chercheur principal initial, travaillant en collaboration étroite avec la Fondation, mais Mathematica a été financée directement par la Fondation et le projet a toujours été dirigé à l'interne.

La réalisation de la PSED et de la KFS selon ces différents modèles a eu, à mon avis, les répercussions suivantes :

- Participation d'universitaires. Le modèle de consortium utilisé pour la PSED non seulement a permis aux universitaires dans le monde entier de participer à un riche débat théorique mais, à mon avis, de concert avec le programme parallèle mais complémentaire Global Entrepreneurship Monitor[7], a été essentiel aux premières répliques à l'échelle internationale[8]. Toutefois, peut-être à cause de la complexité du consortium, de son financement bootstrap et du choc externe de la fermeture de l'établissement de collecte des données, l'avalanche attendue d'articles universitaires (particulièrement dans les revues prestigieuses) a été lente à se matérialiser[9]. L'ensemble de données de la PSED I demeure sous-analysé, les chercheurs ne s'étant pas penchés sur de nombreuses questions. Pour la KFS, nous nous sommes rendu compte durant le deuxième cycle de collecte que l'association étroite avec Kauffman pourrait également présenter des défis. En l'absence d'un responsable des recherches clairement désigné et d'un engagement à fournir l'accès aux données, une importante action directe serait nécessaire pour amener un ensemble initial d'universitaires à utiliser les données aux fins de recherche. Nous avons décidé qu'il serait nécessaire de retenir les services d'un chercheur universitaire principal supplémentaire.

---

[7] Le Global Entrepreneurship Monitor (GEM) a été élaboré par Paul Reynolds, l'un des principaux chercheurs de la PSED et l'architecte théorique de l'un et l'autre projet. Le GEM a été lancé en même temps comme indice et a rapidement été répliqué. On trouvera d'autres renseignements à l'adresse http://www.gemconsortium.org/about.aspx?page=re_about_research.

[8] On trouvera un aperçu des répliques internationales à l'adresse http://www.kauffman.org/Blogs/DataMaven/August-2009/Overview-of-International-Studies-on-Nascent-Entre.aspx.

[9] Les principaux responsables de la PSED ne seraient probablement pas d'accord. On trouvera une bibliographie à l'adresse http://www.psed.isr.umich.edu/psed/download_document/13.

- Courbe d'apprentissage prononcée. Même si nous avions financé des enquêtes, la Fondation ne disposait pas de personnel ayant de l'expérience de la réalisation d'enquêtes avant la KFS. Par conséquent, il a fallu modifier de nombreuses parties du projet. Par exemple, le concept initial approuvé comprenait un panel de quatre ans et le suivi d'une deuxième cohorte d'entreprises pendant deux ans. Le plan a été modifié plusieurs fois pour permettre de mener un projet pilote d'envergure et pour tenir compte des coûts initiaux plus élevés que prévu. Ultérieurement, la mise à l'essai nous a permis d'obtenir des taux de réponse plus élevés tout en veillant à ce que les coûts annuels demeurent faibles.
- Approfondissement des connaissances. La KFS a entraîné un approfondissement des connaissances internes et il est devenu clair que la KFS et la PSED n'étaient pas les seuls projets de données financés par la Fondation. Par conséquent, nous avons centralisé la surveillance des projets de données au lieu de les confier à différents agents de programme. Nous avons pu ainsi continuer de surveiller les investissements dans ces deux programmes tout en aidant à tirer parti de projets plus petits et à orienter les chercheurs vers les données les plus utiles à leur domaine de recherche.

2.3 Activités prévues et réelles de l'échantillon

L'une des constatations étonnantes du programme de recherche PSED est que l'émergence d'entreprises est un processus dont la durée peut varier fortement. Dans le cas de nombreuses entreprises, il peut prendre moins d'un mois, pour la plus petite unité que l'enquête a permis de saisir. Dans le cas d'environ le tiers de l'échantillon, toutefois, l'émergence peut prendre plus de cinq ans sans jamais aboutir (Reynolds et Curtin, 2008). Cette période a été plus longue qu'on ne l'avait envisagé. Elle explique également, entre autres raisons, pourquoi l'enquête longitudinale idéale utilisant le même échantillon pour examiner les entreprises émergentes, les entrepreneurs actuels, les entreprises en exploitation et la croissance éventuelle des entreprises n'a pas été mise en œuvre.

Pour la KFS, les disparitions d'entreprises après le premier cycle étaient moins nombreuses que prévu (voir le tableau 2.4-1). Nous croyons que cela pourrait être attribuable à l'interaction de notre plan d'enquête avec deux autres facteurs, soit la base de sondage de D&B et les entreprises à la durée de vie éphémère. Dans le cas de la base de sondage de D&B, nous soupçonnons un biais en faveur des entreprises ayant besoin de crédit, lesquelles pourraient avoir une meilleure probabilité de survie, du moins initialement[10]. Étant donné que de nombreuses entreprises entrent sur le marché du travail et en sortent assez rapidement, si nous avons pu contacter une entreprise en 2005 pour parler de ses activités menées en 2004, le taux de non-réponse pourrait être plus élevé ou nos questions dans l'enquête de référence pourraient être biaisées en faveur de l'utilisation d'un panel de référence composé d'entrepreneurs actifs. Les écarts en ce qui concerne le taux d'attrition, comparativement aux autres sources gouvernementales, se sont rétrécis de façon appréciable depuis l'enquête de référence.

Tableau 2.3-1
État des efforts de démarrage selon le cycle d'enquête pour la PSED II

	Cycle A	Cycle B	Cycle C	Cycle D
Établissement d'une nouvelle entreprise		128	164	214
Cessation de tous les efforts		231	422	490
Résultat final total		359	586	704
En cours de démarrage	1214	613	472	381
Situation inconnue		242	156	129
Nombre total de cas		1214	1214	1214
Résultat final en pourcentage		26,9 %	48,3 %	57,9 %

Source : Curtin et Reynolds, 2009

2.4 Incitatifs, mode et conception du questionnaire

Lors du lancement de la PSED, la collecte de données de qualité se faisait soit par interview sur place, soit par la poste ou par téléphone. Dans la PSED I, les responsables ont constaté qu'une grande partie des questionnaires envoyés par la poste n'étaient pas remplis d'une manière uniforme, de sorte que ce mode d'enquête a été abandonné pour la PSED II, laissant un seul mode de collecte. Pour la KFS, plusieurs modes ont été envisagés, mais ultérieurement une approche bimodale, soit ITAO et Internet, a été adoptée. En acceptant les réponses du panel fixe fournies au téléphone ou en ligne, nous espérions faciliter aux

[10] Il s'agit seulement d'une hypothèse pour le moment, mais d'autres travaux de recherche sont en cours comparant la base de sondage de D&B à d'autres entreprises figurant dans les dossiers administratifs gouvernementaux durant la même période.

entrepreneurs la participation à l'enquête. Il convient de souligner que les intervieweurs ITAO utilisent pour entrer les données le même système en ligne que celui auquel les utilisateurs ont accès dans le site Web protégé. Initialement, les membres du panel hésitaient à fournir des renseignements en ligne mais, au fil du temps, de nombreux membres du panel sont volontairement passés à la participation en ligne (tableau 2.4-1); ce résultat conceptuel semble indiquer que de nombreuses enquêtes transversales pourraient éventuellement être transformées en enquêtes par panel une fois la confiance établie et que le coût de cette transformation serait raisonnable.

En ce qui a trait à la conception du questionnaire, l'approche du financement par consortium adoptée pour la PSED I a abouti à l'élaboration d'un questionnaire très long et fastidieux. En outre, bon nombre des questions soumises par les universitaires ayant participé initialement à la collecte de fonds pour les recherches n'ont jamais été analysées. Ainsi, pour la PSED I, on s'est penché sur les problèmes de financement au moment d'élaborer le plan d'enquête et on en est arrivé à la conclusion générale qu'il est difficile de procéder par concertation. Pour la PSED II, on a fait appel à un groupe plus petit tout en imposant des limites strictes à la longueur du questionnaire d'enquête.

Tableau 2.4-1
Données de la Kauffman Firm Survey selon le cycle

	Enquête de référence (2004)	**Premier suivi (2005)**	**Deuxième suivi (2006)**	**Troisième suivi (2007)**
Enquête menée	Juin 2005 – juillet 2005	Juin 2006 – janvier 2007	Juin 2007 – décembre 2007	Juin 2008 – décembre 2008
Interviews achevées	4928	3998	3390	2915
Cessation des activités vérifiée	S/O	369	408	540
Interviews téléphoniques achevées	77 %	41 %	37 %	35 %
Interviews par Internet achevées	23 %	59 %	63 %	65 %
Taux de réponse	43 %	89 %	82 %	78 %

Source : Robb, Ballou, DesRoches, Potter, Zhao et Reedy, 2009

Pour la KFS, des universitaires de l'extérieur ont participé à l'élaboration du questionnaire mais, après une mise à l'essai prolongée, environ la moitié des questions ont été supprimées parce qu'elles n'étaient pas pertinentes, parce qu'elles étaient difficiles à comprendre ou pour réduire la longueur du questionnaire (Ballou et coll., 2007). Les chercheurs principaux procèdent à un examen annuel du questionnaire et l'on demande aux chercheurs de proposer des modifications ou des ajouts aux questions en s'appuyant sur les données. En outre, les responsables de la KFS ont procédé à une importante mise à l'essai aléatoire de plusieurs mesures incitatives dans le projet pilote. Cette recherche (voir le tableau 2.4-2) a révélé une augmentation marquée de la participation dans le cas d'un paiement postérieur. Ultérieurement, un paiement postérieur de 50 $ a été choisi comme mesure incitative pour l'enquête de référence et a été maintenu pour chacune des interviews subséquentes.

Tableau 2.4-2
Taux de réponse à la Kauffman Firm Survey – Mise à l'essai de mesures incitatives

	Paiement préalable	
Paiement postérieur	Rien	1$
Rien	20 %	22 %
50 $	28 %	30 %

Source : Ballou, Zhao, DesRoches et Potter, 2007

Une petite digression purement hypothétique en ce qui concerne les mesures incitatives et l'image de marque : dans la KFS, il y a au moins dix cas par an de personnes qui communiquent avec la Kauffman Foundation au moyen de notre site Web au sujet de l'enquête. Dans certains cas, ces personnes ont supprimé leur mot de passe qui leur permet de participer et, dans d'autres cas, elles n'ont pas reçu leur chèque incitatif. Il semble que les membres du panel conservent les documents KFS fournis par Mathematica et, en outre, que de nombreux répondants se souviennent du nom Kauffman Foundation et visitent nos sites Web qui sont distincts de ceux de l'enquête. Étant donné cette situation, les associés de la Kauffman Foundation ont reçu la

formation nécessaire pour leur permettre de reconnaître les personnes qui devraient être réorientées vers Mathematica, de sorte qu'ils ont réussi à conserver bon nombre de ces participants.

2.5 Au-delà des données

En finançant les enquêtes par panel, la Kauffman Foundation avait deux buts, à savoir créer des données pour faire progresser les connaissances et susciter plus d'intérêt dans les milieux universitaires. Nous pensions, naïvement, que les universitaires avaient seulement besoin d'avoir accès aux données et que notre appui des données garantirait leur utilisation dans les meilleurs délais. Cependant, les produits de données à grande diffusion ne sont pas toujours intéressants pour les milieux universitaires où les normes disciplinaires encouragent l'utilisation de données exclusives dans le processus de publication et de titularisation. Nous en sommes arrivés à la conclusion que la collecte des données n'en garantit pas l'utilisation. En outre, les entreprises de recherche par sondages mettent peu d'accent sur la diffusion, laissant ce processus aux chercheurs universitaires principaux qui n'ont qu'une faible connaissance du système.

Avec le programme PSED, les principaux responsables du projet ont pris l'initiative en matière de diffusion, visant les universitaires qui ont participé au projet et qui ont présenté des communications pour des conférences et planifié des ateliers de perfectionnement professionnel dans le cadre de conférences existantes. Toutefois, aucun financement n'a été prévu dans les budgets originaux de la PSED pour la diffusion (conférences, recherches ou autres résultats). Même si certains fonds supplémentaires ont été fournis par Kauffman et d'autres organismes, à notre avis, l'absence de financement de ces diffusions dans le budget initial est regrettable (mais malheureusement chose courante).

Avec la KFS, nous avons inclus la diffusion dans le budget. Nous avons mis à l'essai un programme de subventions de déplacement et un programme de petites subventions, activités venant s'ajouter aux activités d'extension des conférences, aux webinaires de recherche, aux publications destinées au grand public et aux présentations itinérantes dans certaines universités. En outre, les différences nuancées suivantes entre la stratégie de diffusion adoptée pour PSED et celle adoptée pour la KFS sont importantes à mon avis :

- Inscription dans le site Web. Les questionnaires de la PSED et de la KFS peuvent être téléchargés gratuitement. Cependant, dans le cas de la KFS, il faut s'inscrire, ce qui nous permet de communiquer avec les utilisateurs des données et de leur fournir un soutien sous forme de courriels de mise à jour, de formation et d'activités d'extension.
- Production de versions. Nous avons créé plusieurs versions de l'ensemble de données de la KFS, leur utilisation étant assortie d'exigences de plus en plus strictes en matière de sécurité et offrant des avantages de plus en plus importants. Dans la version KFS Data Enclave, l'utilisateur peut avoir accès à des données plus riches dans un environnement protégé à distance. Par exemple, la présence de codes postaux dans le fichier de données à grande diffusion présenterait un risque de divulgation, mais nous pouvons fournir ce renseignement dans la version Enclave. En outre, nous continuons d'améliorer la version Enclave en y ajoutant des codes fournis par plus de 30 chercheurs.

Enfin, nous avons prévu des activités élargies venant s'ajouter aux projets PSED et KFS. Les Kauffman Symposiums on Entrepreneurship and Innovation Data permettent aux utilisateurs éventuels de nouvelles données de rencontrer les producteurs de nouvelles données. La première rencontre a réuni 150 participants et les discussions ont porté sur 38 ensembles de données, dont certains étaient financés par la Kauffman Foundation, mais de nombreux ne l'étaient pas. Réduire les obstacles pour permettre aux universitaires « de mettre à l'essai » les ensembles de données a été une mesure très populaire qui s'est révélée efficace. En outre, le suivi des nouveaux événements dans le domaine de l'entreprenariat et des données sur l'innovation ainsi que la participation aux blogues sur ces événements nous ont permis d'en apprendre davantage sur nos propres projets et sur ceux menés par d'autres.

Bibliographie

Ballou, J., Barton, T., DesRoches, D., Potter, F., Zhao, Z., Santos, B. et Sebastian, J. (2007). Kauffman Firm Survey (KFS) Baseline Methodology Report. Janvier 2007. Disponible au http://ssrn.com/abstract=1024045

Ballou, J., Zhao, Z., DesRoches, D. et Potter, F. (2007). Meeting the Challenges of Designing the Kauffman Firm Survey: Sampling Frame, Definitions, Questionnaire Development, and Respondent Burden (Presentation). Juin 2007. Disponible au http://ssrn.com/abstract=1026393

Curtin, R. et Reynolds, P. (2007). Panel Study of Entrepreneurial Dynamics: History of the Research Paradigm. Octobre 2007. http://www.psed.isr.umich.edu/psed/download_document/1

Curtin, R. et Reynolds, P. (2009). U.S. Panel Study of Entrepreneurial Dynamics: PSED Overview. Août 2009. Disponible au http://www.kauffman.org/Blogs/DataMaven/August-2009/Slides-and-Materials-from-the-Academy-of-Managemen.aspx

Reedy, E. et Frazelle, S. (2009). Entrepreneurship, Innovation, and Economic Growth: A Review of Recent Studies, rapport non publié, Kansas City, Missouri: Kauffman Foundation. Juillet 2009.

Reynolds, P. (208). Current Population Survey: Expanding Understanding of U.S. Business Creation. *Proceedings of the 2008 Kauffman Symposium on Entrepreneurship and Innovation Data*. Novembre 2008. Disponible au http://www.kauffman.org/uploadedFiles/ResearchAndPolicy/EntrepreneurshipData/2008data/current-population-survey.pdf

Reynolds, P. et Curtin, R. (2008). Business Creation in the United States: Panel Study of Entrepreneurial Dynamics II Initial Assessment. *Foundations and Trends in Entrepreneurship*. Vol. 4, No. 3 (2008) 155–307.

Robb, A., Ballou, J., DesRoches, D., Potter, F., Zhao, Z. et Reedy, E. (2009) An Overview of the Kauffman Firm Survey: Results from the 2004–2007 Data. Avril 2009. Available at http://ssrn.com/abstract=1456380

Santos, B. et DesRoches, D. (2009). Exploring Mode Effects in a Panel Survey of New Businesses. Mai 2009. Disponible au http://ssrn.com/abstract=140936

SESSION 7 – ABORDER LA CONFIDENTIALITÉ À L'AIDE DE DONNÉES SYNTHÉTIQUES

La validité analytique et la protection de la confidentialité dans les systèmes de données statistiques à intégration longitudinale

John M. Abowd [1]

Résumé

Cet exposé résume les résultats de six différents projets de données synthétiques menés avec le soutien de la National Science Foundation et utilisant des données statistiques à intégration longitudinale tirées de recensements, d'enquêtes et de dossiers administratifs. Tous ces systèmes ont été conçus pour produire des microdonnées statistiquement valides à diffusion autorisée, protégées par des techniques de données synthétiques. Les systèmes étudiés comprenaient des données longitudinales sur les établissements, des données administratives à intégration longitudinale sur les employeurs et les employés, des données géospatiales intégrées sur le lieu de résidence et le lieu de travail, ainsi que des enquêtes-ménages intégrées à des données administratives longitudinales. Nous analysons les résultats de ces projets en ce qui concerne la validité analytique et la protection de la confidentialité.

[1] John M. Abowd, Cornell University, É.-U. (john.abowd@cornell.edu)

Sommaire des méthodes et évaluation provisoire de l'ensemble de données SIPP Synthetic Beta, version 5.0

Gary Benedetto, Martha Stinson et Melissa Bjelland[1]

Résumé

Cet exposé résume la méthodologie et l'évaluation de la qualité de la plus récente version de SIPP Synthetic Beta (SSB v5.0), ensemble de données à grande diffusion qui combine des variables tirées de la Survey of Income and Program Participation (SIPP) menée par le Census Bureau, des données chronologiques sur les gains tirées des déclarations de revenus de l'Internal Revenue Service (IRS), ainsi que les données de la Social Security Administration (SSA) sur les prestations de retraite et d'invalidité. On a utilisé l'imputation multiple et la synthèse partielle des données pour compléter et perturber les données afin que le produit de données finales (plusieurs ensembles de données appelés implicats, qui ont la même structure que les données confidentielles sous-jacentes) ne compromette pas la confidentialité des données. L'avantage des méthodes utilisées dans le cadre de ce projet est que les utilisateurs de données peuvent analyser chaque implicat synthétique exactement comme ils le feraient s'ils avaient accès à l'ensemble initial de données confidentielles. Après avoir obtenu des résultats pour chaque implicat synthétique, on dispose de formules relativement simples pour combiner ces résultats de manière à obtenir des estimations ponctuelles et des mesures de la variance qui sont pertinentes et qui tiennent compte de l'incertitude inhérente à la modélisation. De plus, comme les valeurs de la vaste majorité des variables du fichier ont été remplacées par des tirages aléatoires provenant d'une distribution de probabilité, les données partiellement synthétiques offrent un niveau très élevé de protection de la confidentialité. Nous tentons également d'évaluer la validité analytique des données partiellement synthétiques et de quantifier le risque de divulgation que comporte la diffusion de ces données au public.

[1] Gary Benedetto (gary.linus.benedetto@census.gov) et Martha Stinson (martha.stinson@census.gov), U.S. Census Bureau, É.-U.; Melissa Bjelland, Cornell University, É.-U.

La création de données synthétiques pour le Cross National Equivalent File

Cynthia Bocci et Jean-François Beaumont[1]

Résumé

Au cours des 15 dernières années, la création de données synthétiques comme méthode de prévention de la divulgation a suscité de plus en plus d'intérêt. Récemment, Statistique Canada a entrepris d'étudier des méthodes de création de donnée synthétiques pour le volet canadien du Cross National Equivalent File (CNEF). Ce volet canadien provient d'un sous-ensemble de variables de l'Enquête sur la dynamique du travail et du revenu, une enquête longitudinale réalisée par Statistique Canada. En raison de contraintes de confidentialité, l'accès aux données canadiennes ne peut avoir lieu qu'aux termes d'ententes spéciales. La création de données synthétiques faciliterait l'accès à ces données et, faut-il l'espérer, augmenterait leur utilisation. Dans le présent article, nous décrivons la méthode de création de données synthétiques longitudinales pour le volet canadien du CNEF et discutons des difficultés que pose la création de ménages synthétiques cohérents dans lesquels sont préservées autant que possible les relations observées entre les données originales tout en maintenant à un faible niveau le risque de divulguer des renseignements confidentiels. Nous présentons certains résultats.

Mots-clés : confidentialité, enquêtes-ménages, régression linéaire, régression logistique, bootstrap paramétrique.

1. Introduction

Au cours des 15 dernières années, la création de données synthétiques comme méthode de prévention de la divulgation a suscité de plus en plus d'intérêt. Récemment, Statistique Canada a entrepris d'étudier des méthodes en vue de créer des données synthétiques pour le volet canadien du Cross National Equivalent File (CNEF). Ce dernier comprend des données d'enquêtes par panel portant sur des variables définies de la même façon réalisées dans six pays. Le volet canadien du CNEF correspond à un sous-ensemble de variables démographiques, de variables d'emploi et de variables de revenu provenant de l'Enquête sur la dynamique du travail et du revenu (EDTR) réalisée par Statistique Canada. L'EDTR est une enquête longitudinale sur six années qui comprend des identificateurs uniques au niveau de la personne et des identificateurs annuels au niveau du ménage. En raison de contraintes de confidentialité, l'accès aux données canadiennes ne peut avoir lieu qu'aux termes d'ententes spéciales, contrairement aux données fournies par d'autres pays, qui sont recueillies par les universités ou des instituts privés. Par conséquent, les données canadiennes sont parfois omises dans les analyses. La création de données synthétiques faciliterait l'accès d'un plus grand nombre de chercheurs aux données et, faut-il l'espérer, augmenterait leur utilisation.

Dans le présent article, nous décrivons la méthode utilisée pour créer des données synthétiques longitudinales spécialement pour le volet canadien du CNEF et discutons des difficultés que pose la création de ménages synthétiques cohérents dans lesquels sont préservées autant que possible les relations entre les données originales, tout en maintenant à un faible niveau le risque de divulguer des renseignements confidentiels. La discussion est donc limitée à cette expérience particulière. Un développement plus étoffé de la théorie qui sous-tend notre approche, ainsi qu'une discussion plus approfondie de la littérature relevée dans ce domaine sont présentés dans un document inédit (Beaumont et Bocci, 2010).

Le plan de l'article est le suivant : à la section 2, nous présentons brièvement les concepts et la théorie sous-jacente. À la section 3, nous décrivons la production de données synthétiques. À la section 4, nous présentons les méthodes de validation et certains résultats pour la première année de données. À la section 5, nous discutons des questions ayant trait aux données longitudinales et à la section 6, nous concluons par un résumé et l'exposé de futurs travaux éventuels.

[1]Jean-François Beaumont, Statistique Canada, Division de la recherche et de l'innovation en statistique, 150, promenade Tunney's Pasture, immeuble R.-H.-Coats 16e étage, Ottawa, Ontario, Canada, K1A 0T6 (Jean-Francois.Beaumont@statcan.gc.ca); Cynthia Bocci, Statistique Canada, Division des méthodes d'enquêtes auprès des entreprises, 150, promenade Tunney's Pasture, immeuble R.-H.-Coats, 11e étage, Ottawa, Ontario, Canada, K1A 0T6.

2. Préliminaires

2.1 Hypothèse principale

Supposons que des données provenant d'une population finie sont générées conformément à un modèle donné. Soit $\boldsymbol{\theta}$ le vecteur de paramètres du modèle d'intérêt et $\hat{\boldsymbol{\theta}}(\mathbf{Y})$, l'estimateur de $\boldsymbol{\theta}$ en utilisant les données d'échantillon $\mathbf{Y}$. À cause de préoccupations concernant la confidentialité des données, ni $\mathbf{Y}$ ni $\hat{\boldsymbol{\theta}}(\mathbf{Y})$ ne peut être diffusé.

L'objectif des méthodes de création de données synthétiques est de générer une matrice échantillon $\mathbf{Y}^*$ afin d'obtenir l'estimateur synthétique $\hat{\boldsymbol{\theta}}(\mathbf{Y}^*)$ de $\boldsymbol{\theta}$. La matrice $\mathbf{Y}^*$ doit être générée de manière à préserver, dans la mesure du possible, les relations entre les variables originales. Nous générons $\mathbf{Y}^*$ en utilisant un modèle paramétrique multivarié, auquel on donne le nom de bootstrap paramétrique. La stratégie générale de génération des données synthétiques est expliquée à la section 3.1. Nous formulons l'hypothèse principale qui suit concernant les données synthétiques.

Hypothèse principale : Les deux premiers moments bootstrap de $\hat{\boldsymbol{\theta}}(\mathbf{Y}^*)-\hat{\boldsymbol{\theta}}(\mathbf{Y})$ sont asymptotiquement sans biais et convergents pour les deux premiers moments de $\hat{\boldsymbol{\theta}}(\mathbf{Y})-\boldsymbol{\theta}$ sous le modèle *m* (conditionnellement à l'échantillon *s*). De manière plus explicite, nous supposons que

i) $E_*\left(\hat{\boldsymbol{\theta}}(\mathbf{Y}^*)-\hat{\boldsymbol{\theta}}(\mathbf{Y})\middle| s,\mathbf{Y}\right)\approx E_m\left(\hat{\boldsymbol{\theta}}(\mathbf{Y})-\boldsymbol{\theta}\middle| s\right)\approx \mathbf{0}$;

ii) $E_m E_*\left\{\left(\hat{\boldsymbol{\theta}}(\mathbf{Y}^*)-\hat{\boldsymbol{\theta}}(\mathbf{Y})\right)^2 \mid s,\mathbf{Y}\right\}\approx E_m\left\{\left(\hat{\boldsymbol{\theta}}(\mathbf{Y})-\boldsymbol{\theta}\right)^2 \mid s\right\}$.

L'indice inférieur * fait référence à la distribution bootstrap. Cette hypothèse devrait tenir raisonnablement bien en pratique si le modèle sélectionné pour générer $\mathbf{Y}^*$ est satisfaisant. Il s'agit d'une hypothèse classique en théorie du bootstrap (paramétrique). La partie (i) de l'hypothèse signifie que la génération de données synthétiques n'introduit aucun biais significatif dans les estimateurs. La protection de la confidentialité est donc réalisée par un accroissement de la variance. La partie (ii) est requise uniquement pour la validité de nos estimateurs de variance proposés en (2).

2.2 Estimation de la variance

Dans la littérature sur les données synthétiques, plusieurs fichiers de données synthétiques sont habituellement créés. Supposons que *M* matrices de données synthétiques, $\mathbf{Y}^{*(1)},...,\mathbf{Y}^{*(M)}$, sont générées indépendamment et donnent lieu à *M* estimateurs synthétiques $\hat{\boldsymbol{\theta}}(\mathbf{Y}^{*(1)}),...,\hat{\boldsymbol{\theta}}(\mathbf{Y}^{*(M)})$. L'estimateur synthétique final est égal à la moyenne sur les *M* ensembles de données

$$\hat{\boldsymbol{\theta}}_M(\mathbf{Y}^*)=\frac{\sum_{i=1}^{M}\hat{\boldsymbol{\theta}}(\mathbf{Y}^{*(i)})}{M}.$$

Alors un estimateur de $\text{var}_{mp*}\{\hat{\boldsymbol{\theta}}_M(\mathbf{Y}^*)\}$, où l'indice inférieur *mp* fait référence à la distribution conjointe induite par le modèle *m* et le plan d'échantillonnage *p*, est donné par

$$V_M=\frac{1}{M}\sum_{i=1}^{M}v(\mathbf{Y}^{*(i)})+\frac{1}{M(M-1)}\sum_{i=1}^{M}\left\{\hat{\boldsymbol{\theta}}(\mathbf{Y}^{*(i)})-\hat{\boldsymbol{\theta}}_M(\mathbf{Y}^*)\right\}^2 \qquad (1)$$

où $v(\mathbf{Y}^{*(i)})$ est un estimateur de variance classique, tel que celui obtenu par linéarisation, bootstrap ou toute autre méthode de rééchantillonnage. La variance de l'estimateur (1) est exactement la même que celle proposée par Reiter (2003), à la différence près qu'il obtient $\mathbf{Y}^{*(i)}$ en utilisant l'imputation multiple et non le bootstrap. L'avantage de l'utilisation d'une procédure bootstrap est qu'il n'est pas nécessaire de refléter l'incertitude dans les estimateurs des paramètres du modèle, ce qui simplifie la production des données synthétiques.

En utilisant l'hypothèse principale, nous obtenons deux options de rechange pour l'estimateur de variance (1) qui sont données par

$$\tilde{V}_{1M} = \left(1+\tfrac{1}{M}\right)\frac{\sum_{i=1}^{M} v(\mathbf{Y}^{*(i)})}{M} \quad \text{et} \quad \tilde{V}_{2M} = \left(1+\tfrac{1}{M}\right)\frac{\sum_{i=1}^{M}\left\{\hat{\theta}(\mathbf{Y}^{*(i)}) - \hat{\theta}_M(\mathbf{Y}^*)\right\}^2}{M-1}. \qquad (2)$$

Un avantage pratique de $\tilde{V}_{1M}$ par rapport à $\tilde{V}_{2M}$ et V_M est qu'il peut être appliqué même quand $M = 1$. Habituellement, $\tilde{V}_{2M}$ et V_M requièrent un grand nombre M d'ensembles de données synthétiques pour être fiables. Comme il est souvent plus commode de produire et d'utiliser un seul fichier de données synthétiques, l'estimateur de variance $\tilde{V}_{1M}$ devient intéressant. Cependant, un inconvénient de l'utilisation d'une petite valeur de M est que l'on obtient ainsi un estimateur synthétique $\hat{\boldsymbol{\theta}}_M(\mathbf{Y}^*)$ moins efficace. Cela pourrait être le prix à payer pour protéger les données confidentielles contre la divulgation.

3. Création de données synthétiques

À la présente section, nous décrivons la méthodologie utilisée pour créer les données synthétiques pour le CNEF. Cette méthodologie comprend certains aspects de la modélisation, des éléments aléatoires et une stratégie de création de valeurs synthétiques cohérentes à l'intérieur d'un ménage.

Les données réelles du CNEF pour la première année de référence, 1999, portent sur environ 38 000 individus formant 15 000 ménages et contiennent des variables démographiques, ainsi que des variables d'emploi et de revenu. La tâche initiale consiste à créer une version synthétique de cet ensemble de données en gardant le même nombre d'individus et de ménages. Des quelque 40 variables du fichier, seuls l'état matrimonial, la relation avec l'indicateur de chef du ménage et une variable de poids ne sont pas synthétisés. Autrement dit, les valeurs réelles de ces trois variables sont gardées pour tous les individus afin de servir de fondement à l'ensemble de données synthétiques. Toutes les autres variables doivent être synthétisées pour tous les individus figurant dans l'ensemble de données synthétiques. Une fois que la création des données pour la première année est achevée, la tâche consiste à produire des ensembles de données longitudinaux pour les cinq autres années.

3.1 Méthodologie

En supposant que les observations sont indépendantes dans l'échantillon, la première étape de la génération des données synthétiques consiste à choisir une distribution multivariée paramétrique de $\mathbf{y}_k' = (y_{1k}, ..., y_{qk})$ conditionnellement à la présence de l'unité k dans l'échantillon s en utilisant les données d'échantillon $\mathbf{Y}$, où q est le nombre de variables d'intérêt qu'il faut synthétiser. Ensuite, on estime les paramètres inconnus de cette distribution multivariée et, enfin, on utilise la distribution ainsi estimée pour produire les données synthétiques. La tâche de modélisation multivariée est décomposée en q modèles conditionnels. Premièrement, on modélise et synthétise y_{1k} sachant $k \in s$. Puis, on modélise et synthétise y_{2k} sachant y_{1k} et $k \in s$, et ainsi de suite jusqu'à ce que soit modélisé et synthétisé y_{qk} sachant $y_{1k}, ..., y_{q-1,k}$ et $k \in s$. Puisque chacun de ces q modèles est conditionnel à l'échantillon, il n'est pas nécessaire d'utiliser les poids de sondage pour choisir les modèles. Cependant, le conditionnement sur l'information concernant le plan de sondage, ou au moins sur les poids de sondage, au moment de la sélection des modèles pourrait être une bonne stratégie. On s'assure ainsi que l'échantillonnage n'est pas informatif, c'est-à-dire que le conditionnement sur l'échantillon ne modifie pas le modèle multivarié inconditionnel qui est vérifié dans la population. Pour cela, on catégorise la variable de poids et on utilise ce poids catégorique dans le processus de modélisation.

La tâche de modélisation multivariée conditionnelle est répétée jusqu'à ce que toutes les variables soient synthétisées. L'ordre dans lequel elles le sont n'a pas d'importance si la modélisation est adéquate. Néanmoins, un classement est nécessaire et peut être déterminé en tenant compte de questions pratiques et en pensant aux analyses que pourraient exécuter les utilisateurs des données. Dans le cas du volet canadien du CNEF, le classement établi pour les variables à synthétiser est démographie, emploi, puis revenu. En général, les modèles multivariés réels utilisés dans notre application ne comportent pas de termes d'interaction d'ordre élevé. Tandis que la sélection proprement dite du modèle est un processus non pondéré s'appuyant sur des données réelles, l'utilisation subséquente de ces modèles pour générer des données synthétiques est un processus pondéré.

3.2 Préservation de la structure des ménages

Les données du CNEF contiennent des identificateurs personnels et des identificateurs du ménage, ce qui permet de produire des estimations au niveau de la personne ainsi qu'au niveau du ménage. Par conséquent, il faut créer des valeurs synthétiques individuelles qui concordent avec celles des autres membres du ménage. Pour préserver les relations entre les membres du ménage, nous synthétisons habituellement les valeurs pour les variables données comme il suit. Premièrement, nous synthétisons une valeur pour le chef du ménage. Ensuite, nous synthétisons une valeur pour le ou la partenaire du chef du

ménage, sachant la valeur pour ce dernier. Ensuite, nous synthétisons les valeurs pour les enfants, sachant les valeurs pour le chef du ménage et son ou sa partenaire. Enfin, nous synthétisons une valeur pour les autres membres du ménage, sachant la valeur pour le chef du ménage. Cette approche structurelle du ménage est semblable à celle adoptée par Reiter (2005).

3.3 Variables catégoriques

Le tableau 3.3-1 énumère les variables catégoriques qu'il faut synthétiser. En général, pour une variable catégorique, nous utilisons un modèle de régression logistique afin d'obtenir une probabilité d'appartenance à chaque catégorie. Pour un enregistrement synthétique donné, l'affectation aux catégories est faite aléatoirement en utilisant ces probabilités.

Tableau 3.3-1
Variables catégoriques à synthétiser

Type de variable	Variable
Démographique	sexe, race, niveau de scolarité, situation d'invalidité, état de santé, province
Emploi	nombre annuel d'heures travaillées, situation d'emploi, niveau d'emploi, activité principale, profession, code d'industrie à un chiffre, code d'industrie à deux chiffres

3.4 Variables quantitatives

Les variables quantitatives englobent deux variables démographiques et toutes les variables de revenu. Le tableau 3.4-1 énumère les variables quantitatives qu'il faut synthétiser.

Tableau 3.4-1
Variables quantitatives à synthétiser

Type de variable	Variable
Démographique	âge, nombre d'années d'études
Revenu	Revenus avant transferts gouvernementaux, après transferts gouvernementaux, revenu d'emploi, avoirs, pension de retraite privée, paiement forfaitaire, pension Transferts : privés, publics Impôts : fédéraux, provinciaux, autres, total

Pour les variables quantitatives démographiques, dans la plupart des cas, nous utilisons un modèle linéaire général pour produire des valeurs prévues pour les unités réelles ainsi que synthétiques, de même que les résidus pour les unités réelles. Les valeurs prévues sont regroupées pour former des classes. Ensuite, un élément aléatoire est introduit et la valeur est encore modifiée pour produire une valeur synthétique finale de la façon suivante. Pour chaque unité synthétique, nous choisissons au hasard un résidu appartenant à une unité réelle dans la classe prise en considération. Puis, nous ajoutons ce résidu à la valeur synthétique prévue et nous arrondissons la somme pour former une valeur synthétique finale. Les règles de vérification sont ensuite appliquées et si elles donnent lieu au rejet d'une unité synthétique particulière, le processus aléatoire est répété. Certaines modifications de la procédure de randomisation pour l'âge des partenaires et des enfants sont nécessaires afin de produire des valeurs de données synthétiques dont les distributions ressemblent de près à celles des valeurs des données réelles.

Les variables de revenu posent certains défis. Ces variables englobent les revenus, les taxes et les transferts, dont les corrélations doivent être maintenues afin que les données synthétiques soient utiles. Ces variables continues, dont la distribution est généralement asymétrique, sont également sujettes à des valeurs aberrantes qui compliquent la modélisation. De surcroît, la variance des valeurs des données est difficile à refléter en utilisant la méthode générale décrite plus haut pour les variables continues. Par conséquent, les variables de revenu sont créées différemment. Nous synthétisons d'abord les revenus d'emploi individuels, puis nous créons simultanément toutes les autres variables de revenu. Il convient de noter que les variables de revenu s'appliquent uniquement aux personnes de plus de 16 ans.

La variable de revenu d'emploi individuel requiert un traitement distinct selon que la personne est occupée ou non. Pour toutes les personnes occupées, nous utilisons un modèle linéaire général pour obtenir les valeurs prévues pour les unités synthétiques ainsi que réelles. Les revenus prévus pour les unités réelles ainsi que synthétiques sont regroupés pour former des classes. Pour chaque unité synthétique, nous sélectionnons une valeur du revenu d'emploi d'une unité réelle au hasard dans la classe et nous l'attribuons à l'unité synthétique. Pour les personnes en chômage, nous commençons par modéliser la probabilité d'avoir un revenu d'emploi individuel nul ou non nul. Pour les unités synthétiques en chômage pour lesquelles il est déterminé aléatoirement que le revenu d'emploi n'est pas nul en se basant sur ces probabilités, nous appliquons un modèle linéaire général des revenus d'emploi et créons une valeur synthétique de la même façon que dans le cas des personnes occupées.

Enfin, nous modifions les revenus d'emploi non nuls de toutes les unités synthétiques en effectuant un ajustement décrit à la section 3.4.1. Cet ajustement a pour but non seulement de réduire le risque de divulgation, mais aussi de forcer les premier et deuxième moments des valeurs des données synthétiques à concorder plus étroitement avec ceux des valeurs des données réelles.

Ensuite, toutes les autres variables de revenu sont synthétisées simultanément comme il suit. Dans les mêmes classes que celles utilisées pour créer les revenus d'emploi non nuls, nous identifions pour chaque unité synthétique un voisin le plus proche réel en ce qui a trait au revenu d'emploi, que nous appelons donneur. Puis, nous appliquons le ratio de la variable de revenu concernée au revenu d'emploi du donneur à la valeur synthétique finale du revenu d'emploi de l'unité synthétique. Pour les unités en chômage dont le revenu d'emploi est nul, nous définissons les classes en groupant les probabilités d'avoir un revenu nul pour les unités réelles ainsi que synthétiques. Ensuite, nous choisissons au hasard une unité réelle dans la classe et attribuons les valeurs réelles des variables de revenu restantes à l'unité synthétique.

Cette procédure s'applique aux variables de revenu au niveau de la personne. À part les revenus d'emploi individuels, toutes les autres variables de revenu figurant dans le fichier du CNEF sont estimées au niveau du ménage. Par conséquent, nous créons la valeur au niveau du ménage simplement en additionnant les valeurs de tous les membres du ménage pour une variable de revenu particulière.

3.4.1 Ajustement du revenu d'emploi individuel

L'analyse univariée pondérée des valeurs synthétiques du revenu d'emploi individuel avant un ajustement donne à penser, en particulier, que le deuxième moment n'est pas reflété adéquatement. Afin d'améliorer ce moment et de déguiser davantage les valeurs du donneur, nous modifions le revenu d'emploi synthétique par transformation linéaire sous deux contraintes. Les ajustements sont faits à l'intérieur d'un groupe g. Soit y une valeur réelle et y^*, une valeur synthétique avant l'ajustement. Définissons la valeur synthétique ajustée finale, y^{**}, pour l'unité $k \in g$ comme étant

$$y_k^{**} = \begin{cases} b_{1g} y_k^* & \text{si } y_k^* \geq 0 \\ b_{2g} y_k^* & \text{si } y_k^* < 0 \end{cases} \qquad (3)$$

pour des constantes réelles b_{1g} et b_{2g} telles que les valeurs y^{**} satisfont les contraintes

$$\frac{\sum_{k \in s_{g,syn}} w_k y_k^{**}}{\sum_{k \in s_{g,syn}} w_k} = \frac{\sum_{k \in s_{g,real}} w_k y_k}{\sum_{k \in s_{g,real}} w_k} \quad \text{et} \quad \frac{\sum_{k \in s_{g,syn}} w_k y_k^{**2}}{\sum_{k \in s_{g,syn}} w_k} = \frac{\sum_{k \in s_{g,real}} w_k y_k^2}{\sum_{k \in s_{g,real}} w_k},$$

où $s_{g,syn}$ et $s_{g,real}$ désignent l'échantillon synthétique appartenant au groupe g et l'échantillon réel appartenant au groupe g, respectivement. La transformation (3) est telle que la valeur synthétique ajustée garde le même signe que la valeur synthétique avant l'ajustement. Il est facile de trouver une solution analytique pour b_{1g} et b_{2g}. Les groupes de calage doivent parfois être manipulés pour s'assurer que les solutions de b_{1g} et b_{2g} ont des valeurs réelles. Ce type d'ajustement est semblable à un ajustement de type calage où ce sont ordinairement les poids qui sont modifiés plutôt que les valeurs de y.

3.5 Limites et questions pratiques

Un certain nombre de limites et de questions pratiques sont associées à la mise en œuvre de l'approche décrite plus haut. La qualité des données synthétiques dépend de la validité du modèle utilisé. Plus le nombre de variables à prendre en considération augmente, plus il devient difficile et peu pratique de trouver tous les termes d'interaction significatifs. En outre, les événements rares peuvent s'avérer difficiles à modéliser.

Une autre difficulté tient aux valeurs extrêmes dans les ensembles de données réelles. Bien qu'il soit possible de générer des valeurs extrêmes pour les variables quantitatives dans l'ensemble de données synthétiques, leur occurrence est peu probable. Des valeurs extrêmes peuvent résulter de différences entre les lois univariées des données réelles et des données synthétiques.

4. Validation empirique

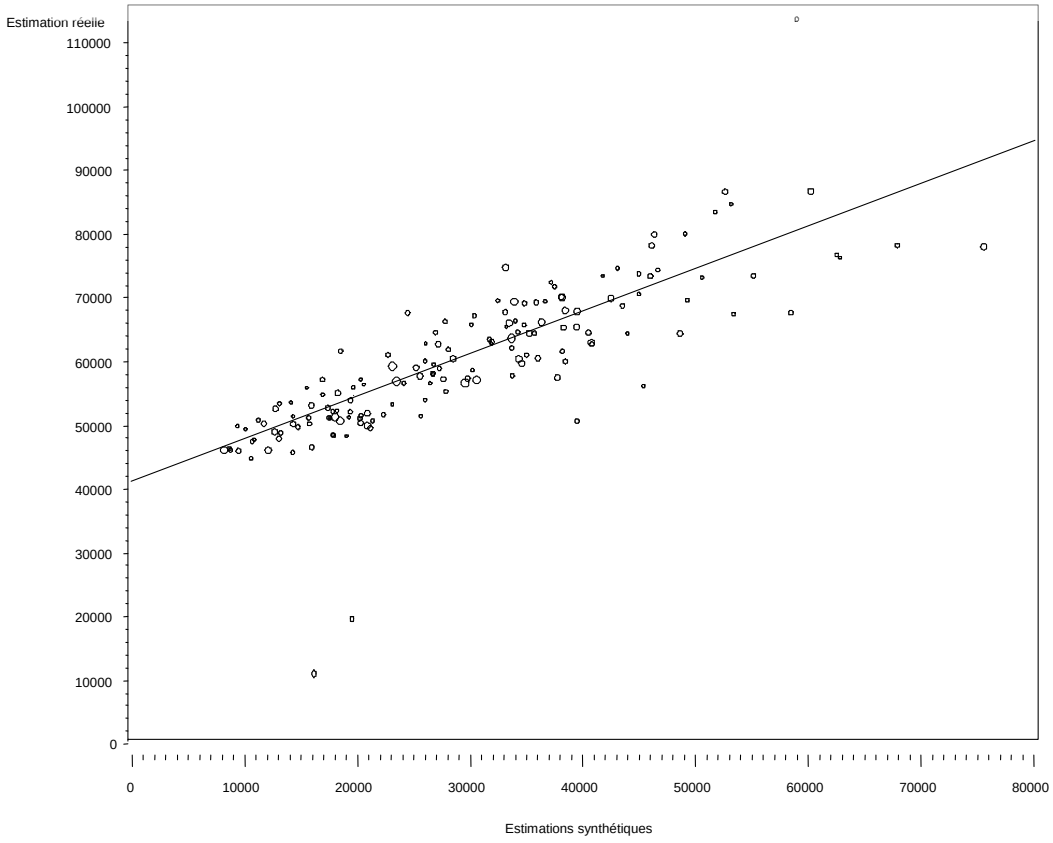
4.1 Validation

Tout au long du processus de création des données synthétiques, certaines validations ont été effectuées après la création de chaque variable synthétique. Diverses fréquences, totalisations croisées et distributions univariées ont été inspectées visuellement pour les données synthétiques ainsi que réelles. Quand les résultats synthétiques étaient insatisfaisants, les modèles ont été perfectionnés ou le mécanisme aléatoire a été modifié pour essayer d'éliminer autant d'incohérences que possible.

4.2 Certains résultats

Les comparaisons entre les données synthétiques créées pour une année et les données réelles correspondantes sont présentées ici. Les résultats sont donnés à titre d'illustration plutôt que pour commenter la qualité globale des données synthétiques.

Figure 4.2-1
Estimations du revenu d'emploi moyen selon l'âge, le sexe et le groupe professionnel



Il semble que nombre de corrélations entre les variables de revenu au niveau du ménage sont bien préservées dans l'ensemble de données synthétiques. Par exemple, les corrélations entre le revenu d'emploi total du ménage et le total des impôts du ménage pour les données synthétiques et réelles, après élimination d'une unité réelle aberrante, sont de 0,825 et 0,829 respectivement. De même, les corrélations entre le revenu de pension de retraite privée du ménage et les transferts de sécurité sociale du ménage pour les données synthétiques et réelles sont de 0,489 et 0,448 respectivement. Ensuite, les adultes de 16 à 69 ans ayant un revenu d'emploi individuel positif sont affectés à des groupes créés par recoupement de six catégories d'âge avec le sexe et avec 25 professions. La figure 4.2-1 donne le graphique des estimations synthétiques en fonction des estimations réelles du revenu d'emploi moyen selon le groupe pour des tailles de groupe supérieures à 30. La grandeur du cercle indique la taille de l'échantillon synthétique du groupe, tandis que la droite représente l'égalité entre les estimations réelles et synthétiques. Bien que les points ne se trouvent pas tous sur la droite, la médiane de l'écart relatif est de 7 %. Le coefficient de corrélation entre ces estimations réelles et synthétiques est de 0,86.

Pour les mêmes points de données, nous calculons la variance en utilisant $\tilde{V}_{1M}$ dans l'équation (2) et les poids bootstrap. Puis, nous calculons le pourcentage de chevauchement des intervalles de confiance; c'est-à-dire le chevauchement entre l'intervalle de confiance synthétique et l'intervalle de confiance réel par rapport à la longueur de l'intervalle de confiance réel. Théoriquement, la probabilité d'un chevauchement complet sous notre hypothèse principale, les hypothèses de normalité et en n'utilisant qu'un seul ensemble de données synthétiques est de 58 % (Beaumont et Bocci, 2010). Dans le présent exemple, la proportion de chevauchement complet est de 41 %, ce qui n'est pas trop éloigné de la valeur théorique.

5. Considérations longitudinales

Le but ultime est de créer un fichier synthétique longitudinal pour le volet canadien du CNEF. Le processus de création d'un fichier synthétique pour toute année *t* après la première vague est triple. Premièrement, certaines particularités de la source des données du volet canadien du CNEF compliquent la détermination exacte des individus réels qui doivent être dans le fichier réel au temps $t+1$ en vue de générer l'ensemble de données synthétiques pour l'année $t+1$. La deuxième étape consiste à appliquer à l'ensemble de données synthétiques des changements d'état matrimonial et de structure du ménage d'une année à l'autre semblables à ceux observés dans l'ensemble de données réelles. Enfin, une fois que les individus et les ménages synthétiques sont déterminés, le long processus de synthèse de toutes les variables peut débuter. Il convient de noter qu'il n'y a aucune non-réponse dans l'ensemble de données synthétiques, puisque les valeurs synthétiques sont créées pour toutes les personnes figurant dans le fichier. Ces trois étapes sont décrites brièvement aux sous-sections 5.1 à 5.3.

5.1 Création d'un ensemble de données réelles dans une année longitudinale

Rappelons que certaines variables de l'ensemble de données du CNEF sont estimées au niveau de la personne tandis que d'autres le sont au niveau du ménage. Notre première étape consiste à créer soigneusement un fichier de données réelles de travail au niveau de la personne contenant des identificateurs du ménage pour refléter i) les décès d'individus longitudinaux, ii) les nouveaux ménages créés par des unités longitudinales qui se sont séparées d'un ménage existant l'année précédente, iii) les nouveaux individus qui englobent les nouveaux entrants dans le ménage de l'unité longitudinale et iv) les cohabitants de retour vivant dans le ménage d'une unité longitudinale. Même si nous ne suivons et ne retenons en fin de compte que l'unité longitudinale, nous devons suivre chacun des membres du ménage, y compris ceux présents dans les groupes iii) et iv) afin de calculer correctement les variables au niveau du ménage pour une année donnée. Un fichier contenant les unités longitudinales, ainsi que les cohabitants est appelé ici un fichier de travail.

5.2 Génération des changements dans les ménages

La création d'une base de données synthétiques pour l'année $t+1$, contenant le même nombre d'individus et de ménages que l'ensemble de données réelles de la même année, débute par l'application des changements de situation et de structure à l'ensemble de données synthétiques de l'année *t*. Les changements de situation et de structure sont causés en partie par les naissances, les décès, les fractionnements des ménages ou des combinaisons de ces trois éléments. Le fichier de travail synthétique doit subir le même type de changement approximativement aux mêmes fréquences que le fichier de travail réel. Pour cela, des classes sont construites soigneusement par croisement de certaines variables au niveau du ménage pour les ménages d'une taille donnée. Les classes sont affectées aux ménages réels ainsi que synthétiques d'après les valeurs des variables de l'année *t* et doivent être suffisamment fines pour produire les changements appropriés. Pour chaque ménage synthétique dans une classe donnée à l'année *t*, un ménage réel est sélectionné au hasard dans la même classe. Tous les changements structurels, si tant est qu'il y en ait, subis par le ménage donneur réel sont imposés aux ménages synthétiques, ce qui crée le nombre correct d'individus et de ménages synthétiques. En outre, les valeurs individuelles d'état matrimonial et de relation avec l'indicateur de chef du ménage du donneur pour l'année $t+1$ sont gardées pour servir de fondement au fichier de données synthétiques de travail.

5.3 Création des valeurs de données longitudinales

L'ensemble de données longitudinales contient les mêmes variables pour chacune des six années de référence et essaye de traduire les relations au cours du temps. En utilisant le même classement des variables que l'année précédente, chaque variable est synthétisée en grande partie de la même manière que pour la première année, excepté que, pour les unités longitudinales, les variables de l'année précédente peuvent maintenant servir de variables explicatives éventuelles dans la modélisation. En particulier, nous choisissons un modèle où la j^{e} variable synthétique, $j = 1,...,q$, pour l'année $t+1$ est une fonction des j-1 première variable synthétique dans l'ensemble de données de l'année $t+1$ et les q-(j-1) dernières variables synthétiques de l'ensemble de données de l'année précédente *t*. Cette approche signifie que nous ne considérons que deux années consécutives

pour créer tout ensemble de données autres que le premier, et que nous limitons les variables explicatives possibles à un nombre pratique.

Pour les nouveaux cohabitants qui apparaissent dans le fichier de travail synthétique, aucune information provenant des années précédentes n'est disponible et les valeurs des données synthétiques sont créées de la même façon qu'ont été créées les valeurs synthétiques de la première année.

6. Conclusion

Jusqu'à présent, nous n'avons créé des données synthétiques que pour une année. Les analyses préliminaires donnent à penser que les données synthétiques créées jusqu'à présent ont une certaine utilité et une stratégie longitudinale a donc été établie dans le but de compléter un ensemble de données longitudinales portant sur six années pour le volet canadien du CNEF. Grâce à la modélisation séquentielle et à la randomisation, nous espérons que les valeurs synthétiques individuelles différeront suffisamment des valeurs réelles pour que le risque de divulgation soit minime, mais que les distributions multivariées des variables demeurent satisfaisantes.

Les futurs travaux comprendront l'achèvement de l'ensemble de données longitudinales et une analyse plus approfondie de la qualité des données synthétiques, qui comprendra des mesures globales de l'utilité des données pour les données synthétiques développées récemment par Woo et coll. (2009). On pourrait aussi étudier l'usage plus important de méthodes non paramétriques en vue de mieux saisir les relations entre les données réelles, quoique ces méthodes puissent accroître le risque de divulgation.

Bibliographie

Beaumont, J.-F. et Bocci, C., (2010). Some Theory on Synthetic Data Generation and Its Application to the Canadian Portion of the Cross National Equivalent File, document non puplié, Ottawa, Canada: Statistique Canada.

Reiter, Jerome P., (2005). Releasing Multiply Imputed, Synthetic Public Use Microdata: an Illustration and Empirical Study, *Journal of the Royal Statistical Society A*, 168, Part 1, pp. 185-205.

Reiter, Jerome P., (2003). Inférence pour les ensembles de microdonnées à grande diffusion partiellement synthétiques, *Techniques d'enquête*, 29, No. 2, pp. 203-211.

Woo, M.-J., Reiter, J. P., Oganian, A. et Karr, A.F., (2009). Global Measures of Data Utility for Microdata Masked for Disclosure Limitation, *The Journal of Privacy and Confidentiality*, 1, No. 1, pp.111-124.

SESSION 8 – DISCOURS DU GAGNANT DU PRIX WAKSBERG

Méthodes pour le suréchantillonnage des sous-populations rares dans les enquêtes sociales

Graham Kalton[1]

Résumé

On requiert de plus en plus des enquêtes sociales qu'elles permettent de produire des estimations pour des sous-populations, et souvent pour des sous-populations rares. De temps à autre, une enquête se concentre sur une seule sous-population, mais dans de nombreux cas, l'enquête doit produire des estimations pour plusieurs sous-populations de même que pour l'ensemble de la population. Lorsque l'appartenance à une sous-population rare peut être déterminée à partir de la base de sondage, il devient relativement simple de choisir un échantillon de la bonne taille. Dans ce cas, le problème principal consiste à déterminer l'ampleur du suréchantillonage à employer lorsque l'enquête vise à produire des estimations pour plusieurs sous-populations et l'ensemble de la population. Suréchantillonner une sous-population rare qui ne peut être identifiée à partir de la base de sondage s'avère un défi majeur. Les méthodes permettant de procéder à ce suréchantillonnage incluent l'échantillonnage stratifié disproportionné, l'échantillonnage à deux phases, l'utilisation de bases multiples, le sondage par réseaux, les enquêtes par panel, et l'emploi d'enquêtes polyvalentes. Cet article décrit ces méthodes et illustre leur application dans une gamme d'enquêtes.

Cet article est publié dans Techniques d'enquêtes, décembre 2009.
http://www.statcan.gc.ca/pub/12-001-x/2009002/article/11036-fra.pdf

[1] Graham kalton, Westat, É.-U. (grahamkalton@westat.com)

SESSION 9 – DONNÉES LONGITUDINALES SUR LA SANTÉ: ENJEUX ET DÉFIS

Établissement d'une méthode longitudinale de recherche sur la santé communautaire : enjeux et difficultés

David Marshall PhD[1]

Résumé

Les méthodes longitudinales peuvent être extrêmement utiles quand on cherche à mieux comprendre la santé et le bien-être des collectivités. En particulier, les études longitudinales axées sur les localités permettent de prendre davantage conscience des conditions sociales, économiques et environnementales locales et de leur influence sur le fardeau de la maladie chronique. Toutefois, jusqu'à présent, elles ont été relativement sous-utilisées dans ce domaine. Le présent article décrit certaines questions méthodologiques prises en considération pour établir l'« Ipswich Study », une étude longitudinale axée sur la localité qu'il est proposé de lancer dans le Queensland, en Australie. L'évolution rapide de la région et sa croissance prévue offrent une occasion unique d'analyser les effets du changement sur la santé communautaire. L'article porte en particulier sur les problèmes éventuels que peut poser l'étude d'une région qui subit d'énormes changements sociaux, économiques et environnementaux.

Mots clés : santé communautaire, axée sur la localité, maladie chronique, changement socioéconomique.

1. Introduction

Le présent article discute des questions que soulève une proposition d'étude longitudinale intitulée l'« Ipswich Study ». Généralement parlant, il s'agira d'une étude longitudinale, axée sur la localité, de la santé et du bien-être dans la collectivité d'Ipswich, une région suburbaine périphérique du Queensland, en Australie. Ipswich est située à environ 40 km au sud-est de Brisbane, la capitale du Queensland. Brisbane et Ipswich sont toutes deux situées dans le sud-est du Queensland, région qui a connu des taux extrêmement élevés de croissance démographique ces dernières décennies, particulièrement dans le secteur d'Ipswich. Le but du présent article est de décrire certains défis logistiques et méthodologiques que pose l'établissement de l'étude.

2. Questions contextuelles générales

Les circonstances contextuelles jouent un rôle essentiel dans le choix de l'approche adoptée pour établir l'Ipswich Study. Non seulement elles fournissent le contexte général dans lequel l'étude se déroulera, mais exercent aussi une influence profonde sur l'élaboration logistique et méthodologique de l'étude. Même si certaines de ces questions sont spécifiques à l'Ipswich Study, elles sont vraisemblablement semblables aux défis soulevés par d'autres méthodes longitudinales.

2.1 Intervenants et financement

L'Ipswich Study découle de la création du Healthy Communities Research Centre (HCRC) sur le campus d'Ipswich de l'Université du Queensland. Le HCRC est une initiative collective de l'Ipswich Hospital Foundation (IHF) et de l'Université. L'étude et ses objectifs sont donc fortement influencés par les intérêts des deux organismes, qui, quoique similaires, ne sont pas parfaitement concordants. L'IHF a pour mission de promouvoir les modes de vie favorables à la bonne santé, d'offrir un soutien aux professionnels de la santé et de construire une collectivité en meilleure santé à Ipswich. Donc, la recherche accomplie dans la collectivité d'Ipswich et en collaboration avec celle-ci produisant des résultats qui ont une application pratique au niveau local est très importante. Par contre, l'Université souhaite effectuer des travaux de recherche de calibre international qui amélioreront l'image du HCRC ainsi que de l'Université dans l'ensemble. Il est donc nécessaire de concevoir une étude qui permet de répondre à des questions particulières d'intérêt local, mais qui fait progresser la connaissance du domaine sur la scène internationale.

[1]David Marshall, Healthy Communities Research Centre, University of Queensland, Salisbury Road, Ipswich Australia 4305.

En plus de ces deux intervenants, il importe de tenir compte des intérêts d'intervenants éventuels, en grande partie parce qu'à l'heure actuelle, l'Ipswich Study n'est pas financée. Le financement offert par les intervenants actuels suffit à l'élaboration de l'étude, mais ne s'étendra pas à sa mise en œuvre. Même si les sources éventuelles de financement de l'étude sont nombreuses (par exemple, divers paliers de gouvernement), tant que les fonds ne sont pas engagés, il est nécessaire de concevoir l'étude de manière à offrir un grand degré de flexibilité en ce qui concerne les sujets étudiés, les échéanciers et les budgets, lesquels ont tous une incidence sur la méthodologie.

2.2 Priorités de recherche, hypothèses et questions

Si les questions concernant les intervenants et le financement du projet présentent des défis d'ordre à la fois méthodologique et logistique, elles influencent aussi l'élaboration d'hypothèses et de questions de recherche. Précisément, comme il est nécessaire que l'étude soit souple pour permettre l'adaptation, l'élaboration de l'Ipswich Study n'est pas guidée par des hypothèses ou des questions de recherche particulières, mais est dictée par les priorités plus générales des intervenants et du HRCR. Cette approche a des aspects positifs, mais elle soulève aussi des défis en ce qui concerne l'élaboration méthodologique de l'étude. Le défi consiste à pouvoir situer l'étude dans un créneau possédant des limites thématiques clairement évidentes, mais qui offre simultanément la souplesse de pouvoir répondre à des questions de recherche particulières à mesure qu'elles émergent ou sont posées par les intervenants. Il s'agit donc de trouver le juste équilibre entre le général et le spécifique.

La grande priorité du HCRC est de mieux comprendre les déterminants sociaux et contextuels de la santé au niveau tant individuel que communautaire. Ce genre de déterminants englobe essentiellement les déterminants non médicaux et non individuels de la santé, en particulier ceux ayant trait aux problèmes de santé chroniques qui peuvent être prévenus. Ils comprennent, entre autres, l'environnement bâti et l'environnement social, les politiques et les programmes qui influent sur la collectivité, les ressources économiques locales et les conditions environnementales générales. Cet ensemble d'objectifs, ainsi que la nécessité de créer un programme de recherche tenant compte à la fois des besoins de niveau local de l'IHF et du mandat international de l'Université ont abouti à la proposition d'une étude longitudinale de la santé et du bien-être à Ipswich dont la vision est la suivante :

- examiner l'interface entre les divers déterminants de la santé;
- intégrer ce que l'on sait des influences contextuelles et environnementales sur la santé des collectivités;
- fournir des éclaircissements sur les différences en matière de santé dans Ipswich et entre les collectivités reflétant diverses caractéristiques et divers environnements contextuels, au fil du temps;
- évaluer comment les interventions et les modifications apportées à la collectivité dans son ensemble se manifestent et comment elles sont vécues localement.

3. Principales questions méthodologiques

En plus des questions contextuelles qui ont une incidence importante sur l'élaboration de l'Ipswich Study, les questions de nature « longitudinale » ayant trait à la santé et au bien-être communautaire ont soulevé de nombreux défis. Ces derniers englobent des problèmes concernant les domaines spécialisés ainsi que des problèmes logistiques et définitionnels qui, en dernière analyse, influent considérablement sur la conception méthodologique globale. Ces défis sont décrits ici.

3.1 Ipswich, une collectivité qui évolue rapidement

Étant donné que l'Ipswich Study doit être réalisée pendant une longue période, comme dans le cas de toutes les études longitudinales, une question essentielle est de veiller à ce que l'étude demeure pertinente au fil du temps. Alors qu'il s'agira toujours d'un défi pour une étude longitudinale axée sur les localités, à Ipswich, le problème pourrait être plus important, étant donné les changements rapides qui surviennent dans la région. Le sud-est du Queensland est l'une des régions du pays dont la croissance est la plus rapide depuis de nombreuses décennies, et il est prévu que cela continue. À l'heure actuelle, la population de la région d'Ipswich est de l'ordre de 150 000, mais on projette qu'elle triplera en 25 ans (State of Queensland 2009). En réponse à ces projections, le gouvernement du Queensland a dressé un plan régional au centre duquel Ipswich est fermement placée. En vertu de ce plan, un certain nombre de villages urbains ou communautés clôturées sont en construction ou désignés pour être développés. Par conséquent, l'Ipswich Study doit permettre de couvrir cette croissance qui, par la voie de nouveaux groupes de population, de l'évolution des caractéristiques démographiques et d'une nouvelle infrastructure, aura un impact important sur la région. Concevoir une étude qui peut tenir compte de nouveaux groupes et de l'évolution contextuelle est donc essentiel. Cependant, ce genre de changements offre aussi l'occasion de comprendre leur effet au cours du temps. Autant que nous sachions, aucune étude permettant de comprendre l'effet de changements environnementaux, sociaux et économiques importants sur la santé et le bien-être des individus et des collectivités n'a été réalisée à l'échelle internationale. Il s'agira d'un

objectif clé de l'Ipswich Study, mais créer la capacité d'intégrer les changements qui surviennent présentera de nombreux défis méthodologiques.

3.2 Variables individuelles et écologiques

Un autre défi consistera à s'assurer que l'étude permette de se pencher sur une vaste gamme de questions mentionnées dans la littérature comme en rapport avec la santé communautaire et individuelle à un niveau local. Une revue de la littérature sur les déterminants sociaux de la santé nous a permis de dégager deux courants de recherche contemporains que nous avons considérés comme étant au centre de nos priorités, à savoir l'approche du cycle de vie et l'approche écologique. L'approche des résultats en matière de santé axée sur le cycle de vie tient compte des parcours que suivent les individus et de l'exposition de ces derniers aux facteurs de risque et aux facteurs de protection au cours du temps. L'approche écologique, quant à elle, s'appuie sur le fait bien prouvé que l'endroit où vit une personne peut avoir sur la santé et le bien-être de celle-ci un effet nettement plus important que les caractéristiques, les expériences, les intentions et les comportements individuels. L'Ipswich Study visera à unifier ces deux thèmes de recherche en une seule étude.

Essentiellement, ces deux thèmes sont apparus à cause d'un manque distinct de travaux de recherche en vue de déterminer la mesure dans laquelle le lieu de résidence aux diverses étapes du cycle de vie peut influer sur la santé (Leyland et Næss, 2009). Comme l'observe Berkman (2009), le fait que ces deux écoles de pensée aient évolué séparément l'une de l'autre explique pourquoi les études aboutissent fréquemment à des conclusions différentes quant aux risques que posent pour la santé les circonstances sociales et comportementales. Afin de comprendre entièrement les effets régionaux sur la santé, il est donc nécessaire de faire intervenir à la fois les facteurs individuels et écologiques (Moon et coll., 2005). De nombreux auteurs ont donc demandé que l'on suive des approches méthodologiques qui combinent celles du cycle de vie et de l'approche écologique en vue de comprendre la santé et le bien-être. Ce genre de recherche aura plus de chance de mieux saisir les relations de causalité (Bauman 2005). L'Ipswich Study adoptera cette approche et :

- intégrera des renseignements sur les caractéristiques de la santé des individus ainsi que des collectivités;
- recueillera des données probantes issues des lieux, des circonstances et des environnements dans lesquels vivent les individus;
- tiendra compte des adaptations et des décisions individuelles qui ont lieu dans des circonstances diverses et changeantes;
- accordera de l'attention aux transitions au cours du cycle de vie, et à l'évolution de la nature et de la forme des collectivités.

Dans ce contexte, le plan de recherche comprendra la collecte de données portant à la fois sur les individus et leur environnement, ce qui pose de nombreux défis. Par exemple, les études cherchant à évaluer des modèles écologiques dans une région locale doivent être capables de permettre des variations aréolaires des variables indépendantes d'intérêt. En outre, dans une étude longitudinale, il est également nécessaire d'inclure des variables qui évoluent au fil du temps et pour lesquelles le changement est effectivement décelable. Il s'agit toutefois d'un des principaux avantages de l'Ipswich Study, parce que l'on sait d'avance qu'Ipswich évoluera spectaculairement et rapidement. Il s'agit donc là d'une occasion idéale d'établir une étude de cohorte conçue en vue de comprendre l'effet des changements contextuels sur la santé à mesure qu'ils se produisent.

3.3 Facteurs constitutionnels, contextuels et collectifs

De surcroît, alors que la décision d'embrasser à la fois l'approche du cycle de vie et l'approche écologique situera définitivement l'étude à l'avant-garde de la recherche sur la santé communautaire, d'énormes défis logistiques se posent immédiatement, principalement en raison du très grand nombre de variables du cycle de vie et de variables écologiques mentionnées dans la littérature comme étant en relation avec les résultats en matière de santé et de bien-être. On cherchera certes à inclure dans l'étude autant de variables essentielles fondées sur des données probantes que possible, mais il est également crucial que l'étude demeure gérable et que l'on n'impose pas un fardeau excessif aux participants à cause d'un trop grand nombre de questions et de protocoles méthodologiques complexes.

Il existe donc une vaste gamme de facteurs couvrant à la fois les attributs physiques et sociaux des quartiers qui pourraient vraisemblablement affecter la santé des individus (Diez-Roux, 2007). Toutefois, comme l'étude s'intéresse aux localités et à l'évolution de ces dernières, la sélection des variables peut être limitée au tirage dans trois catégories de variables environnementales souvent mentionnées dans la littérature écologique. Ces catégories sont les suivantes :

- Facteurs constitutionnels – constitution de la localité telle qu'elle est déterminée par les caractéristiques des personnes qui y vivent (par exemple, caractéristiques socioéconomiques et démographiques);
- Facteurs contextuels - variables physiques et non physiques de l'environnement (par exemple, environnements bâti, naturel et social qui englobent, entre autres, les bâtiments, les politiques et les programmes);

- Facteurs collectifs – caractéristiques qui existent au-delà des individus, mais qui ne sont pas des facteurs contextuels tangibles non plus (par exemple, normes sociales, capital social et efficacité collective).

Il sera nécessaire de sélectionner des indicateurs de chacune de ces trois catégories de variables, mais aussi de reconnaître qu'il ne sera pas simple de recueillir des données sur toutes les variables comprises dans chaque catégorie. Par exemple, alors que certaines variables peuvent être mesurées de manière objective en recourant à l'observation, à des interviews et à des sources de données existantes, nombre d'entre elles pouvaient être de nature hautement subjective, et donc, différer de localité en localité, de personne en personne et au cours du temps. Le cas échéant, c'est souvent la perception plutôt que la réalité qui influence le plus le comportement, et donc, les résultats en matière de santé. Dans certains cas et pour certains types de variables, il serait peut-être prudent, voire même nécessaire, de recueillir à la fois des données objectives et subjectives se rapportant à la même variable. Par exemple, les taux réels de criminalité ainsi que la peur de la criminalité, telle qu'elle est perçue par les résidents, pourrait avoir des effets indépendants sur le comportement, donc sur les résultats en matière de santé. Dans le cadre de l'Ipswich Study, il est nécessaire d'identifier les variables les plus importantes pouvant être légitimement mesurées pour lesquelles les changements spatiaux et temporels sont décelables.

3.4 Relations dynamiques et réflexives

La détermination d'un ensemble clé de variables et d'indicateurs se complique encore davantage du fait qu'il est de plus en plus reconnu que les relations réflexives entre les individus et leur environnement doivent être examinées minutieusement afin de comprendre comment les circonstances contextuelles affectent la santé. En effet, les relations entre les individus et les circonstances dans lesquelles ils vivent sont à la fois bidirectionnelles et dynamiques. Autrement dit, chaque action ou chaque expérience vécue par un individu est, d'une certaine façon, influencée par l'environnement dans lequel se déroule l'action, mais l'action a également un effet dans l'autre direction. Par conséquent, les relations entre les personnes et leurs circonstances contextuelles et environnementales ne doivent pas être traitées comme passives ou statiques si l'on veut arriver à les comprendre pleinement. Contrairement aux études transversales qui produisent un instantané à un point particulier dans le temps, les études longitudinales portant sur les relations entre les personnes, leur comportement et leur environnement au cours du temps doivent permettre de comprendre la nature réflexive et dynamique de ces relations.

Le défi consiste à concevoir une étude permettant de recueillir et d'analyser des données qui reflètent les relations dynamiques susmentionnées. Afin de mieux comprendre ce genre de relations, il est vraisemblable que des données qualitatives ainsi que quantitatives seront nécessaires pour saisir les processus de prise de décisions qui influencent ces relations. Cela permettrait d'apporter des éclaircissements au sujet des mécanismes qui influent sur les relations entre les localités et la santé des gents qui y vivent.

3.5 Définition de la collectivité et du quartier

La dernière question importante soulevée dans le présent article est celle, toujours présente, consistant à définir ce que l'on entend par collectivité ou par quartier. Cela est tient au fait que ce qui constitue un quartier ou une collectivité peut varier en fonction de divers facteurs, dont la nature de la localité, la personne ou le groupe concerné ou la question d'intérêt. Par exemple, les jeunes peuvent avoir une idée de la collectivité locale fort différente de celle des personnes plus âgées. De même, ce qui constitue une collectivité pourrait varier selon le quartier, la compréhension de ce qu'est une communauté étant plus restreinte et plus rigide dans certains quartiers que dans d'autres. Considérées sous l'angle d'un service ou d'une installation, les définitions seront influencées par la clientèle, la charte de l'organisme ou ce qu'il existe d'autre dans le secteur.

Un autre élément critique est la reconnaissance que ces questions peuvent varier non seulement de localité en localité et de variable en variable, mais aussi au cours du temps. Cet élément est essentiel parce que différentes variables pourraient influer sur la santé à des degrés divers, mais pourrait également varier chez les individus et dans les collectivités au cours du temps. Par exemple, si l'on pose aux participants à l'étude des questions sur les problèmes dans la « zone locale », nous ne devons pas perdre de vue que ce qui constitue cette « zone locale » peut évoluer au cours de l'étude.

Dans le contexte d'une étude écologique transversale, l'approche simple consiste à utiliser des définitions préexistantes et des données connexes qui ont été recueillies pour des raisons administratives ou autres. Cependant, bien qu'elles soient faciles à obtenir et à utiliser, ces données ne tiennent pas entièrement compte de la diversité des collectivités et des quartiers dans lesquelles les personnes vivent. En outre, les données de ce genre pourraient ne pas être suffisamment détaillées pour déceler les variations en cas de différences entre de petits domaines, et elles ne permettent certainement pas de déceler l'évolution temporelle au niveau du petit domaine. Notre défi consiste à pouvoir utiliser des définitions de la collectivité et du quartier qui reflètent cette diversité.

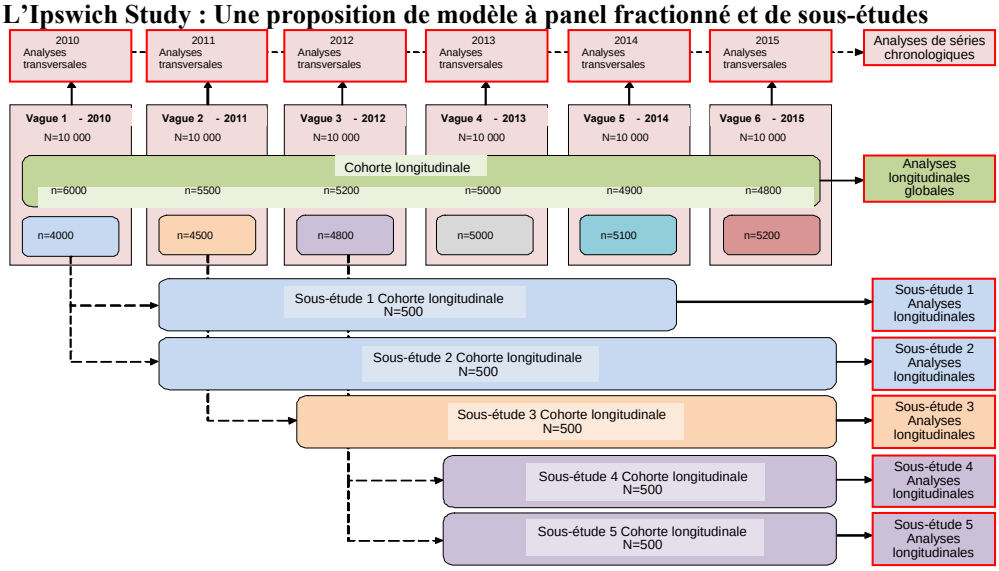
4. Discussion et conclusion

Malgré toutes les incertitudes et les paramètres inconnus dans le contexte desquels est élaborée l'Ipswich Study, il est nécessaire d'amener le projet à un niveau avancé de préparation en prévision du moment où le financement deviendra disponible. Par conséquent, une structure méthodologique générale a été élaborée et proposée pour aider le projet à réaliser ses nombreux objectifs, à satisfaire les divers intervenants et à permettre d'étudier une vaste gamme de sujets liés à la santé, ainsi que l'influence des contextes environnementaux sur le comportement lié à la santé au cours du temps. Ce genre de modèle doit être adaptable et capable de demeurer pertinent au niveau tant local qu'international au cours du temps.

4.1 Une proposition de modèle méthodologique

La figure 4.1-1 dépeint un modèle qu'il est envisagé d'utiliser pour exécuter l'Ipswich Study. Le modèle s'appuie sur un plan à panel fractionné et, pour les besoins de l'illustration, a été conçu en se servant de six vagues annuelles de collecte des données et d'une cohorte annuelle de 10 000 personnes. Une composante importante du modèle correspond aux sous-études qui, au lieu d'être ajoutées par après, sont intégrées dans le cadre de modélisation en tant que composantes essentielles conçues pour permettre de relever plus facilement nombre des défis mentionnés dans le présent article.

Figure 4.1-1
L'Ipswich Study : Une proposition de modèle à panel fractionné et de sous-études



Au cœur du panel fractionné se trouve une cohorte longitudinale qui, après être recrutée, sera retenue pendant une longue période. Au départ, les données recueillies auprès de ce groupe auront une grande portée, mais elles seront moins détaillées – elles auront trait à une vaste gamme de questions et de priorités liées à la santé, sans beaucoup approfondir bon nombre d'entre elles. Cette cohorte de participants ne sera pas remise à un niveau au cours du temps, et il ne sera demandé à aucun des membres de participer à une sous-étude. Toutefois, il est reconnu qu'après un certain temps, il pourrait être prudent de recruter une nouvelle cohorte longitudinale qui sera étudiée en parallèle.

L'autre composante du panel fractionné est le recrutement annuel des nouveaux participants. Ce groupe complétera la cohorte de base de plusieurs façons et, en bout de ligne, contribuera à la réalisation des multiples objectifs de l'étude et à la satisfaction des besoins des intervenants. En particulier, comme la région évolue rapidement, le recrutement transversal annuel donnera la chance de recruter des personnes qui, pour diverses raisons, sont sous-représentées dans la cohorte principale. Il serait alors possible de produire des estimations transversales plus exactes en vue de répondre, à chaque vague, aux besoins concernant l'élaboration des politiques et des programmes de l'IHF. En outre, les personnes recrutées chaque année pour faire partie de l'échantillon transversal de la vague en question seraient utilisées pour sélectionner les participants à de nombreuses sous-études. Les sous-études pourraient être transversales ou longitudinales, et qualitatives ou quantitatives. Néanmoins, elles seraient toutes utilisées pour examiner des questions particulières de manière beaucoup plus approfondie que ne permet de le faire la cohorte longitudinale de portée plus générale. Par exemple, une sous-étude pourrait se concentrer sur un groupe démographique particulier (disons les jeunes), tandis qu'une autre pourrait porter sur un problème de santé particulier. Un grand avantage de cette approche est que le nombre de sous-études ne sera pas limité par la nécessité de réduire au minimum le fardeau de la cohorte longitudinale. Au moment de chaque vague, on disposera d'un nouvel échantillon auprès duquel sélectionner les participants aux sous-études.

En bout de ligne, les divers aspects de l'étude, à savoir le panel longitudinal, l'échantillon transversal et les sous-études, fonctionneront collectivement en se fournissant mutuellement des renseignements. En particulier, on s'attend à ce que les sous-études fournissent des éclaircissements concernant la cohorte principale en permettant de perfectionner les indicateurs utilisés, tandis que la cohorte principale permettra de déterminer quelles sous-études particulières doivent être réalisées.

4.2 Résumé et conclusion

L'objectif de l'Ipswich Study est d'arriver à comprendre comment la santé des individus et des collectivités est affectée par le cycle de vie et par les localités dans lesquelles vivent les individus à mesure qu'ils évoluent au fil du temps. Fondés sur une étude bibliographique de grande portée, le modèle socioécologique et le modèle du cycle de vie, qui sont habituellement utilisées indépendamment l'un de l'autre, devraient être intégrées en vue d'adopter une meilleure approche méthodologique. Ce genre d'approche permettrait de mieux comprendre les effets de l'environnement des quartiers sur la santé au cours du cycle de vie, l'effet du déménagement d'un quartier dans un autre et les changements que les quartiers eux-mêmes subissent au cours du temps. Les études de ce genre nécessiteront un plan permettant de suivre l'évolution des quartiers et celle des individus au cours du temps (Diez-Roux, 2003). Dans une région telle que Ipswich, l'évolution rapide qui est prévue offre une excellente opportunité d'établir une étude de ce genre.

Cependant, le présent article vise aussi à souligner que, si la possibilité de comprendre les déterminants sociaux de la santé grâce à des études de cohorte longitudinales axées sur la localité est grande, d'importants défis doivent être surmontés en vue de concevoir une étude qui permet de tenir compte de la vaste gamme de nuances et de questions devant y être incluses. En outre l'Ipswich Study pose des défis supplémentaires qui tiennent à la nécessité de satisfaire aux intérêts de divers intervenants tout en maintenant une flexibilité de court terme. Étant donné les multiples défis que pose l'établissement d'une telle étude, ainsi que les paramètres particuliers dans le cadre desquels doit être établie l'Ipswich Study, il est nécessaire de suivre le conseil de (Byles et coll., 2007) selon lesquels [traduction] « comme la fabrication d'un vin fin, les études longitudinales nécessitent une planification et une gestion minutieuses aux étapes de travail préparatoire et ne produisent pas leurs meilleurs résultats avant de nombreuses années ». Une planification minutieuse couplée à des plans novateurs et souples tels que ceux proposés ici permettra de surmonter les problèmes et les défis mentionnés et, en dernière analyse, de profiter des avantages de l'Ipswich Study pour produire des données utiles au niveau local, mais significatives à l'échelon international pendant de nombreuses années à venir.

Remerciements

Le présent article est fondé sur du matériel préparé par Marshall et Bush (2009a; 2009b). La participation de M. David Marshall au XXV^{e} Symposium international sur les questions de méthodologie a été financée en partie par la Ian Potter Foundation.

Bibliographie

Bauman, A. (2005). The physical environment and physical activity: moving from ecological associations to intervention evidence, *Journal of Epidemiology & Community Health,* 59, pp. 535-536.

Berkman, L. F. (2009). Social epidemiology: Social determinants of health in the United States: Are we losing ground?, *Annual Review of Public Health,* 30(1), pp. 27-41.

Byles, J., Dobson, A., Bryson, L. et Brown, W. (2007). Getting started: 'Preparing the ground' and 'planting the vines' for longitudinal research, *International Journal of Multiple Research Approaches,* 7(2), pp. 80-91.

Diez-Roux, A. V. (2003). The examination of neighbourhood effects on health: Conceptual and methodological issues related to the presence of multiple levels of organization, in I. Kawachi and L. Berkman (eds.) *Neighborhoods and Health*, New York: Oxford University Press, pp. 45-64.

Diez-Roux, A. V. (2007). Neighborhoods and health: where are we and where do we go from here?, *Rev Epidemiol Sante Publique,* 55(1), pp. 13-21.

Leyland, A. H. et Næss, Ø. (2009). The effect of area of residence over the life-course on subsequent mortality, *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 172(3), pp. 555-578.

Marshall, D. et Bush, R. (2009a). The Ipswich Study: A Review of Longitudinal Methodology, unpublished monograph, Healthy Communities Research Centre. http://www.uq.edu.au/health/healthycommunities

Marshall, D. et Bush, R. (2009b). The Ipswich Study: Guiding Approaches and New Agendas, unpublished monograph, Healthy Communities Research Centre. http://www.uq.edu.au/health/healthycommunities

Moon, G., Subramanian, S., Jones, K., Duncan, C. and Twigg, L. (2005). Area-based studies and the evaluation of multilevel influences on health outcomes, in A. Bowling and S. Ebrahim (eds.) *Handbook of Health Research Methods: Investigation, Measurement and Analysis*, Maidenhead: Open University Press, pp. 266-292.

State of Queensland (2009). *South East Queensland Regional Plan 2009-2031*, Brisbane: Department of Infrastructure and Planning.

Analyse de l'approche longitudinale de la santé mise en place en Belgique

Ingenbleek Ann, Coppieters Yves, Lammens Lies, Deboosere Patrick,
Cols Florence, D'hoore William, Levêque Alain[1]

Résumé

La mise en place du "gouvernement électronique" en Belgique est l'occasion de moderniser les secteurs de la santé et de la sécurité sociale. Le système national d'information sanitaire devrait être complété par une vision dynamique. Certaines initiatives belges récentes sont propices à une approche longitudinale. Il s'agit notamment des dossiers médicaux informatisés, des registres clinico-administratifs, ainsi que de l'Echantillon Permanent constitué sur base de l'assurance maladie. Les révisions de la législation, l'usage du numéro personnel d'identification et la disponibilité de moyens substantiels sont des facteurs favorisant une vision prospective de la santé. Les défis afférents impliquent des choix de société, la propagation d'une vision globale de santé publique et l'assentiment des professionnels aux changements induits.

Mots clefs : système d'information sanitaire, Belgique, longitudinal, indicateurs, informatisation des données.

1. Introduction : cadre de la recherche

Le système d'information sanitaire (SIS) belge actuel est capable de donner une image assez complète de la consommation des soins de santé, mais il n'est pas encore ajusté pour décrire les dynamiques sociosanitaires - individuelles et environnementales - qui conduisent les personnes dans des délais variables vers des états de bonne ou de mauvaise santé. Le Service Publique Fédéral Sécurité Sociale a dès lors souhaité enrichir sa connaissance de la santé de la population et a chargé le projet BeLHIS[2] de dresser un état des lieux de la question d'une part et d'autre part, de proposer des stratégies adaptées au contexte belge et permettant l'interaction et la coordination des différentes initiatives de suivi prospectif de santé. Etablir une vision longitudinale de la santé implique que l'information générée illustre les états de santé et leurs déterminants, au même titre que l'utilisation du système de santé. A terme, cette capacité d'observation élargie devrait induire une transition des politiques de gestion des soins de santé vers des approches de promotion de la santé et de prévention des maladies. Le projet BeLHIS propose aux institutions et acteurs concernés un cadre de référence, un lieu et une occasion de débat où organiser un système national d'information sanitaire longitudinale. A brève échéance, l'objectif est aussi de démontrer que l'approche longitudinale de la santé est réalisable sur base du couplage des données.

En effet, la Belgique dispose de peu de cohortes et de peu de moyens pour en constituer. Les autorités compétentes cherchent donc à compléter les méthodes habituelles de suivi des personnes par la réutilisation de données et éventuellement par des couplages ponctuels, à des fins d'analyses secondaires.

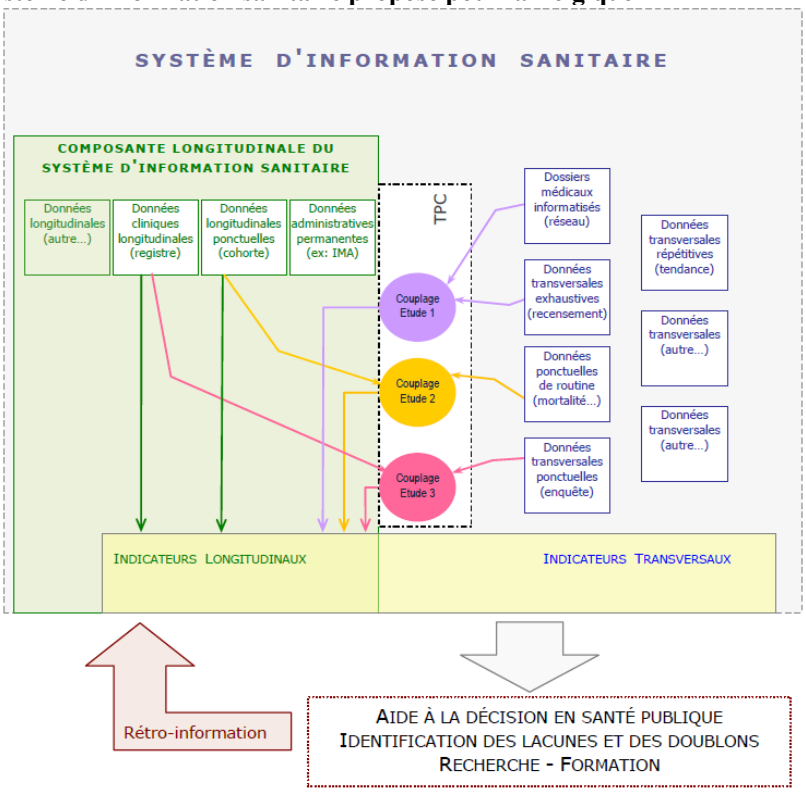
2. Modèle proposé pour le système belge d'information sanitaire longitudinale

Dans le modèle conceptuel proposé, le vaste SIS est doté d'une composante longitudinale (Coppieters, Bazelmans et Levêque, année non précisée). Les études ponctuelles ou des récoltes de données transversales (telles que les enregistrements de décès) fournissent les indicateurs sanitaires usuels. De leur côté, les suivis longitudinaux résument des phénomènes se déroulant dans le temps, illustrant une évolution. Ils aboutissent à des indicateurs mentionnés dans la **Figure 2-1** sous le terme de "longitudinaux". Il s'agit par exemple du taux de reprise du tabagisme, du taux de survie cinq ans après traitement d'un cancer, etc. Ensemble, les indicateurs transversaux et "longitudinaux" ont pour rôle de permettre l'évaluation du système de santé, l'orientation de ses objectifs, l'identification des priorités et des modes d'action adéquats.

---

[1] Ingenbleek Ann, Ecole de Santé Publique, Université Libre de Bruxelles, 808 Route de Lennik, B-1070 Bruxelles, Belgique (aingenbl@ulb.ac.be) ; Coppieters Yves, Université Libre de Bruxelles ; Lammens Lies, Vrije Universiteit Brussel, Belgium ; Deboosere Patrick, Vrije Universiteit Brussel, Belgium ; Cols Florence, Université Catholique de Louvain, Belgique ; D'hoore William, Université Catholique de Louvain, Belgique ; Levêque Alain, Université Libre de Bruxelles.
[2] **Be**lgian **L**ongitudinal **H**ealth **I**nformation **S**ystem

Figure 2-1
Schéma conceptuel du système d'information sanitaire proposé pour la Belgique



La réalisation d'un couplage de données est une opération restreinte dans le temps et dans ses finalités. Les études de couplage sont réalisées conformément à la législation et sous le contrôle d'un organe de contrôle indépendant. La tierce partie de confiance (TPC) agréée par l'Etat codifie de manière irréversible l'identité des personnes participant à l'étude. Son intervention garantit le respect des libertés individuelles. Le couplage permet aux chercheurs d'accéder à des fragments d'histoires de santé individuelles, à partir de données ponctuelles ou prélevées en routine et ceci, sans enfreindre le principe de vie privée.

3. Evolutions récentes du contexte de l'information à caractère personnel en Belgique

3.1 L'*e-government* en général

Les nouvelles technologies de l'information et de la communication (TIC) transforment la relation citoyen-administration. A la recherche de plus d'efficacité et de maîtrise des budgets, les services publics sont réorganisés, déploient de nouveaux outils électroniques et développent l'offre de services "en ligne" afin de se rapprocher de la population. Depuis le milieu des années 80, les différents niveaux administratifs du pays se sont largement engagés dans cette réforme ; plusieurs lois fédérales et adaptations légales vont dès lors entériner cette volonté.

La stratégie principale appliquée en Belgique repose sur la collecte unique de l'information puis sa validation. Ensuite, l'agence qui a recueilli l'information pourra en faire usage et la partager avec d'autres agences de l'Etat. Si nécessaire, les données ainsi validées seront réutilisées plutôt que collectées à nouveau (Robben, Desterbecq et Maes., 2005). Parallèlement, la protection de l'information - c'est-à-dire sa sécurité, son intégrité, sa confidentialité - est assurée par des mesures structurelles, organisationnelles, ainsi que par des technologies idoines.

3.2 L'*e-government* appliqué au secteur de la santé

Le secteur de la santé, ou plus exactement de la sécurité sociale, est le premier à bénéficier de cette stratégie de l'*e-gouvernment* (Robben, Desterbecq et Maes., 2005). En effet, on constate une accélération dans le processus d'actualisation du système de santé qui traduit le fait que l'enjeu de la gestion des systèmes de santé est monté dans les priorités politiques, en Belgique comme au niveau international. Ce sont surtout les perspectives financières, tirées d'observations démographiques et épidémiologiques qui justifient cette accélération.

Ce processus s'accompagne comme signalé plus haut, d'adaptions législatives précisant les moyens et les finalités autorisés, dans les domaines du droit du patient, de la vie privée, de la protection des données, des certifications de matériel et de logiciels, etc. (Devlies, Thienpont et De Moor, 2006). De façon complémentaire, la modernisation du secteur consiste également à mieux décrire et à mieux gérer le système de santé, selon les recommandations internationales. C'est ainsi qu'une liste d'indicateurs de performances du système de santé est actuellement en cours d'élaboration (Vlayen et al., 2009).

Objet concret de cette modernisation, la Belgique dispose depuis juillet 2008 d'une nouvelle institution appelée eHealth[3], qui est à la fois une plate-forme d'échange électronique de données mise à la disposition des acteurs de la santé et un outil de simplification administrative entre prestataires de santé et gestionnaires de la sécurité sociale. La plate-forme eHealth offre une infrastructure technique qui facilite le flux d'une partie de l'information sanitaire et qui s'intègre dans un environnement juridique conforme aux décisions nationales et aux réglementations européennes. Sa mise en œuvre entraîne plusieurs autres initiatives à l'échelle du pays, de nature à favoriser une perspective dynamique de santé publique au sens large.

4. Exemples de sources de données sanitaires à potentiel ou à objectif longitudinal

Les trois projets présentés ci-après peuvent s'articuler autour de la plate-forme eHealth. Ils fournissent des informations sociosanitaires de qualité, objectives, actualisées et centrées sur les personnes. Ils constituent donc un potentiel propice à l'approche longitudinale.

4.1 Le Dossier Médical Informatisé (DMI)

Il s'agit de la version digitale du dossier médical global créé et tenu à jour par un médecin généraliste à propos d'un patient afin de consigner l'information médicale importante le concernant. Les données peuvent être partagées avec ou provenir d'autres professionnels de la santé. Ce dossier n'est pas seulement un système d'archivage ; c'est aussi un outil d'amélioration de la qualité des soins fournis, particulièrement en cas de maladie chronique. Le DMI est un recueil centré sur le patient, un outil de mise au point des processus décisionnels et éventuellement, une source partielle d'évaluation en économie de la santé.

Le dossier médical global été lancé en Belgique en 1999 et sa version électronique quelques années plus tard. Il est principalement utilisé pour les soins de santé primaire, mais la possibilité existe d'y ajouter des sous-dossiers par spécialité médicale et paramédicale. La difficulté majeure qu'il soulève est qu'il requière des normes éthiques d'utilisation et l'interopérabilité des systèmes qui l'exploitent. C'est pourquoi le DMI est soumis à une procédure d'homologation, réalisée sous l'autorité du Ministère de la Santé publique.

Par ailleurs, il faut noter que sa finalité première est clinique et non pas épidémiologique. Une étude récente menée par un réseau belge de médecins généralistes relève que les données enregistrées dans le DMI sont pour l'instant encore peu exploitables en termes d'analyse épidémiologique ou de politiques gouvernementales (De Clercq E. et al., 2009). Cette situation ne semble pourtant pas inéluctable ; si elle est résolue, les DMI deviendront une source cruciale d'informations cliniques et personnelles objectivées, pouvant s'étendre à toute la population.

4.2 Les registres cliniques

Notre pays dispose de registres de différents types. Le recueil d'une information ciblée doit être systématique lorsque son but est d'élaborer des indicateurs épidémiologiques populationnels. Une obligation légale d'enregistrement des données favorise l'exhaustivité d'un tel registre. C'est dans ce contexte qu'a été organisé le Registre du Cancer[4], qui consigne les données initiales et le suivi thérapeutique des nouvelles tumeurs cancéreuses, après codification de l'identité du patient atteint. Cette base de données a un rôle de dénombrement et de contrôle global et statistique de la qualité de la prise en charge oncologique.

[3] https://www.ehealth.fgov.be
[4] http://www.registreducancer.be

Si la participation à un registre n'est pas obligatoire, son enjeu est principalement clinique. Le nouveau registre belge des arthroplasties permettra aux orthopédistes participant à l'initiative de suivre l'efficacité des techniques utilisées ainsi que la qualité du matériel prosthétique qu'ils auront implanté (INAMI, 2009). L'application électronique SAFE[5] concerne la prise en charge de l'arthrite ; elle permet notamment de suivre l'évolution d'un patient, d'accéder au registre de la polyarthrite juvénile et de signaler l'efficacité et les effets secondaires de médicaments. Parallèlement à ces aspects cliniques, sa finalité est aussi administrative, puisqu'il vérifie les conditions de prescriptions de certains médicaments et facilite l'accès aux remboursements prévus.

Les registres peuvent être construits en lien avec le DMI. Lors de l'encodage des données, celles-ci sont inscrites simultanément dans un classeur spécialisé du DMI et dans le registre. Ceci évite la double entrée, les risques d'erreur et permet une évaluation clinique qui tienne compte des complications et co-morbidités de chaque patient. Au-delà de l'intérêt immédiat pour les individus concernés, les registres offrent à long terme, une information détaillée et spécifique, autrement dit, une vision macroscopique de pathologies ou d'évènements pouvant être rares.

4.3 L'Echantillon Permanent (EP)

Il s'agit d'une banque de données constituant un échantillon anonymisé et représentatif de la population belge. Elle est composée de quelques données démographiques, d'un identifiant personnel et des données administratives récoltées dans le cadre de l'assurance maladie-invalidité obligatoire, qui couvre plus de 98% des résidents. L'échantillon de base est composé par 1/40 des bénéficiaires ; ceux-ci sont suivis pendant 10 ans. A la population des plus de 65 ans, correspond un sur-échantillonnage : une personne sur 20 de cette tranche d'âge participe à l'EP. L'échantillon total est estimé à 300.000 personnes, soit environ 3% de la population nationale. Le niveau d'observation est celui du bénéficiaire. Les données sont actualisées annuellement et sont mises à la disposition de quelques institutions en charge de la santé publique et de la sécurité sociale (INAMI, 2008).

Cette source d'information est donc orientée vers le suivi de la consommation des soins ; elle ne donne pas la possibilité de suivre la morbidité ni les habitudes individuelles. Toutefois, elle pourrait devenir une base de travail pour des couplages avec des données provenant d'autres sources sociosanitaires.

5. Les opportunités concernant la mise en œuvre de l'approche longitudinale de santé en Belgique

5.1 Mise en place d'une infrastructure technologique

Les adaptations législatives qui ont été apportées permettent d'avoir recours aux TIC pour réorganiser les flux de l'information et pour gérer les données de santé à caractère personnel. Elles le permettent car elles définissent un cadre d'utilisation et de régulation, c'est-à-dire qu'elles établissent les modalités et les conditions d'usage, les normes de sécurité et de protection ainsi que les recours possibles. Ce cadre législatif se complète d'un organe de contrôle indépendant, la Commission de la Protection de la Vie Privée[6], constituée de membres experts de la société civile et faisant rapport au Parlement fédéral.

L'infrastructure technologique qui commence ainsi à se déployer doit s'incorporer aux architectures électroniques locorégionales d'information sanitaire comme par exemple celles des hôpitaux, qui sous l'impulsion des autorités fédérales, ont déjà développé des environnements informatiques pour gérer les dossiers de leurs patients et conduire leurs activités. D'une part la production primaire de données de santé est stimulée et d'autre part, grâce à l'interconnexion des réseaux et avec l'accord des protagonistes, les données peuvent emprunter un chemin virtuel ininterrompu et sécurisé, d'un bout à l'autre du réseau.

5.2 Existence d'un numéro personnel d'identification

En 1983, un Registre National des Personnes Physiques est officiellement créé, qui répertorie chaque Belge (puis les résidents étrangers) de manière univoque. Même si la Belgique s'oriente vers une gestion décentralisée de l'information (en principe, chaque structure qui génère des données en garde la maîtrise), il reste possible de relier l'information grâce à l'existence de ce numéro personnel d'identification. Bien entendu, la liaison de l'information est réalisée selon la loi, ce qui peut impliquer une

[5] https://www.ehealth.fgov.be/fr/application/applications/SAFE.html
[6] http://www.privacycommission.be

étape d'anonymisation ou de codification de l'identité. Les différents aspects de la vie du citoyen ne sont plus compartimentés ; il est théoriquement possible de mettre des données de santé en rapport avec des données d'ordre social ou environnemental par exemple.

5.3 Contexte international incitatif

Enfin, le contexte international offre de nombreux exemples intéressants d'approche longitudinale, notamment dans les domaines de l'économie et de la santé. Les institutions internationales sont favorables à cette méthodologie, car elle donne lieu à une vérification au niveau des individus de la pertinence et de l'adéquation des actions entreprises par les gouvernements.

6. Les obstacles à cette mise en œuvre

6.1 Processus décisionnel éclaté

La structure fédérale de la Belgique est construite de telle sorte qu'à l'instar de tous les domaines de la vie citoyenne, la santé et l'économie de la santé sont segmentées entre plusieurs niveaux de pouvoir. Les compétences sont réparties entre les différentes entités fédérées de façon exclusive (il n'y a pas de concertation prévue à priori entre pouvoirs régional, communautaire et fédéral) et globale (c'est-à-dire que chaque entité assume seule le rôle normatif, la fonction exécutive et les responsabilités budgétaires). Du fait de l'équipollence des normes, chaque niveau est entièrement autonome dans ses décisions et leurs applications. La plupart des résolutions politiques se heurtent donc, lors de leur réalisation, à d'autres priorités divergentes.

6.2 Défaut de stratégie globale

Par conséquent, le pays ne bénéficie pas d'une stratégie de santé globale dont l'opérationnalisation serait finement coordonnée au travers des différents échelons de la pyramide étatique. Plutôt, la Belgique module sa progression en fonction des opportunités qui se présentent, sans calendrier fixé mais dans une direction consensuelle, établie par des experts du domaine (Devlies, Thienpont et De Moor, 2006). Dans un premier temps, le travail de pionniers de l'administration et de quelques cliniciens permet de concrétiser des applications nécessaires. Dans un deuxième temps, celles-ci sont améliorées sur une base pragmatique, en fonction des ressources disponibles. L'appropriation de ces résultats concrets par les professionnels de la santé est parfois relative et doit être stimulée par des incitants financiers ou par la réglementation.

6.3 Logiques divergentes

Les efforts importants fournis par les administrations en charge de la sécurité sociale et de la santé publique doivent trouver une contrepartie au niveau des acteurs du terrain. L'information et la formation des professionnels sont nécessaires pour bâtir la confiance. Or, l'adhésion des professionnels aux changements induits ne fait pas encore clairement partie des priorités fédérales. On peut, au contraire, regretter que professionnels et décideurs de la santé aient des logiques différentes - les uns axés vers la médecine libérale, les autres cherchant à maîtriser les budgets - et que les réserves exprimées par ces premiers ne soient pas suffisamment levées au moyen de concertations et d'échanges préparatoires par exemple.

Enfin, concernant le domaine de la vie privée et de l'utilisation des données à caractère personnel, les choix de société qui sont en jeu attendent que des décisions cruciales soient prises. Les délais constatés quant à la participation de la société à ces débats ne font que retarder l'acceptation du système dans son ensemble[7].

7. Conclusion

Si l'intérêt du développement d'une vision longitudinale, dans la santé comme dans d'autres domaines, n'est plus matière à discussion chez les scientifiques, la traduction de cette conviction en termes de priorité politique est pour l'instant peu évidente en Belgique. L'ajout d'une composante longitudinale au SIS national nécessite encore de convaincre un certain nombre de partenaires.

Le projet BeLHIS recommande que l'architecture de l'information sanitaire de notre pays soit renforcée, au moyen de liens approfondis et de coordination entre structures productrices et institutions utilisatrices de données et qu'une plus grande

[7] Voir aussi Lammens L, et al. Ethical implications of longitudinal data collection on both the individual and the societal level. Proceedings of Statistics Canada Symposium 2009.

compatibilité entre les données soit développée. Il encourage les suivis prospectifs au sein d'enquêtes récurrentes et recommande un meilleur accès aux données sociosanitaires.

Le projet cherche aussi à mettre en valeur l'intérêt de l'approche longitudinale dans toutes les réflexions qui ont trait aux performances et aux indicateurs du système de santé, c'est-à-dire que des indicateurs "longitudinaux" soient intégrés dans la grille d'évaluation du secteur de la santé. Il insiste notamment pour que soient prises en compte, au travers des indicateurs retenus, des mesures autres que strictement économiques et hospitalières, comme les dimensions sociodémographiques et environnementales des thématiques décrites.

Dans les limites méthodologiques connues, la possibilité de réutiliser l'information de santé à des fins d'analyses prospectives permet d'obtenir des indicateurs "longitudinaux" à moindre coût qu'à partir des suivis classiques de cohortes, ce qui réduit d'autant les obstacles à la mise en œuvre d'un SIS longitudinal.

Bibliographie

Coppieters, Y., Bazelmans, C. et Levêque, A. (année non précisée). Etude préparatoire en vue de compléter le système d'information sur la santé au moyen de données provenant d'une perspective longitudinale dynamique. Rapport de recherche. Projet AGORA AG/II/131.

De Clercq, E., Van Casteren, V., Jonckheer, P., Burggraeve, P., Lafontaine, M.F., Artoisenet, C., Lorant, V. (2009). Are GPs' Electronic Health Records suitable for use in Public Health Research ? Scientific Institute of Public Health, April 2009, Brussels, Belgium. Accessible sur : http://www.iph.fgov.be/epidemio/epien/re_pr_en/D_2009_2505_20.pdf

Devlies, J., Thienpont, G. et De Moor, G. (2006). eHealth strategy and implementation activities in Belgium. Report in the framework of the eHealth ERA project. December 2006.

INAMI (2008). Echantillon Permanent. Rapport des travaux de la Commission technique de l'échantillon permanent au Conseil général de l'INAMI. Note CTPS 2008/26. INAMI, Bruxelles, Belgique. Accessible sur : http://www.inami.fgov.be/information/fr/sampling/pdf/report.pdf

INAMI (2009). Communiqué de Presse 06/05/2009. INAMI, Bruxelles, Belgique. Accessible sur : http://www.inami.fgov.be/news/fr/press/pdf/press20090506.pdf

Robben, F., Desterbecq, T. et Maes, P. (2005). Les apports de l'e-sécurité sociale à la simplification et à l'harmonisation de l'application de la sécurité sociale. *Actes des Journées Juridiques Jean Dabin - Colloque 60ème anniversaire de la sécurité sociale*, 2005. To be published. Accessible sur : http://www.law.kuleuven.be/icri/frobben/publication%20list.htm

Vlayen, J., Leonard, C., et al. (2009). Belgian Health System Performance : How do we do? Projet N° 2008-50 (KCE-ISP-INAMI). Rapport intermédiaire. Mai 2009. Non diffusé.

Les aspects éthiques de la collecte de données longitudinales aux niveaux individuel et sociétal

Lies Lammens, Patrick Deboosere, Florence Cols, William D'hoore, Ann Ingenbleek, Yves Coppieters't Wallant et Alain Levêque[1]

Résumé

Les innovations technologiques font en sorte que les chercheurs des sociétés industrialisées font face à un dilemme éthique fondamental qui oppose le savoir à la vie privée. Les données longitudinales offrent des avantages particulièrement intéressants pour la recherche, mais elles comprennent des éléments qui pourraient être menaçants (ou perçus comme tels) pour la vie privée des personnes. Dans une étude antérieure[2], nous avons élaboré un cadre conceptuel des répercussions éthiques de projets de collecte de données (sur la santé). Nous avons réfléchi sur le contenu des objectifs stratégiques élaborés dans le contexte d'une collecte de données encore plus détaillées, et nous avons souligné les risques possibles de l'utilisation abusive des données recueillies à court terme (menace pour la vie privée) et à long terme (menace pour la démocratie). Dans le présent document, nous évaluons ce cadre en rapport avec deux systèmes statistiques organisés différemment, celui du Royaume-Uni et celui du Danemark.

Mots-clés : collecte des données, vie privée, démocratie, longitudinal, système statistique.

1. Introduction : recherche dans les sociétés modernes

1.1 Contexte technologique et sociétal

Les progrès technologiques ont modifié considérablement l'environnement de recherche traditionnel. De plus en plus de renseignements sont recueillis, analysés et entreposés sous forme numérique. Internet permet la transmission de l'information au niveau mondial, plus rapidement que jamais auparavant. Des techniques statistiques de plus en plus perfectionnées servent à gérer l'information.

Par conséquent, les coûts de la collecte, de l'entreposage, de l'analyse, de l'intégration et de la diffusion des données ont diminué, et la population peut être suivie de façon plus systématique et détaillée. Parallèlement, ces innovations technologiques présentent un défi pour nos sociétés : les droits des personnes relatifs à la vie privée peuvent être compromis. Les préoccupations concernant la vie privée sont présentes depuis au moins 50 ans, c'est-à-dire depuis l'établissement des grandes bases de données électroniques. Comme on recueille une somme croissante de données détaillées sur la vie et le comportement des personnes et que les données électroniques peuvent être plus facilement copiées, transférées et diffusées que les dossiers sur papier, les conséquences d'une utilisation abusive peuvent être plus importantes maintenant que par le passé (Myers et coll., 2008, p. 794).

Par ailleurs, le contexte sociétal de la collecte des données a changé. Depuis les attaques du 11 septembre, le contrôle par les administrations publiques des menaces possibles pour la sécurité nationale semble être devenu plus important que la protection de la vie privée des personnes. Le 11 septembre a changé considérablement le rôle de l'information et la façon dont les personnes perçoivent leurs droits relatifs à la vie privée (Duncan, 2004, p. 4-5).

1.2 Collecte des données

Parallèlement, la façon dont les données sont recueillies a changé considérablement. De plus en plus de bases de données englobent l'ensemble de la population. La plupart des données qu'elles contiennent sont obtenues par l'entremise de systèmes

[1]Lies Lammens, Vrije Universiteit Brussel, Pleinlaan 2 - 1050 Brussels, Belgium (Lies.Lammens@vub.ac.be); Patrick Deboosere, Vrije Universiteit Brussel, Belgium; Florence Cols, Université Catholique de Louvain, Belgium; William D'hoore, Université Catholique de Louvain, Belgium; Ann Ingenbleek, Université Libre de Bruxelles, Belgium; Yves Coppieters 't Wallant, Université Libre de Bruxelles, Belgium; Alain Levêque, Université Libre de Bruxelles, Belgium.

[2] Policy versus privacy: an analysis of ethical issues in a health monitoring project (XXVI IUSSP International Population Conference, Marrakech, 2009)

de dossiers administratifs. Une somme croissante de dossiers sont couplés d'une base de données à l'autre. En outre, il existe un intérêt croissant pour l'analyse de la trajectoire de vie des personnes au moyen de données longitudinales.

Les divers avantages que comportent ces techniques de recherches récentes représentent un défi pour la confidentialité des données. Par exemple, l'utilisation de données administratives et de processus de couplage, qui nécessitent l'accès aux microdonnées, présente un grand défi lorsqu'il s'agit d'obtenir un « consentement éclairé », c'est-à-dire le droit pour les personnes d'accorder ou de refuser leur consentement pour l'utilisation des données les concernant, la solution classique pour surmonter le dilemme qui oppose le savoir et la vie privée. En outre, les données longitudinales comprennent des renseignements beaucoup plus détaillés sur les caractéristiques et le comportement des personnes que les données transversales. C'est donc dire que le risque de divulgation est plus grand.

1.3 Approches sociétales différentes

Les politiques sont modelées par leur contexte historique. Les administrations publiques, qui sont confrontées à une évolution technologique similaire, élaborent des politiques qui leur sont propres et qui sont parfois très différentes, en conformité avec les institutions existantes, le système statistique et les usages qui prévalent concernant la vie privée.

Le système danois est reconnu internationalement pour sa production de statistiques très centralisée. Le principal producteur de statistiques est Statistique Danemark (°1850), qui est chargé de la coordination de toutes les statistiques officielles du Danemark et de la supervision des registres publics centraux, et qui doit être consulté par tous les organismes publics en ce qui a trait à la conception et au contenu des nouveaux registres (Møller et coll., 2001, p. 2). Depuis la création du système, le niveau élevé de centralisation a été compensé par des règles très strictes d'accès aux données. Pendant une longue période, cela a donné lieu à un accès très restreint aux données pour la recherche scientifique, à tout le moins pour les chercheurs à l'extérieur de l'organisme statistique. Graduellement, Statistique Danemark a élargi l'accès à ses données. De nos jours, les règles d'accès aux données sont encore très strictes, mais elles s'accompagnent d'une politique active de diffusion des données pour la recherche. Le système très centralisé du Danemark permet une stratégie cohérente à cet égard.

Le système du Royaume-Uni est décentralisé, ce qui rend compte de la structure et de l'évolution du gouvernement de ce pays, et peut-être aussi de l'hypothèse historique selon laquelle les statistiques officielles ont comme principale raison d'être de répondre aux besoins du gouvernement. Graduellement, cette perception s'est transformée en une acceptation répandue que les statistiques doivent servir la société dans son ensemble. Aujourd'hui, l'Office for National Statistics (°1996) est le principal institut statistique au Royaume-Uni, mais près de 80 % des statistiques de ce pays relèvent d'organisations autres que l'ONS, par exemple, les administrations et les ministères gouvernementaux à qui des pouvoirs ont été dévolus (Dunnell, 2007, p. 1).

Ce contexte historique a modulé les systèmes statistiques actuels du Danemark et du Royaume-Uni, ainsi que l'attitude générale et la sensibilité à l'égard de la collecte, de la divulgation et de la protection des données dans les deux pays.

Le Danemark a élaboré un système statistique qui est presque entièrement fondé sur des registres administratifs, et il est considéré comme un pionnier à l'échelle mondiale en ce qui a trait aux recensements fondés sur des registres. Cette utilisation des données administratives est facilitée par l'adoption d'identificateurs uniques (pour les personnes, les biens immobiliers et les entreprises), qui sont utilisés dans toutes les administrations publiques. Le caractère unique de ces identificateurs permet de coupler directement l'information des différentes unités statistiques et/ou sujets.

Au Royaume-Uni, la protection de la vie privée des personnes a toujours été considérée comme plus importante que la collecte de masse de données sur la population. Néanmoins, le nombre de bases de données administratives a augmenté graduellement, celles-ci mettant toutefois l'accent sur la fourniture d'un service plutôt que sur l'identité des citoyens, n'étant pas accessibles dans le public et n'offrant pas de couverture complète de la population. La plupart des données produites au Royaume-Uni aujourd'hui proviennent de recensement décennaux traditionnels et, à l'occasion, de très larges enquêtes sur une vaste gamme de sujets (Dunnell, 2007, p. 2). Par suite de cette politique, le Royaume-Uni possède probablement l'un des ensembles les plus importants et meilleurs d'enquêtes sur la population en Europe.

2. Cadre des répercussions éthiques de la surveillance de la santé

2.1 Quelles sont les répercussions de ces différentes stratégies sur la collecte des données dans le domaine de la surveillance de la santé?

L'évaluation de la santé de la population est devenue une préoccupation majeure pour les administrations publiques modernes. Cela tient non seulement à des raisons budgétaires, mais aussi au fait que les populations vieillissantes exercent de toute

évidence une pression plus grande sur les dépenses en soins de santé. Dans une démocratie moderne, le bien-être de la population est l'objectif stratégique principal et la justification ultime des politiques gouvernementales, et la santé représente sans nul doute la pierre angulaire de ce bien-être. Les attentes du public, les contraintes budgétaires et la complexité d'une politique sur la santé moderne fondée sur des faits probants ont fait en sorte de rehausser la place qu'occupe la surveillance de la santé de la population dans le programme statistique.

Les suivis longitudinaux individuels présentent un intérêt particulier en matière de surveillance de la santé. Ils sont non seulement essentiels pour évaluer correctement l'évolution de la santé de la population et pour éviter les sophismes liés à l'utilisation de données agrégées sur la population, mais ils dépendent aussi des quelques méthodes disponibles pour établir des rapports de cause à effet. Il est évident que les données individuelles et la recherche longitudinale, qu'il s'agisse d'un suivi périodique ou du couplage de données après le fait, représentent des défis particuliers du point de vue de la vie privée et de la confidentialité des données et suscitent à juste titre des préoccupations dans la communauté scientifique et dans le grand public.

Toutefois, au moment de l'évaluation des répercussions de la collecte des données, il est important de ne pas tenir compte seulement des tensions actuelles. Une évaluation appropriée devrait aussi comprendre une perspective à long terme. Ainsi, nous soutenons qu'il est important d'ajouter une dimension temporelle à l'évaluation des processus de collecte de données très exhaustifs. Nous pouvons résumer notre approche dans un tableau de contingence 2x2, dans lequel le temps introduit un niveau d'incertitude illustré par le point d'interrogation.

	Surveillance de la santé	**Répercussions sociétales**
Dilemme à court terme	1. Politique en matière de santé	2. Déficit au chapitre de la vie privée
	↓	↓
Risques à long terme	3. « Police de la santé »?	4. Déficit au chapitre de la démocratie?

2.2 Dilemme à court terme entre la politique et la vie privée

À court terme, les projets de surveillance de la santé peuvent susciter un dilemme entre les blocs 1 et 2. D'une part, l'accumulation de données personnelles à l'intérieur d'une infrastructure d'information sur la santé de plus en plus perfectionnée et automatisée a entraîné une meilleure surveillance et connaissance de la population. Par conséquent, des politiques plus appropriées peuvent être mises en œuvre, qui sont susceptibles d'entraîner des avantages importants pour la santé et une plus grande qualité de vie. Par ailleurs, les possibilités technologiques récentes et la surveillance de plus en plus détaillée de la vie des personnes et du comportement en matière de santé comprennent des éléments qui pourraient menacer, ou être perçus comme menaçant, la vie privée de citoyens. Ceux-ci pourraient commencer à s'inquiéter du contrôle exercé sur eux ou d'être « surveillés ».

Le fait que le respect de la vie privée représente un droit fondamental pour chaque citoyen semble être une idée répandue et universellement acceptée. Toutefois, la mise en application de ce droit suscite un débat considérable. De nos jours, l'équilibre entre le droit à la vie privée du public et le besoin de la société de savoir entraîne un dilemme éthique important au chapitre de la recherche. Le secteur de la santé est particulièrement touché par ce dilemme, la recherche dans ce secteur utilisant des données qui rendent compte de certains des aspects les plus personnels et délicats de la vie des personnes. Même si les articles qui paraissent dans la presse concernant la perte de données et les infractions à la sécurité peuvent donner l'impression que la vie privée est plus souvent sacrifiée au profit de la politique en matière de santé, les responsables semblent globalement conscients de cette tension et ont souvent tendance à pencher pour les préoccupations en matière de vie privée. Il est étrange de constater que, dans la foulée du 11 septembre et parce qu'ils sont confrontés à la réalité du terrorisme, les gouvernements et le grand public ont accepté des infractions graves au chapitre de la protection de la vie privée, au nom de la sécurité.

2.3 Risques à long terme en matière de politique et préoccupations concernant la vie privée

Jusqu'à maintenant, nous avons abordé le débat concernant la protection des données dans une perspective limitée, en mettant principalement l'accent sur le droit à la vie privée des personnes, c'est-à-dire « leur droit d'être protégées contre la communication à quiconque de renseignements concernant leur vie personnelle ». Toutefois, les répercussions à long terme de la façon dont les statistiques, la collecte des données personnelles et les systèmes administratifs sont organisés, devraient susciter encore plus d'attention (bloc 4). Par exemple, on sous-estime souvent le fait que les processus d'information pourraient, à long terme, contribuer à l'adoption de pratiques ou de politiques qui ne correspondent pas à notre perception traditionnelle de la démocratie, notamment la discrimination à l'endroit de personnes ou de sous-populations particulières ou la

stigmatisation de celles-ci. Dans le contexte de la recherche sur la santé, les données personnelles de nature délicate, comme l'orientation sexuelle d'une personne, son ethnicité ou ses antécédents culturels, peuvent être particulièrement pertinentes, mais il s'agit exactement du type de données personnelles qui servent souvent de base à la discrimination.

Nous ne pouvons pas être certains que les données recueillies et protégées de façon appropriée aujourd'hui ne feront pas l'objet d'utilisations abusives à l'avenir. Nous distinguons au moins trois dangers différents. Tout d'abord, le fait de posséder des données concernant des personnes donne un pouvoir qui peut être utilisé pour contrôler ces mêmes personnes. En deuxième lieu, plus les données sont importantes pour avoir du pouvoir, plus ceux qui détiennent ce pouvoir sont incités à veiller à ce que les données recueillies présentent un tableau favorable (Prewitt, 1985, p. 116). En troisième lieu, il peut devenir tentant de transformer les dossiers statistiques en dossiers administratifs et de surveillance. L'utilisation des données statistiques à des fins administratives est non seulement désastreuse pour la perception du public de la collecte impartiale de données et de la production des statistiques, mais les limites plus floues entraînent une concentration du pouvoir et pourraient faciliter l'utilisation abusive. En temps de crise ou d'instabilité politique, particulièrement, il pourrait être tentant pour les personnes qui possèdent les données de les utiliser à des fins non démocratiques, par exemple, pour pénaliser ou favoriser certains groupes (Seltzer et Anderson, 2001).

La mise en œuvre de mesures visant à prévenir l'utilisation abusive non démocratique possible des données est essentielle à court et à long terme. Parmi les mesures de protection fondamentales figurent la réglementation, les obstacles techniques, les mesures organisationnelles et les lignes directrices éthiques. Il est clair qu'aucune de ces mesures n'offre de garantie absolue d'une utilisation démocratique des données. Toutefois, toutes ces mesures mises ensemble peuvent contribuer à éliminer les utilisations abusives des données en faisant augmenter les coûts financiers, personnels et politiques de telles utilisations.

Historiquement, les préoccupations concernant l'utilisation abusive non démocratique des données ont souvent été à l'avant-plan des interventions parlementaires dans presque tous les pays où l'on a abordé la question de la technologie de l'information et de la collecte des données. Étrangement, cette préoccupation s'est atténuée, au fur et à mesure que nous nous sommes familiarisés avec l'utilisation quotidienne des technologies de l'information, mais cela a aussi mené graduellement à une discussion sur la vie privée plutôt que sur la démocratie. Les préoccupations à long terme concernant les répercussions démocratiques des TI ont tendance à disparaître, du fait notamment qu'on présume souvent que la protection de la vie privée permet aussi la protection de la démocratie et des valeurs démocratiques. Toutefois, même si de nombreuses mesures de protection de la vie privée peuvent aider à maintenir la démocratie, elles sont définitivement conçues dans une perspective différente et ne sont pas nécessairement les mieux adaptées pour protéger la démocratie à long terme. Il est nécessaire de faire une distinction entre ces deux objectifs ou risques et d'évaluer la mise en œuvre des systèmes statistiques en gardant les deux à l'esprit.

3. Comment les systèmes statistiques du Danemark et du Royaume-Uni traitent-ils ces aspirations conflictuelles?

3.1 Dilemme à court terme entre la politique et la vie privée

Même si les préoccupations mentionnées précédemment ont été exprimées à la fois au Royaume-Uni et au Danemark, ces deux pays ont élaboré des méthodes différentes de mise en œuvre de leurs systèmes statistiques. L'analyse des deux systèmes permet-elle de conclure qu'un modèle est plus approprié que l'autre pour répondre aux préoccupations ou objectifs concurrents liés au traitement moderne des données?

À court terme, le système statistique danois est reconnu pour offrir de nombreux avantages au chapitre de la recherche ainsi que des politiques, par exemple, la possibilité de recueillir une vaste gamme de données détaillées sur divers domaines touchant l'ensemble de la population, et le fait que le système est relativement peu coûteux et hautement efficace.

Par ailleurs, le système est souvent perçu comme menaçant la vie privée. Cela vient principalement de l'organisation centralisée, les données étant concentrées en un endroit ou relevant d'un seul responsable, de l'utilisation d'identificateurs uniques et de la possibilité de coupler les données de différentes bases de données. Cela mis à part, la perception du caractère menaçant pour la vie privée de ce système a servi de base à de nombreuses mesures de protection et a limité l'accès aux microdonnées danoises pour la recherche pendant de nombreuses années. Ce système statistique efficace a été pendant longtemps un terrain peu fertile pour la recherche pour l'ensemble de la collectivité des chercheurs.

Le système du Royaume-Uni est moins efficace et plus coûteux que le système danois. Toutefois, il offre certains avantages particuliers, par exemple, l'utilisation exhaustive de données d'enquête comprenant des renseignements beaucoup plus détaillés sur des sujets particuliers. En ce qui a trait à la vie privée, le système décentralisé est perçu comme plus sécuritaire. Le système

du Royaume-Uni offre moins de possibilités de couplage des données, ce qui écarte certains problèmes importants liés à la vie privée.

Étonnamment, la confiance du public à l'égard du système du Royaume-Uni n'est pas très grande. L'épisode récent de perte de données ou d'information, qui a fait l'objet de reportages dans la presse du Royaume-Uni, a été à l'origine de cela. Des règles strictes et une politique transparente, par ailleurs, ont entraîné une confiance élevée du public à l'égard du système danois.

3.2 Risques à long terme pour la politique et préoccupations concernant la vie privée

À long terme, il est difficile de déterminer comment les deux systèmes évolueront. La technologie progresse extrêmement rapidement, ce qui facilite la collecte des données à grande échelle et la mise en correspondance de la situation et des caractéristiques d'une population. Par conséquent, les possibilités de recherche futures, ainsi que leurs répercussions sur la confidentialité des données, sont difficiles à prévoir.

Les éléments fondamentaux du système statistique actuel du Danemark ont été établis au moment de l'avènement de la démocratie au Danemark, à la fin du XIX^{e} siècle, et ils semblent avoir été influencés profondément par cette idéologie. Les besoins de transparence, d'ouverture, de fiabilité, de convivialité et de compétence professionnelle de la part des responsables statistiques semblaient appropriés par rapport à l'ancien régime, dans lequel les statistiques étaient gardées secrètes à certains moments. Toutefois, que pourrait-il arriver au système, qui comprend une somme considérable de données personnelles concernant la population danoise, si les conditions sociétales et le contexte politique devenaient instables, en temps de guerre, par exemple, ou par suite de la modification des relations de pouvoir?

Le système statistique du Royaume-Uni s'est édifié graduellement, plutôt guidé par des réformes successives correspondant au contexte politique qui prédominait que par une idéologie ferme ou un changement radical à un certain point de l'histoire. Depuis le 11 septembre, l'accent mis sur la sécurité nationale a augmenté considérablement au Royaume-Uni, alors que l'importance accordée à la protection de la vie privée des personnes a diminué. En outre, les mesures récentes en vue d'un système d'information plus intégré semblent être prises graduellement.

Le Danemark et le Royaume-Uni ont une approche très différente, et parfois opposée, en ce qui a trait à l'organisation de leur système statistique et à la question de la confidentialité des données. Aucun système ne peut empêcher totalement l'utilisation abusive possible des données, et le système du Danemark et du Royaume-Uni comportent tous les deux des avantages particuliers. Toutefois, dans l'ensemble, tant pour la recherche et les politiques, que pour la protection des données recueillies, le système danois semble comporter plus d'avantages que le système du Royaume-Uni, même si au premier abord, il semble plus dangereux. En débattant publiquement et explicitement de la question de la protection des données et en mettant en œuvre des mesures de protection appropriées, le Danemark a jusqu'à maintenant réussi à prévenir les utilisations abusives des données et les importantes pertes de données, même si son système statistique est très centralisé et comprend plus de renseignements détaillés sur ses citoyens que celui du Royaume-Uni. Ainsi, l'exemple danois semble démontrer que la collecte exhaustive de données personnelles peut être compatible à court et à long terme avec une société dans laquelle la protection de la vie privée des personnes et des valeurs démocratiques est perçue comme fondamentale.

4. Conclusion

La réussite de nos sociétés industrialisées, qui est illustrée par la croissance constante du niveau de vie moyen, et plus particulièrement de l'espérance de vie en santé, est principalement liée à leur organisation et à leur fonctionnement proprement dits. Toutefois, l'organisation du savoir dans nos sociétés, c'est-à-dire la façon dont les connaissances sont diffusées, la formation est assurée et la recherche est organisée, a aussi contribué considérablement à ce succès.

Du fait des progrès technologiques dans les sociétés modernes, il a été et il deviendra plus facile de recueillir des données sur la santé au niveau de la population. C'est pourquoi il est maintenant plus important que jamais de tenir compte des conséquences de la collecte en masse de données parfois très personnelles sur les particuliers. Les données longitudinales offrent des avantages particulièrement intéressants au chapitre de la recherche, du fait qu'elles permettent d'établir un lien entre les résultats sociaux et les causes sous-jacentes, mais elles posent un défi de taille en ce qui a trait à la confidentialité. Elles comprennent des renseignements détaillés sur la vie et le comportement des personnes et peuvent révéler des modèles permettant de connaître l'identité de certaines. En outre, pour élaborer des données longitudinales, des données individuelles doivent souvent être couplées à répétition, ce qui augmente aussi le risque d'identification.

Compte tenu de la capacité croissante de nos systèmes d'information de saisir la réalité dans les données, il est de première importance d'inclure dans les projets de collecte des données les évolutions possibles à long terme. L'établissement du bon

équilibre entre la demande de savoir dans nos sociétés et la protection de la vie privée des personnes et une utilisation démocratique des données semble être difficile, voire impossible. Il ne s'agit certainement pas d'une solution idéale pour résoudre ce dilemme.

Outre des mesures techniques, légales et organisationnelles à jour, nous percevons un contrôle démocratique des données comme crucial pour assurer la protection contre l'utilisation abusive possible des données, particulièrement du fait que les possibilités technologiques progressent à une rapidité imprévisible. Après tout, ce n'est pas l'information proprement dite qui est dangereuse pour la vie privée des personnes ou le système démocratique. En outre, un niveau élevé et exact d'information du public représente une condition fondamentale de la démocratie. Le problème se pose lorsque seulement un groupe de personnes a accès aux données recueillies et les utilise à des fins non démocratiques. Par contre, si les données recueillies sont partagées par de nombreuses personnes, et que l'analyse et l'utilisation de ces données se font de façon transparente et ouverte, l'information proprement dite n'est pas incompatible avec les valeurs liées à la vie privée et à la démocratie. En outre, lorsque l'on recherche un équilibre entre la collecte des données ou la protection contre les utilisations abusives possibles, il est important de discuter de cette question éthique de façon explicite et d'augmenter la sensibilisation à l'égard des problèmes possibles.

Bibliographie

Cols, F., Lammens, L., Ingenbleek, A., D'hoore, W., Deboosere, P, Coppieters, Y. et Levêque, A. (2009). *Gestion des données à caractère personnel. Réflexions sur le système d'information sanitaire belge, à partir d'une analyse comparative des systèmes britannique et danois.* Working paper N° 2, Projet BeLHIS, AGORA AG/00/139, Bruxelles, 2009.

Duncan, G.T. (2004). *Exploring the Tension Between Privacy and the Social Benefits of Governmental Databases*, paper presented at the Georgetown University Law Center Conference on Security, Technology, and Privacy: Shaping a 21st Century Public Information Policy, Washington DC.

Dunnell, K. (2007). *Evolution of the United Kingdom statistical system*, paper presented at the UN Seminar on Evolution of National Statistical Systems, New York, USA.

Myers, J., Frieden, T.R., Bherwani, K.M. et Henning, K.J. (2008). Privacy and Public Health at Risk: Public Health Confidentiality in the Digital Age, *American Journal of Public Health,* 98, 5, pp. 793-801.

Prewitt, K. (1985). Public statistics and democratic politics, in J.J. Smelser and D.R. Gerstein (eds.), *Behavioral and Social Science: Fifty years of Discovery*, Washington, D.C.: Nat. Academic Press.

Seltzer, W. et Anderson, M. (2001). The Dark Side of Numbers: The Role of Population Data Systems in Human Rights Abuses, *Social research*, 68, 2, pp. 481-513.

Statistics Denmark (2008). *Numbers on time. Introduction to Statistics Denmark*, [http://www.dst.dk/HomeUK/About/introSD.aspx, consulted 12/04/2009].

L'apport des bases administratives et médico-administratives dans la cohorte CONSTANCES

Alice Guéguen, Rémi Sitta, Laetitia Bénézet, Gaëlle Santin, Jean-Louis Lanoë, Marcel Goldberg , Marie Zins et les Centres d'examen de santé de la Sécurité sociale[1]

Résumé

Constances est une cohorte française qui inclura 200 000 personnes, et qui a un double objectif de santé publique et de recherche épidémiologique. Les informations d'intérêt seront non seulement recueillies auprès des personnes, mais seront également issues de bases administratives et médico-administratives françaises. En ce qui concerne le recueil auprès des personnes, le taux de réponse attendu est faible (environ 15%) et des corrections pour non-réponse seront effectuées à l'aide d'informations auxiliaires disponibles chez les répondants et les non répondants, prédictrices de la non réponse, et qui seront extraites des bases de données évoquées précédemment.

Mots Clés : données administratives, non-réponse, épidémiologie, enquête longitudinale, cohorte, pondération.

1. Introduction

L'épidémiologie est la science qui étudie la fréquence et la distribution, dans le temps et dans l'espace, des problèmes de santé des populations (épidémiologie descriptive), ainsi que le rôle des facteurs qui les déterminent (épidémiologie étiologique, ou analytique). Actuellement, les questions de recherche en épidémiologie consistent principalement à mettre en évidence des risques de faible ampleur, concernant des expositions peu fréquentes ou de faible intensité et pour lesquels les temps de latence entre exposition et maladie peuvent être très longs. Il est donc nécessaire de mettre en place des cohortes (également appelées panels) avec suivi individuel, de taille importante et dont la durée de suivi peut se compter en dizaines d'années.

Constances est une cohorte épidémiologique mise en place par l'Institut national de la santé et de la recherche médicale (Inserm). Elle comprendra 200 000 personnes âgées de 18 à 69 ans à l'inclusion et couvertes par le régime général de la Sécurité sociale. Elle est destinée à fournir des informations à visée de santé publique et contribuera au développement de la recherche épidémiologique concernant principalement les déterminants sociaux et professionnels de la santé. Une phase pilote incluant environ 3 000 sujets a débuté en septembre 2009 et prendra fin en mars 2010. La mise en place de la cohorte commencera à la fin de l'année 2010. On attend un taux de non réponse élevé, de l'ordre de 85%, ce qui est courant dans les enquêtes épidémiologiques, car elles sont habituellement fondées sur le volontariat ; on emploiera ainsi les termes de participation/non-participation au lieu de ceux de réponse/non réponse. Les biais dus à la non-participation seront pris en compte grâce à des données auxiliaires disponibles sur les participants et les non participants, et issues de bases de données nationales administratives et médico-administratives françaises.

La deuxième partie aborde les biais dus aux effets de sélection en épidémiologie. Les bases de données françaises utilisées pour corriger les biais de non participation seront présentées dans la troisième partie. La quatrième partie sera consacrée à la présentation de la cohorte Constances. Enfin, la dernière partie contiendra des éléments de discussion.

[1] Alice Guéguen, Rémi Sitta, Jean-Louis Lanoë, Marcel Goldberg, Marie Zins, Unité mixte Inserm-Cnamts 687 / Centre de recherche en Epidémiologie et Santé des Populations, Hôpital Paul Brousse, Bâtiment 15/16, 16 avenue Paul Vaillant Couturier, 94807 Villejuif Cedex, France ; Laetitia Bénézet, Gaëlle Santin, Marcel Goldberg, Institut de veille sanitaire, 12 rue du Val d'Osne 94415 Saint-Maurice Cedex, France.

2. Biais dus aux effets de sélection en épidémiologie

2.1 Description des biais

Dans les enquêtes épidémiologiques, les données sont le plus souvent recueillies directement auprès des sujets eux-mêmes du fait de la nature des informations recueillies (consommation d'alcool et de tabac, alimentation, etc.). Or, pour des raisons d'éthique et de législation, on ne peut rendre obligatoire la participation active des personnes à l'enquête : il en résulte que la population enquêtée diffère de la population cible en raison de phénomènes liés à la non participation à l'inclusion et à l'attrition (encore appelée érosion) au cours du suivi.

Les biais induits par la non participation peuvent concerner, dans le cas d'un objectif descriptif, l'estimation de la fréquence de la maladie ou de l'exposition à un facteur de risque ou, lorsque l'objectif est analytique, l'estimation de l'association entre exposition et maladie. L'estimation de la fréquence de la maladie ou de l'exposition, ou de l'association exposition – maladie, est biaisée si la probabilité d'être malade (ou exposé) n'est pas indépendante de la probabilité d'être inclus dans l'étude, ou si la relation exposition – maladie est différente entre les sujets inclus et ceux qui ne le sont pas. Le biais est en général plus important lorsqu'on s'intéresse à l'estimation d'une fréquence que lorsqu'on s'intéresse à l'estimation d'une association.

2. 2 Les solutions

Quelle que soit la solution mise en œuvre, elle nécessite d'avoir une connaissance approfondie des facteurs prédicteurs de la participation à des enquêtes épidémiologiques. Les études de Goldberg et al (2001) et de Zins et al (2008) ont montré que les facteurs liés à la participation sont essentiellement l'âge, la catégorie sociale, le statut marital, alors que les facteurs liés à l'attrition sont avant tout les comportements de santé (consommation d'alcool et de tabac), et la survenue d'événements de santé sérieux.

En épidémiologie analytique, on utilise habituellement la méthode qui consiste à inclure les facteurs liés à la participation dans le modèle de régression expliquant la maladie par l'exposition. L'avantage de cette méthode est qu'elle permet de se dispenser de la mesure de ces facteurs sur les non participants. Il faut cependant noter que dans certains cas, ce procédé peut poser des problèmes (Hernan, 2004) et ne peut donc pas être utilisé.

Dans les enquêtes par sondage aléatoire, la méthode de repondération des poids d'échantillonnage permet de prendre en compte les biais dus à la non participation, qu'il s'agisse d'études descriptives ou analytiques. Mais elle implique de disposer de la mesure de ces facteurs à la fois chez les participants et les non participants. Les bases de données administratives et médico-administratives présentées ci après répondent à ce besoin, car elles sont exhaustives et contiennent de nombreuses informations qui sont des proxys des facteurs de participation aux enquêtes épidémiologiques.

3. Les bases de données administratives et médico-administratives françaises

Il existe en France plusieurs bases de données nationales et exhaustives, contenant de nombreuses informations susceptibles d'être liées aux facteurs de participation. Il s'agit des bases de données socio professionnelles provenant de la Caisse Nationale d'Assurance Vieillesse (Cnav), de données de santé issues du Système national d'information inter-régimes de l'assurance maladie (SNIIR-AM), ainsi que de la base de données du Centre d'épidémiologie des Causes de Décès de l'Inserm (CépiDc). Toutes ces sources d'informations pourront être appariées grâce à l'utilisation d'une forme cryptée du Numéro d'identification au Répertoire (NIR), identifiant national unique employé dans toutes ces bases de données (Goldberg, 2008).

3.1 Données socioprofessionnelles de la Caisse nationale d'assurance vieillesse

Le rôle de la Cnav est notamment d'assurer le droit au paiement de la retraite pour tout individu ayant appartenu au moins une fois à un moment de sa vie au régime général de la Sécurité sociale, régime qui assure environ 85% de la population française. Pour cela, la Cnav a mis en place un système permettant de collecter et traiter les données sociales issues de différents organismes et régimes gestionnaires des prestations sociales. Pour la constitution et l'enrichissement de ses bases de données, la Cnav reçoit régulièrement des données transmises par les employeurs, y compris les employeurs de personnel de maison. Elle reçoit également des informations sur les périodes d'inactivité (chômage, absence pour maladie, maternité) de la part de plusieurs organismes. Ces données, enregistrées de façon prospective, servent de base pour le calcul des retraites et sont de ce fait complètes et particulièrement bien validées, notamment pour les périodes les plus récentes, et leur qualité (complétude et exactitude) s'améliore régulièrement au fil des années avec l'informatisation du recueil à la source.

3.2 Données issues du Système national d'information inter-régimes de l'assurance maladie

Le SNIIR-AM contient des données individuelles médicalisées, structurées et codées de manière standardisée :

- données de consommation de soins : ces données proviennent de la base de données de remboursement de l'Assurance Maladie ; elles ne comportent pas d'information sur la nature des maladies traitées, et excluent par définition l'automédication et les prestations non présentées au remboursement,
- données d'hospitalisation : la base de données du PMSI (Programme de médicalisation des systèmes d'information des hôpitaux) comprend pour chaque séjour hospitalier le recueil du diagnostic principal, du ou des diagnostics associés, et des actes diagnostiques et thérapeutiques les plus coûteux. Les diagnostics et les actes médicaux sont codés de façon standardisée,
- données concernant les Accidents du Travail, les Maladies Professionnelles et les Affections de Longue Durée, ces dernières représentant les 30 maladies les plus graves.

3.3 Données de Mortalité provenant du Centre d'épidémiologie des causes de décès de l'Inserm

Les données de mortalité - statut vital et causes de décès - peuvent être obtenues auprès du Centre d'épidémiologie des causes de décès de l'Inserm.

4. La cohorte Constances

4.1 Les objectifs

Constances (www.constances.fr) est conçue comme une cohorte généraliste, à vocation très large ; c'est-à-dire qu'elle n'a pas d'objectifs spécifiques en termes d'hypothèses concernant des pathologies et/ou des facteurs de risque précis, mais vise à constituer une « plateforme de recherche épidémiologique ». De plus, la durée du projet n'est pas définie et Constances fera l'objet d'un suivi longitudinal sans limite de temps, pour pouvoir étudier les effets de facteurs de risque à très long terme, et pour tenir compte de l'évolution des connaissances et des techniques, qui suscitent constamment de nouvelles questions scientifiques.

Constances a un double objectif de santé publique et de recherche, et elle sera ouverte à la communauté scientifique. Les objectifs de recherche de Constances sont largement centrés sur l'étude des déterminants professionnels et sociaux de la santé à travers quatre thèmes principaux : les risques professionnels, les inégalités sociales de santé, le vieillissement et la santé des femmes. Constances a également des objectifs de santé publique ; en ce qui concerne les objectifs de surveillance épidémiologique, une collaboration avec l'Institut de veille sanitaire sera mise en place, et plus particulièrement avec le Département santé travail pour la surveillance des risques professionnels à travers le programme Coset (Santin, 2009).

4.2 Le réseau des Centres d'examen de santé

La Cohorte Constances (CONSultants des Centres d'Examen de Santé) s'appuiera sur le réseau des Centres d'examen de santé (CES) ouverts aux assurés du régime général de la Sécurité sociale. Ces derniers peuvent bénéficier tous les cinq ans d'un examen de santé approfondi en se rendant dans le centre dont ils dépendent. Dix-sept centres, répartis sur tout le territoire, ont accepté de participer à Constances. La population cible, habitant les 17 départements dans lesquels ces centres sont situés, est proportionnelle à la population française en ce qui concerne l'age, le sexe et la catégorie sociale.

4.3 Le plan de sondage et le recueil des données

Un échantillon de 2 millions de sujets appartenant à la population couverte par le régime général de la Sécurité sociale et habitant les 17 départements Constances, sera tiré au sort par sondage aléatoire avec probabilités inégales et stratifié sur l'âge, le sexe, le statut d'activité/inactivité, la catégorie sociale et le département. Les personnes tirées au sort seront invitées à participer à Constances dans le cadre d'un bilan de santé passé dans l'un des 17 CES. Grâce à des enquêtes antérieures, dans lesquelles les sujets étaient invités à se déplacer dans un centre d'examen de santé, on a pu estimer que la participation à Constances serait environ de 10-15%, et on dispose d'estimations de la probabilité de participation en fonction de l'âge, du sexe, et de la catégorie sociale du sujet. En conséquence, les probabilités de tirage seront inégales, de manière à ce que les sujets ayant une participation attendue plus faible aient un poids de tirage proportionnellement plus élevé. L'objectif est que les effectifs attendus de participants soient proportionnels à ceux la population cible selon l'âge, le sexe et la catégorie sociale.

Au total, on attend environ 200 000 participants à Constances, pour lesquels on disposera des données du bilan de santé ainsi que de données recueillies par questionnaires et auto-questionnaires. Un échantillon de 400 000 personnes sera tiré au sort parmi les 1 800 000 non participants. Pour les 600 000 personnes des deux échantillons, participants et non participants, on disposera des données issues la Cnav sans limite de recul et de celles du SNIIR-AM avec deux ans de recul. En effet, pour ces dernières, l'information étant stockée sur trois ans glissants, on disposera de ces données pour l'année en cours et pour les deux années précédant l'inclusion (ou la non inclusion) dans Constances. Au cours du suivi longitudinal, les deux échantillons (participants et non participants) seront suivis de manière passive dans les bases de données de la Cnav et du SNIIR-AM. Pour les participants, un suivi actif sera mis en place au moyen d'auto-questionnaires postaux annuels. Il est également prévu que les participants soient réinvités à passer un examen de santé tous les cinq ans. Le statut vital des participants et des non participants, ainsi que les causes de décès, seront recherchés chaque année auprès du Cépidc.

4-4 Corrections pour non-participation

La richesse des informations contenues dans les bases de données de la Cnav et du SNIIR-AM sera suffisante pour que la probabilité de participation puisse être modélisée de façon acceptable. Des groupes de réponse homogène (Eltinge, 1997) seront constitués, afin de calculer des corrections pour non-participation. Cette opération pourra être suivie d'un calage sur marges, après que les marges auront été précisées. Les probabilités de participation seront tout d'abord estimées à l'aide des informations disponibles à l'inclusion, mais elles pourront également être réactualisées grâce aux informations recueillies au cours du suivi longitudinal. Par exemple, on sait que certains comportements de santé, tels qu'une consommation d'alcool et/ou de tabac élevée, sont prédicteurs de non-participation aux enquêtes épidémiologiques. Or, les données relatives à ces consommations ne seront disponibles que pour les participants à Constances, car recueillies par auto-questionnaire ; de plus, aucun proxy de ces variables n'existe, tout au moins au moment de l'inclusion, dans les bases de données du SNIIR-AM. En revanche, on sait que ces comportements peuvent générer des événements de santé à plus ou moins long terme, comme des cancers du poumon ou des voies aérodigestives supérieures ; ces événements, postérieurs à l'inclusion ou à la non inclusion dans Constances, seront quant à eux recueillis pour tous les sujets dans les bases de données du SNIIR-AM. Il sera ainsi possible d'affiner les probabilités de participation au fur et à mesure du suivi.

Des corrections pour attrition (également appelée érosion) seront calculées par des méthodes analogues, et on disposera de données plus nombreuses sur les participants tant à l'inclusion qu'au cours du suivi.

5. Discussion

Les bases de données administratives et médico-administratives de la Cnav, du SNIIR-AM et du CépiDc font partie intégrante du dispositif Constances : non seulement elles contribueront à la prise en compte des effets de sélection dus au volontariat, mais de plus elles fourniront des données d'intérêt utiles en elles-mêmes.

Elles ont cependant des limites : par exemple, les diagnostics médicaux figurant dans les bases du SNIIR-AM ne sont pas systématiquement validés, du fait que ces bases sont à vocation de gestion. Alors que cela pourrait poser un problème si l'on s'intéressait aux diagnostics médicaux en tant que tels, cela ne sera en revanche pas le cas si ces diagnostics sont utilisés pour estimer les probabilités de participation, car on peut faire l'hypothèse que les erreurs attachées à ces diagnostics ne seront pas distribuées différentiellement chez les participants et les non participants.

Les bases de données administratives et médico-administratives utilisées dans Constances possèdent plusieurs avantages :

- les bases de données sont nationales et exhaustives,
- les informations sont disponibles avec deux ans de recul avant l'inclusion pour les données SNIIR-AM et sans limite de recul pour les données de la Cnav,
- les informations, recueillies de façon prospective seront disponibles à l'inclusion et au cours du suivi,
- les informations sont codées de façon standardisée et indépendante du sujet.

Outre la stratégie décrite ci-dessus visant à prendre en compte les biais de sélection dus au volontariat, des enquêtes complémentaires seront menées en parallèle auprès des non participants. Il est prévu de recueillir des informations sur le motif de non-participation, certaines expositions professionnelles et maladies, et quelques comportements de santé connus pour être liés à la participation. Ces études apporteront d'autres éléments de discussion sur les biais dus au volontariat.

Bibliographie

Eltinge, J. L., Yansaneh, I.S. (1997). Méthodes diagnostiques pour la construction de cellules de correction pour la nonréponse avec application à la nonréponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey, *Techniques d'enquête*, 23, p. 37-45.

Goldberg ,M., Chastang, J.F., Leclerc, A., Zins, M., Bonenfant S., Bugel, I., Kaniewski, N., Schmaus, A., Niedhammer, I., Piciotti, M., Chevalier, A., Godard, C., Imbernon, E. (2001). Socioeconomic, demographic, occupational, and health factors associated with participation in a long-term epidemiologic survey: a prospective study of the French GAZEL cohort and its target population, *Am J Epidemiol,* 154, p. 373-84.

Goldberg, M. et Luce, D. (2001). Les effets de sélection dans les cohortes épidémiologiques. Nature, causes et consequences, *Rev Epidemiol Santé Publ*, 49, p. 477-492.

Hernan, M.A., Hernandez-Diaz, S. et Robins, J.M. (2004). A structural approach to selection bias, *Epidemiology,* 15, p 615-625.

Goldberg, M., Quantin, C., Guéguen, A. et Zins, M. (2008). Bases de données médico-administratives en épidémiologie: intérêts et limites, Courrier des statistiques, 124, p59-70.

Santin, G., Bénézet, L., Guéguen, A., Sitta, R., Gauvin, S., Sarter, H., Razafindratsima, N., Zins, M., Geoffroy-Perez, B. et Goldberg, M. (2009). Stratégies pour étudier le biais de non réponse dans le programme Coset et la cohorte Constances, *Recueil : Symposium 2009, Les enquêtes longitudinales: de la conception à l'analyse,* Statistique Canada.

Zins, M., Chastang, J.F., Coeuret-Pellicer, M., Leclerc, A., Bonenfant, S., Guéguen, A., Ozguler, A. et Goldberg, M. (2009). Facteurs associés à la participation dans la cohorte GAZEL, *Recueil : Symposium 2009, Les enquêtes longitudinales: de la conception à l'analyse,* Statistique Canada.

SESSION 10 – ENJEUX LIÉS À LA COLLECTE DE DONNÉES D'ENQUÊTES LONGITUDINALES

Maintien du contact avec les familles mobiles de la Millennium Cohort Study du Royaume-Uni

Lisa Calderwood[1]

Résumé

La présente communication est axée sur le problème de la localisation des familles mobiles dans le cadre de la Millennium Cohort Study (MCS), une étude longitudinale à grande échelle menée au Royaume-Uni, et tente de déterminer la proportion de familles qui déménagent entre les vagues et qui sont localisées avec succès grâce aux procédures de dépistage de l'étude. Elle porte en outre sur l'efficacité des techniques conçues pour prendre connaissance des changements d'adresse avant le début des travaux sur le terrain pour une vague particulière, comparativement au dépistage sur le terrain effectué par les intervieweurs, et examine certains des facteurs liés au succès ou à l'échec de la localisation des familles mobiles. Elle montre que plus de neuf familles mobiles sur dix ont été localisées avec succès entre la vague 2 et la vague 3 de l'étude, la majorité (55 %) l'ayant été avant le début des travaux sur le terrain de la deuxième vague. Même si certaines différences sont notées dans les caractéristiques démographiques et socioéconomiques observables des familles mobiles et non mobiles, très peu de ces caractéristiques sont liées au succès ou à l'échec de la localisation de ces familles.

Mots-clés : dépistage, attrition, mobilité, non-réponse, Millennium Cohort Study.

1. Introduction

L'un des principaux avantages des enquêtes longitudinales au chapitre de l'analyse est qu'elles offrent aux chercheurs la possibilité d'étudier les changements au fil du temps. L'attrition découlant des enquêtes longitudinales peut mener à un biais dans les résultats d'une étude, si les membres de l'échantillon qui abandonnent au fil du temps diffèrent de façon systématique de ceux qui demeurent dans l'étude. Une des préoccupations particulières est que, si les facteurs liés à la perte d'échantillon sont eux-mêmes associés aux processus de fond que l'étude vise à mesurer au fil du temps, cela peut mener à des estimations biaisées des changements. Lepkowski et Couper (2002) font une distinction entre trois sources différentes d'attrition : impossibilité de localiser, impossibilité de prendre contact après la localisation et impossibilité de collaborer avec les personnes contactées. La présente communication est axée sur l'attrition de l'échantillon attribuable à l'impossibilité de localiser. L'une des principales raisons pour lesquelles les études longitudinales visent à dépister les membres de l'échantillon qui déménagent vient de ce que la dynamique de la mobilité résidentielle et les processus s'y rapportant, comme les relations et les changements d'emploi, représentent un intérêt de fond, et que l'impossibilité de localiser les membres de l'échantillon qui déménagent peut mener à des estimations biaisées des changements qui touchent ces domaines et d'autres domaines importants.

Les données de la Millennium Cohort Study (MCS), la quatrième d'une série d'études de cohortes de naissances de réputation internationale réalisées au Royaume-Uni, serviront à examiner le problème de la localisation. Des recherches antérieures sur la MCS, dont il est question dans Plewis (2007a) et Plewis et coll. (2008), ont montré que la mobilité résidentielle entre la vague 1 et la vague 2 est un prédicteur de la non-réponse à la vague 2, mais que lorsque la localisation est réussie, les familles mobiles ne sont pas plus susceptibles de refuser d'être interviewées à la vague 2 que les familles non mobiles. Les recherches ont aussi démontré que, même si la mobilité résidentielle est associée à la non-réponse à la vague 2, il ne s'agit pas d'un prédicteur de l'abandon permanent de l'étude.

La présente communication examinera la proportion des familles qui déménagent entre la vague 2 et la vague 3 et qui sont localisées avec succès grâce aux procédures de dépistage de l'étude et, plus particulièrement, l'efficacité des techniques conçues pour prendre connaissance des changements d'adresse avant le début des travaux sur le terrain pour une vague par rapport au dépistage effectué par les intervieweurs sur le terrain. Elle examinera en outre certains des facteurs liés au succès ou à l'échec de la localisation des familles mobiles.

[1]Lisa Calderwood, Institute of Education, 20 Bedford Way, Londres, Royaume-Uni, WC1H 0AL (l.calderwood@ioe.ac.uk)

2. Localisation des membres de l'échantillon des enquêtes longitudinales

Le problème de la localisation des membres de l'échantillon des enquêtes longitudinales est lié à la propension des personnes à déménager et, par suite du déménagement, à être localisées. Couper et Ofstedal (2009) présentent un modèle général pour aider à comprendre le processus de localisation qui repose sur l'hypothèse que les principaux facteurs touchant la propension à déménager sont des facteurs au niveau de la personne, comme l'âge, la situation familiale, l'emploi et la situation de logement, et des facteurs au niveau sociétal, comme le niveau général de mobilité et d'urbanisation. La propension à être localisé, par ailleurs, est influencée par des facteurs de conception de l'enquête, comme l'intervalle entre les vagues et les procédures de dépistage, ainsi que des facteurs structurels, comme la disponibilité de registres de la population, les règles d'envoi par la poste et la transférabilité du numéro de téléphone. Couper et Ofstedal font un examen des ouvrages publiés concernant la probabilité de déménagement et montrent que les taux de mobilité varient à la fois à l'intérieur d'un pays et entre les pays, et qu'une gamme variée de facteurs démographiques et socioéconomiques sont liés à la mobilité. Ils évaluent en outre les facteurs structurels ou les facteurs de conception de l'enquête qui sont susceptibles d'être associés à la capacité de localiser les membres de l'échantillon qui déménagent. Cela comprend un examen utile des procédures de dépistage qui sont couramment utilisées dans les enquêtes longitudinales, examen qui comporte une distinction entre le dépistage rétrospectif, qui est conçu pour localiser des membres de l'échantillon avec qui on a perdu le contact, et le dépistage prospectif, qui est conçu pour prévenir la perte de contact, grâce à la mise à jour des coordonnées et au dépistage par les bureaux et sur le terrain. Les auteurs soulignent que, même si la plupart des enquêtes longitudinales consacrent des ressources considérables au dépistage des membres mobiles de l'échantillon et que des procédures très efficaces ont été élaborées pour réduire l'attrition découlant de l'impossibilité de localiser[2], il existe très peu de preuves méthodologiques du succès relatif et de la rentabilité des différentes procédures de dépistage.

La présente communication, qui rend compte de l'efficacité globale des procédures de dépistage de la MCS, vise à faire progresser les travaux dans ce domaine de deux façons; tout d'abord, grâce à l'examen de la proportion des membres mobiles de l'échantillon qui sont localisés avant le début des travaux sur le terrain comparativement à pendant les travaux sur le terrain pour une vague et, en deuxième lieu, grâce à l'évaluation de l'association entre les variables démographiques et socioéconomiques individuelles et la propension à être localisés. Il est bien établi que ces types de variables sont liés à la propension à déménager, mais il existe très peu de données théoriques ou de preuves au sujet de leur lien avec la propension à être localisés. Dans certains cas, le dépistage est caractérisé comme un processus « subi » par les membres de l'échantillon, mais ceux-ci sont en fait des agents actifs de ce processus, et il est donc raisonnable de partir de l'hypothèse que les procédures de dépistage peuvent être plus efficaces pour certains types de personnes, qui peuvent être distinguées par leurs caractéristiques démographiques et socioéconomiques observables, que pour d'autres.

3. La Millennium Cohort Study

La Millennium Cohort Study est une étude longitudinale par cohorte de naissances qui suit plus de 19 000 enfants au Royaume-Uni qui sont nés en 2000 et 2001. L'échantillon a été tiré du registre des prestations pour enfants et a été mis en grappe géographiquement au départ, par circonscription électorale, avec une surreprésentation des régions comptant des proportions élevées de familles noires et asiatiques, de régions défavorisées et de régions situées dans les trois pays plus petits du Royaume-Uni. Les prestations pour enfants sont des prestations universelles versées aux familles qui ont des enfants, et les paiements commencent à la naissance de l'enfant. Jusqu'à maintenant, quatre vagues de l'étude ont été menées, pour les membres de la cohorte âgés de neuf mois, trois ans, cinq ans et sept ans. À toutes les vagues, des interviews ont été menées auprès des deux parents résidant avec l'enfant et, à partir de la deuxième vague, directement auprès du membre de la cohorte. L'étude a aussi permis de recueillir des données auprès des frères et sœurs et des enseignants, et comportait une demande de consentement en vue du couplage des données administratives pour les membres de la cohorte, les parents et les frères et sœurs. De plus amples renseignements concernant la conception de l'étude se trouvent dans Plewis (2007b).

La MCS utilise une gamme variée de procédures de dépistage prospectif et rétrospectif. L'étude comporte un numéro de téléphone sans frais, une adresse de courriel et un site Web, grâce auxquels les familles comprises dans la cohorte peuvent nous informer si elles changent d'adresse ou de coordonnées. Les coordonnées des membres de l'étude sont mises à jour chaque année. Les années d'enquête, cela se fait au moment de l'interview. Les autres années, cela se fait grâce à l'envoi par la poste d'un coupon de réponse sur lequel sont imprimés tous les détails des coordonnées des familles, par exemple, adresse, nom,

[2] Par exemple, dans le cadre de la Panel Survey of Income Dynamics et de la Health and Retirement Study, on a localisé avec succès de 97 à 98 % des membres de l'échantillon qui ont déménagé entre les vagues de 2003-2005 et 2002-2004 de ces études, et l'enquête socioéconomique par panel allemande et la British Household Panel ont eu des taux de dépistage de 96 % entre 2003 et 2005 et de 94 % entre 2003 et 2004, respectivement (Couper et Ofstedal, 2009).

numéro de téléphone, adresse de courriel et coordonnées d'un contact stable. Les familles doivent renvoyer le formulaire, avec des corrections et/ou des ajouts, ou doivent confirmer que les données sont correctes et complètes. Les formulaires sont renvoyés par environ 75 % des familles, après deux rappels. Les formulaires non livrés, qui montrent généralement qu'une famille a déménagé, sont renvoyés aux responsables de l'étude par le bureau de poste, ce qui constitue l'amorce d'un dépistage rétrospectif. Plusieurs tentatives sont faites pour contacter les membres de l'échantillon, la personne à contacter « stable » désignée et les occupants actuels du logement occupé précédemment par les membres de l'échantillon, par téléphone, par courrier, par courriel et par message texte. Nous utilisons aussi les répertoires des bureaux de poste, électoraux ou téléphoniques, qui sont disponibles dans Internet ou grâce à des logiciels spécialisés, ainsi que d'autres sources de données administratives, comme le National Health Service Central Register et les dossiers des prestations pour enfants. Pendant les travaux sur le terrain de l'étude, les intervieweurs tentent aussi de dépister les familles qui ont déménagé. Les intervieweurs sur le terrain peuvent effectuer des visites sur place à la dernière adresse connue des membres de la cohorte et à l'adresse des contacts stables, s'il s'agit d'une adresse locale, en plus des tentatives de contact par téléphone et par courrier. Les intervieweurs peuvent aussi tenter de dépister les familles grâce aux voisins, à des indices visuels dans la propriété, par exemple, des affiches « à vendre », qui peuvent mener au dépistage grâce aux agents immobiliers, et utiliser d'autres sources d'information disponibles localement.

Du point de vue de la gestion de l'enquête et du budget, il est beaucoup plus souhaitable de déterminer qu'un membre de l'échantillon a déménagé et, idéalement, de trouver sa nouvelle adresse avant les travaux sur le terrain pour une vague, que pendant ces travaux, le dépistage sur le terrain par les intervieweurs étant généralement plus coûteux que le dépistage assuré par le bureau et pouvant mener à des retards dans les travaux sur le terrain, en raison du délai additionnel nécessaire pour la localisation. Il restera toujours des personnes qui ont déménagé et qu'il est impossible de localiser avant le début des travaux sur le terrain, soit parce que le déménagement s'est produit après le début des travaux sur le terrain ou parce que l'intervieweur en prend connaissance seulement lorsqu'il tente d'établir le contact.

4. Résultats

La présente communication examine la mobilité entre la vague 2 (3 ans) et la vague 3 (5 ans) de l'étude et utilise les données du processus d'enquête pour déterminer si une famille a déménagé entre la vague 2 et la vague 3, si elle a été localisée, le cas échéant, et dans les cas où elle a déménagé et a été localisée, si la localisation a eu lieu avant ou après le début des travaux sur le terrain. La première partie de la présente section comporte des résultats descriptifs, et la deuxième présente des résultats de modèles statistiques qui utilisent des variables d'intérêt des données d'enquête recueillies à la vague 2 pour prédire la propension à déménager, à être localisés et à être localisés avant le début des travaux sur le terrain. C'est pourquoi l'échantillon d'analyse se limite aux familles qui ont participé à la vague 2 (15 590).

4.1 Résultats descriptifs

Le tableau 4.1-1 montre que 21 pour cent des familles qui ont collaboré à la vague 2 (MCS2) avaient déménagé à la vague 3 (MCS3). Pour un petit nombre de familles (169), il n'est pas possible de savoir avec certitude si elles ont déménagé ou non. On note une combinaison de cas non admissibles et de cas de refus et de nature délicate qui ne sont pas allés sur le terrain. Pour tous les autres cas, il est possible de savoir avec un niveau élevé de certitude s'ils ont déménagé ou non parce que, même s'ils n'ont pas participé à l'enquête, un intervieweur s'est rendu à leur adresse et a établi qu'ils y résidaient toujours ou non.

Tableau 4.1-1
Mobilité entre la MCS2 et la MCS3 pour les familles qui ont collaboré à la MCS2

	Familles qui ont collaboré à la MCS2
Ont déménagé entre la MCS2 et la MCS3	3278 (21 %)
N'ont pas déménagé entre la MCS2 et la MCS3	12 143 (78 %)
Ne sait pas s'ils ont déménagé entre la MCS2 et la MCS3	169 (1 %)
Base	15 590

La première colonne du tableau 4.1-2 montre qu'une proportion extrêmement élevée de familles mobiles (93 %) ont été localisées. Elle montre en outre que les familles mobiles étaient beaucoup moins susceptibles que les familles non mobiles d'être localisées et de collaborer à la vague 3 : 84 % comparativement à 91 %.

Tableau 4.1-2
Localisation et collaboration à la MCS3 dans le cas des familles qui ont collaboré à la MCS2, selon qu'elles ont déménagé depuis la MCS2

	Familles mobiles, c.-à-d. qui ont déménagé depuis la MCS2	Familles non mobiles, c.-à-d. qui n'ont pas déménagé depuis la MCS2
Ont été localisées et ont collaboré	2766 (84 %)	11 036 (91 %)
Ont été localisées et n'ont pas collaboré	284 (9 %)	1 107 (9 %)
N'ont pas été localisées	228 (7 %)	0 (0 %)
Base	3278	12 143

Le tableau 4.1-3 montre, qu'à condition de réussir à les localiser, les familles mobiles ne sont pas moins susceptibles que les familles non mobiles de collaborer à la vague 3 : 91 % dans les deux cas.

Tableau 4.1-3
Collaboration à la MCS3 pour les familles qui ont collaboré à la MCS2 et qui ont été localisées pour la MCS3, selon qu'elles ont déménagé depuis la MCS2

	Familles mobiles localisées, c.-à-d. qui ont déménagé depuis la MCS2 et qui ont été localisées pour la MCS3	Familles non mobiles, c.-à-d. qui n'ont pas déménagé depuis la MCS2
Ont collaboré	2766 (91 %)	11 036 (91 %)
N'ont pas collaboré	284 (9 %)	1107 (9 %)
Base	3050	12 143

Globalement, 55 % de toutes les familles mobiles ont été localisées avant le début des travaux sur le terrain, et 38 %, pendant les travaux sur le terrain (tandis que 7 % n'ont pas été localisées). Le tableau 4.1-4 montre que les familles mobiles qui ont été localisées avant le début des travaux sur le terrain étaient tout aussi susceptibles que les familles mobiles qui ont été localisées pendant les travaux sur le terrain de participer à une interview : 90 % et 91 % respectivement.

Tableau 4.1-4
Collaboration à la MCS3 pour les familles qui ont collaboré à la MCS2 et qui ont été localisées pour la MCS3, selon qu'elles ont déménagé depuis la MCS2 et le moment où elles ont été localisées

	Familles mobiles localisées, c.-à-d. qui ont déménagé depuis la MCS2 et qui ont été localisées avant le début des travaux sur le terrain de la MCS3	Familles mobiles localisées, c.-à-d. qui ont déménagé depuis la MCS2 et qui ont été localisées pendant les travaux sur le terrain de la MCS3	Familles non mobiles, c.-à-d. qui n'ont pas déménagé depuis la MCS2
Ont collaboré	1635 (90 %)	1131 (91 %)	11 036 (91 %)
N'ont pas collaboré	175 (10 %)	109 (9 %)	1107 (9 %)
Base	1810	1240	12 143

Globalement, ces résultats descriptifs montrent que la MCS comporte des taux de dépistage qui sont élevés et des procédures de dépistage qui permettent de localiser une proportion élevée de familles mobiles entre les vagues de travaux sur le terrain.

4.2 Modélisation statistique

La présente section rend compte des résultats de modèles de régression logistique qui ont servi à prédire la propension à déménager, à être localisé et à être localisé avant le début des travaux sur le terrain. Toutes les variables figurant dans le tableau étaient statistiquement significatives (test de Wald; $p<0,05$) et des intervalles de confiance de 95 % sont fournis pour toutes les catégories.

4.2.1 Prédicteurs de la mobilité entre la vague 2 et la vague 3

La présente section examine comment les caractéristiques des familles mobiles diffèrent de celles des familles non mobiles. Une gamme variée de facteurs géographiques, démographiques, socioéconomiques et d'attitude ont été examinés, et des résultats non corrigés et corrigés sous forme de rapports de cotes sont présentés dans le tableau 4.2.1-1.

Globalement, les résultats ne rendent pas compte d'une situation entièrement cohérente. Certains des résultats montrent que les familles moins avantagées étaient plus susceptibles de déménager que les familles plus avantagées. Les familles qui étaient les plus susceptibles de déménager étaient celles dont la mère était jeune (moins de 25 ans), celles ayant un enfant plus jeune que celui de la cohorte (3 ans) et celles qui vivaient dans un logement loué dont elles étaient insatisfaites à la vague 2. Toutefois, d'autres indicateurs de la situation socioéconomique ont montré que les familles plus avantagées, c'est-à-dire celles se situant au-dessus du seuil de pauvreté et celles comptant des nombres plus élevés de véhicules, étaient plus susceptibles de déménager[3]. En outre, même si le fait d'avoir un autre enfant plus jeune que l'enfant de la cohorte était associé à une propension plus grande à déménager, le fait d'avoir d'autres enfants, outre un enfant de la cohorte (plus d'un enfant dans la famille), était associé à une plus faible propension à déménager.

Le modèle de régression a aussi montré que les familles d'Écosse étaient légèrement moins susceptibles de déménager que les familles d'Angleterre et que les familles qui étaient insatisfaites de la région où elles vivaient étaient plus susceptibles de déménager que celles qui étaient satisfaites. Les familles dont la mère appartenait à un groupe de minorité ethnique, quel qu'il soit, sauf mixte, étaient moins susceptibles de déménager que celles dont les mères étaient de race blanche dans les statistiques non corrigées, même si seulement celles dont les mères étaient de race noire étaient significativement moins susceptibles de le faire dans les modèles. Les familles qui étaient passées de deux parents à la vague 1 à un parent à la vague 2 étaient plus susceptibles de déménager que les familles qui étaient demeurées monoparentales dans les deux vagues, ce qui rend peut-être compte des répercussions tardives de la fin d'une relation.

Parmi les autres variables qui ont été incluses dans le modèle, mais qui ne comportent pas de rapport significatif avec la mobilité, figuraient le type de famille, le niveau de scolarité de la mère et la situation d'emploi du ménage.

Tableau 4.2.1-1
Pourcentage des familles qui ont collaboré à la MCS2 qui ont déménagé entre la MCS2 et la MCS3 et rapports de cotes du déménagement à partir d'un modèle de régression logistique, selon les variables de la MCS2

Variable de la MCS2	% non corrigé de personnes qui ont déménagé	Rapports de cotes (RC)	Intervalle de confiance de 95 % pour les RC
Pays			
Angleterre	21,0	1	Fixe
Pays de Galles	17,5	0,87	(0,73 , 1,02)
Écosse	24,0	1,16	(1,01 , 1,35)
Irlande du Nord	19,5	1,14	(0,95 , 1,37)
Âge de la mère			
16 à 24 ans	33,3	1	Fixe
25 à 29 ans	26,2	0,89	(0,75 , 1,05)
30 à 34 ans	20,6	0,76	(0,64 , 0,91)
35 à 39 ans	15,3	0,58	(0,46 , 0,72)
40 ans et plus	14,1	0,51	(0,40 , 0,68)
Groupe ethnique de la mère			
Blanc	21,0	1	Fixe

[3] Il est intéressant de constater que la direction du rapport entre la mobilité et ces deux variables est inverse dans le modèle statistique comparativement à l'analyse unidimensionnelle.

Mixte	27,9	1,14	(0,68 , 1,91)
Indien	16,6	0,82	(0,57 , 1,16)
Pakistanais et Bangladais	18,6	0,80	(0,63 , 1,02)
Noir ou Noir britannique	19,7	0,62	(0,46 , 0,83)
Autre	21,7	0,93	(0,55 , 1,57)
Nombre d'enfants dans le ménage (y compris le membre de la cohorte)			
Un	24,2	1	Fixe
Deux	20,5	0,85	(0,74 , 0,99)
Trois	18,7	0,79	(0,66 , 0,95)
Quatre ou plus	18,9	0,71	(0,56 , 0,91)
Membre de la cohorte ayant des frères et sœurs plus jeunes			
Frères et sœurs plus jeunes	24	1,34	(1,17 , 1,53)
Pas de frères et sœurs plus jeunes	19,8	1	Fixe
Changement familial depuis la MCS1			
Même famille biparentale	18,4	0,29	(0,08 , 1,11)
Famille biparentale à famille monoparentale	34,9	1	Fixe
Famille monoparentale à famille biparentale	28,9	0,28	(0,07 , 1,06)
Même famille monoparentale	28,1	0,67	(0,52 , 0,86)
Autre	39,7	0,45	(0,12 , 1,71)
Pauvreté de la famille			
Inconnue	22	1,17	(0,97 , 1,42)
Supérieure à la médiane de 60 %	19,6	1,29	(1,10 , 1,51)
Inférieure à la médiane de 60 %	24,6	1	Fixe
Mode d'occupation du logement			
Propriétaire	16,3	1	Fixe
Locataire	29,9	1,49	(1,23 , 1,79)
Autre	38,9	2,53	(1,92 , 3,35)
Type de logement			
Maison	19,1	1	Fixe
Appartement	39,6	1,87	(1,58 , 2,22)
Propriété d'une automobile			
Aucune	28,1	1	Fixe
Une	22,2	1,24	(1,06 , 1,46)
Deux	17,7	1,39	(1,14,1.70)
Trois ou plus	20,5	1,53	(1,12,2.10)
Satisfaction à l'égard du logement			
Très satisfait	14,0	1	Fixe
Assez satisfait	21,6	1,51	(1,33 , 1,70)
Ni satisfait ni insatisfait	34,0	2,30	(1,89 , 2,79)
Assez insatisfait	33,0	2,25	(1,78 , 2,85)
Très insatisfait	46,7	3,62	(2,82 , 4,64)
Satisfaction à l'égard de la région			
Très satisfait	17,2	1	Fixe
Assez satisfait	21,5	1,10	(0,99 , 1,23)
Ni satisfait ni insatisfait	30,7	1,43	(1,01 , 1,86)
Assez insatisfait	32,0	1,53	(1,23 , 1,90)
Très insatisfait	38,4	1,48	(1,16 , 1,90)

4.2.2 Prédicteurs de la localisation à la vague 3, conditionnels à la mobilité

La présente section examine comment les caractéristiques des familles mobiles qui sont localisées diffèrent de celles des familles mobiles qui ne sont pas localisées. Le même ensemble de facteurs géographiques, démographiques, socioéconomiques et d'attitude examinés dans les sections précédentes sont évalués ici, et les résultats sous forme de rapports de cotes sont présentés dans le tableau 4.2.2-1. Comme il est démontré dans la section 2, on est parti de l'hypothèse que les procédures de dépistage peuvent être plus efficaces pour certains types de familles que pour d'autres.

Dans l'ensemble, les résultats vont très peu dans le sens de cette hypothèse, les seules variables représentant des prédicteurs significatifs d'être localisés étant l'origine ethnique de la mère et le type de logement. Les familles dans lesquelles la mère appartient à un groupe ethnique non blanc, sauf mixte, sont beaucoup plus susceptibles d'être localisées que celles dont la mère est de race blanche, et celles qui vivent en appartement sont moins susceptibles d'être localisées que celles qui vivent dans une maison. Les procédures de dépistage de l'étude sont de toute évidence moins efficaces pour les groupes de minorité ethnique, ce qui peut avoir un lien avec les obstacles linguistiques, et pour les personnes vivant en appartement, ce qui peut avoir un lien avec la stabilité résidentielle plus faible de cette partie du secteur du logement et avec les obstacles au contact, comme les numéros de téléphone.

Tableau 4.2.2-1
Rapports de cotes de la localisation, conditionnels au déménagement, à partir d'un modèle de régression logistique, selon les variables de la MCS2

Variable de la MCS2	Rapports de cotes (RC)	Intervalle de confiance de 95 % pour les RC
Groupe ethnique de la mère		
Blanc	1	Fixe
Mixte	0,41	(0,14 , 1,24)
Indien	0,19	(0,07 , 0,51)
Pakistanais et Bangladais	0,29	(0,13 , 0,63)
Noir ou Noir britannique	0,21	(0,10 , 0,44)
Autre	0,08	(0,03 , 0,19)
Type de logement		
Maison	1	Fixe
Appartement	0,60	(0,36 , 0,99)

4.2.3 Prédicteurs de la localisation avant le début des travaux sur le terrain pour la vague 3, conditionnels à la mobilité et à la localisation

La présente section examine comment les caractéristiques des familles mobiles qui ont été localisées avant le début des travaux sur le terrain pour la vague 3 diffèrent de celles des familles mobiles qui ont été localisées pendant les travaux sur le terrain pour la vague 3. Le même ensemble de facteurs géographiques, démographiques, socioéconomiques et d'attitude examinés dans les sections précédentes sont évalués ici, et les résultats sous forme de rapports de cotes sont présentés dans le tableau 4.2.3-1. Comme il est démontré dans la section 2, on est parti de l'hypothèse que les procédures de dépistage qui donnent lieu à la localisation des familles avant le début des travaux sur le terrain peuvent être plus efficaces pour certains types de familles que pour d'autres.

Dans l'ensemble, les résultats vont très peu dans le sens de cette hypothèse, les seules variables représentant des prédicteurs significatifs de la localisation avant le début des travaux sur le terrain étant l'âge de la mère, le nombre d'enfants et l'enfant de la cohorte ayant un frère ou une sœur plus jeune. Les familles dont les mères sont plus âgées et les frères et sœurs plus jeunes sont plus susceptibles d'être localisées avant le début des travaux sur le terrain, et les familles comptant un plus grand nombre d'enfants sont moins susceptibles d'être localisées avant le début des travaux sur le terrain.

Tableau 4.2.3-1
Rapports de cotes de la localisation avant le début des travaux sur le terrain de la MCS3, conditionnels au déménagement et à la localisation, à partir d'un modèle de régression logistique, selon les variables de la MCS2

Variable de la MCS2	Rapports de cotes (RC)	Intervalle de confiance de 95 % des RC
Âge de la mère		
16 à 24 ans	1	Fixe
25 à 29 ans	1,44	(1,05 , 1,98)
30 à 34 ans	1,59	(1,12 , 2,26)
35 à 39 ans	2,01	(1,42 , 3,01)
40 ans et plus	2,35	(1,39 , 3,98)
Nombre d'enfants dans le ménage (y compris le membre de la cohorte)		
Un	1	Fixe
Deux	0,82	(0,64 , 1,05)
Trois	0,62	(0,46 , 0,82)
Quatre ou plus	0,72	(0,48 , 1,09)
Membre de la cohorte ayant des frères et sœurs plus jeunes		
Frères et sœurs plus jeunes	1,39	(1,10 , 1,76)
Pas de frères et sœurs plus jeunes	1	Fixe

5. Discussion

La présente communication, qui a été motivée par Couper et Ofstedal (2009), a démontré que tout comme de nombreuses autres enquêtes longitudinales majeures, la Millennium Cohort Study comporte des procédures très efficaces pour maintenir le contact avec les familles mobiles. Plus de neuf familles sur dix (93 %) qui ont déménagé entre la vague 2 et la vague 3 ont été localisées, dont plus de la moitié (55 %), avant le début des travaux sur le terrain. Elle montre en outre qu'à condition que l'on ait réussi à les localiser, les familles mobiles n'étaient pas moins susceptibles de collaborer que les familles non mobiles, c'est-à-dire que les efforts de dépistage mènent à des interviews. Cela, ainsi que la conclusion selon laquelle les familles mobiles comportent des caractéristiques différentes des familles non mobiles, fournit une justification scientifique pour les ressources consacrées par l'étude au dépistage.

Comme il fallait s'y attendre, plusieurs caractéristiques démographiques et socioéconomiques étaient liées à la mobilité résidentielle. Toutefois, très peu de ces facteurs étaient liés à la localisation réussie des familles mobiles, que ce soit globalement et avant le début des travaux sur le terrain. Cela est rassurant, étant donné que cela démontre que les procédures de dépistage de l'étude n'échouent pas systématiquement lorsqu'il s'agit de joindre certains types de répondants, sauf les familles de groupes de minorité ethnique.

Les prochaines étapes prévues pour cette recherche sont l'utilisation des données sur le moment du déménagement de la vague 3 (disponibles pour les familles répondantes seulement), afin d'estimer la proportion des familles mobiles qui déménagent avant le début des travaux sur le terrain et qui sont localisées au moyen des procédures de dépistage de l'étude avant le début des travaux sur le terrain, et d'utiliser les données concernant la distance du déménagement (disponibles pour les familles localisées seulement), en vue de déterminer si les familles mobiles qui sont localisées avant le début des travaux sur le terrain diffèrent des familles mobiles qui sont localisées pendant les travaux sur le terrain, selon la distance du déménagement. Nous aimerions aussi utiliser les données du processus d'enquête, c'est-à-dire les données concernant la façon dont les familles mobiles sont localisées, pour tenter d'évaluer l'efficacité et l'efficience relatives des différentes procédures de dépistage.

Bibliographie

Couper, M.P. et Ofstedal, M.B. (2009). Keeping in Contact with Mobile Sample Members, in P.Lynn (ed.) *Methodology of Longitudinal Surveys,* Chichester: John Wiley & Sons, Inc, pp. 183-203.

Lepkowski, J.M. et Couper, M.P. (2002). Nonresponse in longitudinal household surveys, in R.M.Groves et al. (eds.) *Survey Nonresponse,* New York: John Wiley & Sons, Inc, pp. 259-272.

Plewis, I. (2007a). Non-response in a Birth Cohort Study: The Case of the Millennium Cohort Study, *International Journal of Social Research Methodology*, 10, pp.325-334.

Plewis, I. (ed.) (2007b). *The Millennium Cohort Study: Technical Report on Sampling (4th. ed.),* London: Institute of Education, University of London.

Plewis, I., Ketende, S.C., Joshi, H. et Hughes, G. (2008). The Contribution of Residential Mobility to Sample Loss in a British Cohort Study: Evidence from the First Two Waves of the Millennium Cohort Study, *Journal of Official Statistics,* 24, pp. 365-385.

Recueil du Symposium 2009 de Statistique Canada
Les enquêtes longitudinales: de la conception à l'analyse

Organisation et monitoring du terrain d'enquête : impact sur la qualité des estimateurs pour un panel rotatif de ménages

Thomas Christin, Stéphane Fleury, Johan Pea[1]

Résumé

En présence de non-réponse totale non ignorable et imparfaitement modélisée, les mesures permettant de maximiser les taux de réponse sont de fait susceptibles de diminuer la variance et le biais de nos estimations. Dans cette optique, nous proposons d'évaluer l'impact que peut avoir l'organisation du terrain d'enquête sur les taux de réponse et plus particulièrement le monitoring journalier du terrain d'enquête et l'anticipation du degré de coopération des ménages. Notre contexte de travail est celui de l'enquête SILC en Suisse.

Mots Clés : terrain d'enquête, monitoring d'enquête, degré de coopération, non-réponse.

1. Introduction

Comprendre au mieux les paramètres expliquant le processus de réponse est un enjeu important et récurrent des enquêtes statistiques. Pour les enquêtes longitudinales, le caractère cumulatif de la non-réponse totale augmente encore l'importance de cet enjeu. Le traitement de la non-réponse est en priorité orientée vers la correction du biais de non-réponse totale via la pondération. Nous proposons d'y ajouter la mise en place de mesures visant à maximiser les probabilités de réponse. En admettant que la non-réponse totale est non ignorable et imparfaitement modélisée, elle contribue à augmenter la variance de nos estimateurs et crée un biais malgré les ajustements apportés par la pondération. Habituellement, la probabilité de réponse est expliquée puis corrigée selon des caractéristiques individuelles inhérentes au répondant. Légitime dans une optique de pondération, cette approche ne permet guère d'identifier les facteurs influençant la probabilité de réponse qui dépendraient avant tout du travail du statisticien et sur lesquels nous pourrions agir. Dans cette idée, nous proposons d'évaluer dans quelle mesure l'organisation et le monitoring du terrain d'enquête peuvent significativement influencer les taux de réponse.

Nous commençons par présenter le contexte dans lequel cette démarche a été entreprise en abordant brièvement les caractéristiques de l'enquête SILC en Suisse (Statistics on Income and Living Conditions). L'ampleur et les caractéristiques du processus de non-réponse totale de l'enquête SILC sont également décrites. Dans la section 3, nous présentons les principes et hypothèses qui ont présidé à la mise en place d'instruments de monitoring journalier du terrain d'enquête SILC. Après une description plus précise de ces instruments, nous évaluons leur efficacité afin de comprendre dans quelle mesure un bon monitoring du terrain est susceptible d'augmenter les probabilités de réponse. Dans la section 4, nous proposons d'évaluer comment les informations des vagues précédentes permettent d'estimer le degré de coopération des ménages. Nous examinons alors s'il peut être efficace d'utiliser cette estimation afin d'augmenter les probabilités de réponse des non coopératifs. Pour conclure, nous évaluons dans quelle mesure en optimisant la relation entre les caractéristiques des répondants et leur probabilité de réponse nous ne remplaçons pas un problème par un autre. En effet, ce que l'on gagne en termes de qualité des estimateurs en minimisant la non-réponse totale, peut être en partie reperdu via un taux de non-réponse partielle ou une erreur de réponse plus élevés chez les unités concernées.

2. L'enquête SILC

2.1 Un projet européen

L'enquête SILC est un projet européen coordonné par Eurostat sous le nom d'EU-SILC. Démarré en 2003, EU-SILC relève des informations sur plus de 250'000 individus dans plus de 25 pays. Son principal objectif est de produire annuellement des statistiques communautaires sur la distribution des revenus, les conditions de vie, la pauvreté et l'exclusion sociale avec des

[1]Thomas Christin, Office Fédéral de la Statistique, Espace de l'Europe 10, Suisse, 2010 (Thomas.Christin@bfs.admin.ch); Stéphane Fleury, Office Fédéral de la Statistique, Espace de l'Europe 10, Suisse, 2010 (Stephane.fleury@bfs.admin.ch); Johan Pea, Office Fédéral de la Statistique, Espace de l'Europe 10, Suisse, 2010 (Johan.Pea@bfs.admin.ch).

données transversales et longitudinales comparables et actualisées. L'enquête SILC a été lancée en Suisse en 2007 sous la forme d'un panel rotatif sur 4 ans.

2.2 Caractéristiques du terrain SILC en Suisse

En Suisse, nous travaillons avec un échantillon stratifié proportionnel par grande région d'environ 6'000 ménages. Ainsi, plus de 11'000 personnes sont interrogées par interview téléphonique chaque année pour SILC, soit en vague 1, 2, 3 ou 4. L'utilisation de données de registres pour certaines sources de revenus est en phase de test et de mise en place. La particularité de l'enquête SILC est d'interroger toutes les personnes âgées de 16 ans et plus dans les ménages éligibles.

Les adresses de l'échantillon brut sont réparties en 4 paquets, activés successivement dans l'enquête de façon à minimiser l'intervalle de temps entre la réception de la lettre d'annonce et le premier contact téléphonique. Un institut de sondages privé est mandaté par l'Office fédéral de la Statistique (OFS) pour réaliser les interviews. Conformément au caractère multilingue de la Suisse, les interviews sont réalisées en français, en allemand ou en italien dans deux centrales téléphoniques qui sont situées dans les deux plus grandes régions linguistiques, une en Suisse alémanique (Berne) qui gère les interviews en allemand et une en Suisse romande (Lausanne) pour le français et l'italien. Ce contexte pose la question de l'unicité de notre terrain d'enquête, ceci d'autant plus que les deux centrales d'interviews présentent des différences de tailles et de puissance d'enquête importantes. Plus de 70% des adresses sont traitées par la centrale de Berne (allemand) qui bénéficie d'un plus grand nombre d'enquêteur et d'enquêtrices ainsi que d'une infrastructure plus moderne que celle de Lausanne. De plus, on peut également noter des différences en termes de culture du travail ou de processus de recrutement entre ces deux centrales.

L'enquête SILC en Suisse comprend une articulation de plusieurs questionnaires (composition du ménage dénommée « Grille », questionnaire ménage, questionnaires individuels, enfants et proxy) avec un large éventail de thématiques abordées. La charge d'interview pour SILC en Suisse est considérée comme importante. En vague 1, les interviews pour un ménage d'une personne durent environ 40 minutes et presque 1 heure 30 minutes pour un ménage composé de 3 personnes éligibles (plus de 16 ans).

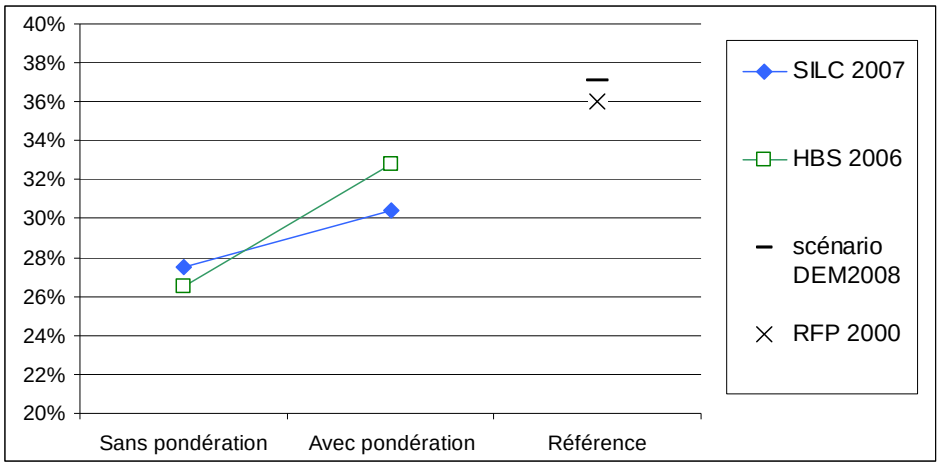
2.3 La non-réponse totale pour SILC en Suisse

Malgré un fardeau de réponse élevé, les taux de réponse en vague 1 de SILC en Suisse s'élevaient en 2007 à 68%[2], ce qui est considéré comme un bon résultat dans le contexte suisse. En comparaison européenne, la Suisse se situe au centre de la distribution des taux de réponse pour EU-SILC. Le processus de non-réponse est modélisé par segmentation pour ajuster les poids. Les principales caractéristiques pénalisant les probabilités de réponse en vague 1 sont un faible niveau de formation, le fait qu'aucun membre du ménage ne soit de nationalité suisse et le fait de vivre dans un ménage d'une seule personne. Les indicateurs principaux estimés par SILC s'attachant aux aspects de revenu, on peut sans autre admettre que les probabilités de réponse sont liées à nos variables d'intérêt entrainant ainsi une non-réponse non ignorable.

En plus d'être non ignorable, notre postulat de travail est que la probabilité de répondre est imparfaitement modélisée. En effet, la complexité du processus de réponse a pour conséquence qu'une part des variables auxiliaires explicatives de la non-réponse totale est inconnue. De plus, la probabilité est élevée qu'une partie du processus de réponse ne soit pas lié à des variables structurelles mais à des éléments subjectifs ou difficilement observables tels que l'humeur au moment de l'appel ou la nature de l'événement précédent l'appel téléphonique. Nous admettrons ainsi que le biais dû à la non-réponse ne peut pas être totalement corrigé par la pondération. Pour soutenir cette hypothèse, la figure 2.3-1 montre pour SILC et l'enquête sur le budget des ménages (HBS) les estimations du pourcentage de ménages composés d'une seule personne en Suisse avant et après pondération.

[2] Ce taux de réponse est calculé en rapportant le nombre de ménages exploitables (valides) sur le total des ménages éligibles. Les adresses non contactées pour lesquelles le statut d'éligibilité n'est pas connu sont considérées comme éligibles.

Figure 2.3-1
Différence d'estimation du pourcentage de ménages d'une personne en Suisse



Sources :
OFS, SILC 2007, HBS 2006, scénarios démographiques 2008, Recensement Fédéral de la Population (RFP) 2000.

Cette figure montre que les estimations pondérées d'un paramètre, même relativement basique, tel que le pourcentage de ménages d'une seule personne sont significativement différentes selon la source de données utilisée. En outre on sait que le risque de pauvreté, qui représente la dimension d'intérêt principale de SILC, dépend notamment de la taille du ménage.

Ces constats nous encouragent à tout faire pour optimiser, dans la mesure du possible, la relation entre le degré de coopération des individus face à la statistique et leur probabilité de réponse à l'enquête SILC. On considère le degré de coopération d'un individu face à la statistique comme étant relativement stable à court terme. Il dépend de son histoire de vie et de ses caractéristiques socio-économiques. Bien que la relation entre le degré de coopération et la probabilité de réponse soit généralement considérée comme forte, elle s'avère tout de même variable. Elle peut par exemple être influencée par l'organisation du terrain d'enquête. C'est précisément ce contexte qui nous a encouragés à investir dans le monitoring de terrain d'enquête ; l'objectif étant de minimiser la non-réponse totale qui serait due à des défauts d'organisation.

3. Monitoring du terrain d'enquête et implications sur la non-réponse totale

3.1 Objectifs et principes de base

Le constat initial qui a conduit à la mise en place du monitoring de terrain SILC en Suisse est de considérer le terrain d'enquête comme une grosse machine complexe à produire des interviews. Comme toute machinerie de production sophistiquée, notre terrain d'enquête est susceptible de cacher des défauts de réglages détériorant son niveau de performance et se mesurant par une augmentation de la non-réponse. Notre terrain d'enquête renferme ainsi un potentiel d'optimisation des réglages par exemple en terme d'adéquation entre le volume de matière en input (ménages à interroger) et le volume de main d'œuvre nécessaire (enquêteurs-trices), ou en terme de rythme d'activation des paquets d'adresses. Ainsi, la recherche des réglages optimaux via un huilage approprié au bon moment au bon endroit a pour but de diminuer la non-réponse totale due aux défauts de gestion de notre terrain d'enquête.

L'objectif du monitoring d'enquête est de nous permettre d'obtenir une vue journalière de la progression de l'enquête, de déterminer le moment adéquat pour activer les paquets d'adresses et de définir la durée d'enquête optimale. Il doit également servir à identifier rapidement les décalages de performance d'une centrale à l'autre ou d'un paquet d'adresse à l'autre et permettre de réagir rapidement en alertant les centrales d'appel de l'institut de sondage. Il permet également non seulement de visualiser concrètement les problèmes mais aussi de légitimer nos interventions face à l'institut lorsque par exemple nous demandons d'augmenter le nombre d'enquêteurs-trices dans une centrale d'interview. En visualisant les étapes clefs du terrain d'enquête, il fournit une base d'information importante pour améliorer la gestion du terrain d'une année à l'autre. En affinant nos écrans de contrôle, le monitoring d'enquête nous permet de définir empiriquement les objectifs de performance demandés à l'institut mandaté.

Les outils de monitoring sont développés sur la base de statistiques quotidiennes fournies par les centrales d'appel. Elles comportent des statuts de traitement des adresses activées dans l'enquête au niveau ménage et individu tant au sujet des

contacts initiaux, des statuts d'éligibilité, des prises de rendez-vous, des rappels de refus que de la réalisation des interviews. Ces informations sont livrées dans un format Excel avec les sous-totaux par statut de contact et de réponse et dans un format SAS, niveau ménage et individu, avec les statuts de contact et de réponse pour chacune des unités. Cela représente une masse d'information considérable de laquelle il n'est pas simple d'extraire l'information la plus pertinente. L'enjeu devient donc de développer un tableau de bord permettant de piloter le plus efficacement possible la machinerie d'enquête sans pour autant que cela exige de notre part des ressources d'analyses trop importantes. Pour cela, nous travaillons avec des macros SAS qui produisent quotidiennement des graphiques visualisant les points critiques de la progression du terrain d'enquête. Nous cherchons dans ces graphiques les signes de mauvais réglages de la machinerie d'enquête comme les différences de performance entre centrales ou entre paquets d'activation d'adresses.

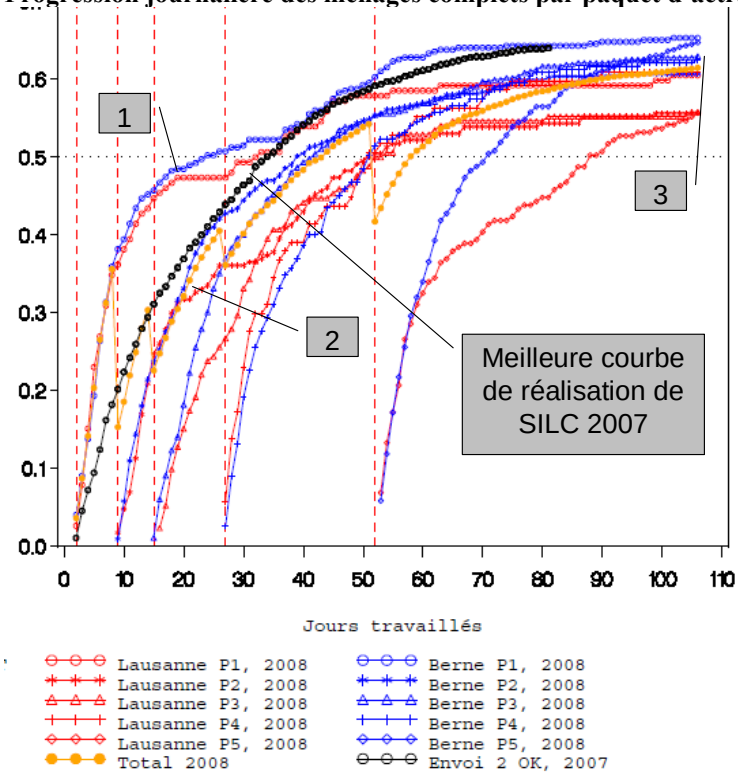
3.2 Enseignements tirés du monitoring SILC 2008 et 2009

Dans cette section, nous présentons quelques exemples des enseignements tirés du monitoring des enquêtes SILC 2008 et 2009 en Suisse. Le graphique 3.2.1 montre les taux journaliers de ménages valides[3] par paquet d'activation et par centrale d'appel pour la cohorte des ménages introduits dans le panel SILC en 2008. L'échantillon brut 2008 de vague 1 a été activé en quatre paquets successifs. Le panel SILC ayant débuté en 2007 en Suisse, l'enquête 2008 était la première à combiner des ménages interrogés en vague 1 et en vague 2. Un paquet de réserve a dû être activé en raison du taux de réponse 2008 légèrement inférieur à celui estimé sur la base de celui observé en 2007.

La figure 3.2-1 montre une progression exemplaire du premier paquet activé en 2008 les quinze premiers jours, avec une forte progression des ménages interrogés. La marque 1 identifie un problème de fléchissement de la progression au moment de l'activation des paquets 2 et 3. Cela indique une difficulté en centrale à gérer en parallèle les adresses encore à traiter des anciens paquets et les nouvelles adresses activées. Ce problème de « bourrage de la machinerie d'enquête » a été plus important dans la plus petite centrale de Lausanne (courbe rouge). Le même problème est observé à la marque 2 qui souligne un fort ralentissement de la progression du deuxième paquet au moment de l'activation du troisième, mais cette fois-ci uniquement à la centrale de Lausanne. Au final, on remarque un décalage systématique de performance entre les centrales de Berne et de Lausanne, avec des taux de ménages interrogés inférieurs à Lausanne en vague 1 en 2008 (marque 3). On voit aussi que l'activation tardive du paquet de réserve n'a pas porté à conséquence pour la centrale de Berne, mais a entraîné un grand défaut de performance à Lausanne.

[3] Pour SILC, un ménage est considéré comme valide si les questionnaires grille, ménage et au moins un questionnaire individu ont été répondus.

Figure 3.2-1
Progression journalière des ménages complets par paquet d'activation, vague 1, SILC 2008



La visualisation des problèmes rencontrés en 2008 (figure 3.2-1) et les échanges avec l'institut responsable des interviews en fin 2008 ont permis de les corriger en 2009. On remarque dans la figure 3.2-2 que les ralentissements de progression des anciens paquets d'adresses au moment de l'activation des nouveaux sont moins marqués et surtout que les différences entre les centrales de Lausanne et Berne ont été en grande partie gommées. En revanche, l'institut responsable des interviews a été nettement plus performant sur le premier paquet d'activation que sur les suivants. Les adresses étant réparties aléatoirement entre les paquets d'activation, ces décalages s'expliquent en priorité par des problèmes d'organisation du terrain d'enquête.

La figure 3.2-3 fournit un exemple d'une bonne progression d'un terrain d'enquête sans différence de performance selon la centrale d'interview ou le paquet d'activation.

Figure 3.2-2
Progression journalière des grilles réalisées par paquet d'activation, vague 1, SILC 2009

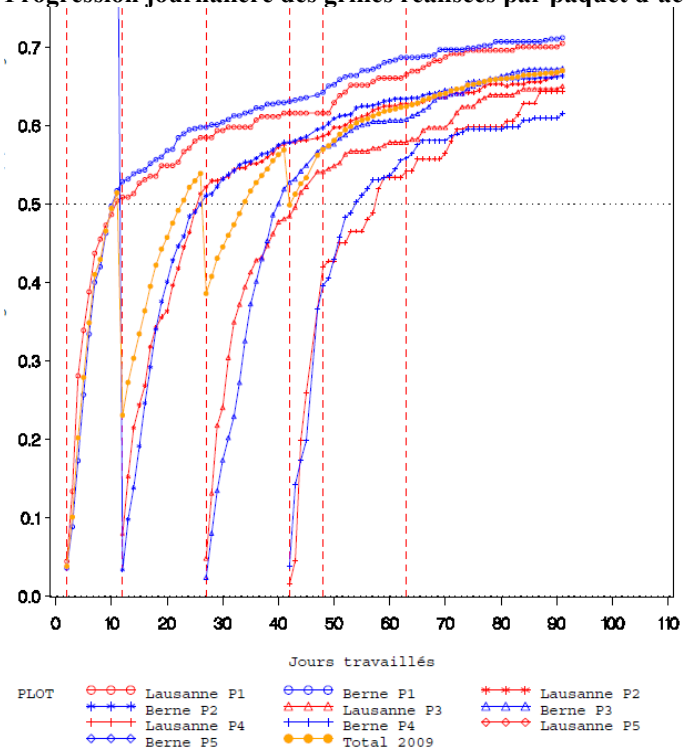
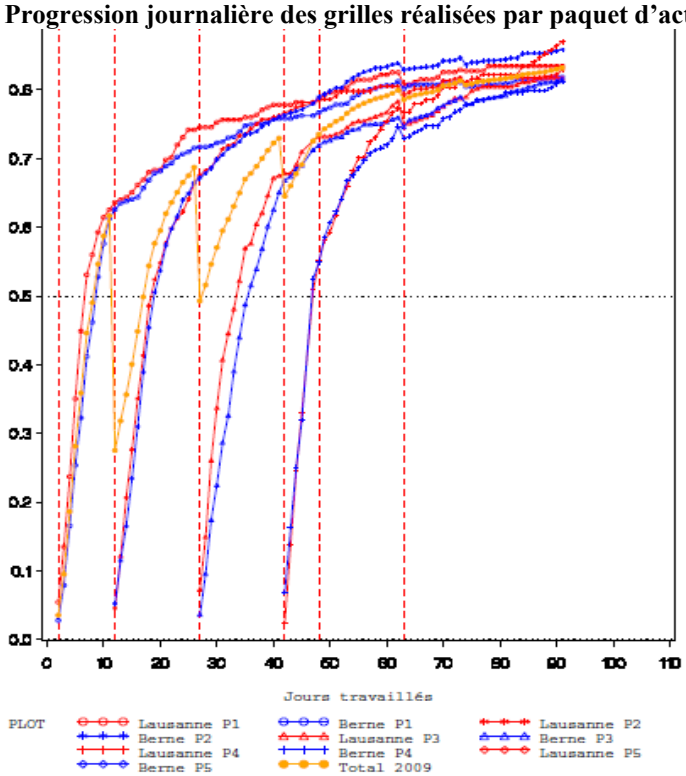


Figure 3.2.3
Progression journalière des grilles réalisées par paquet d'activation, vague 3, SILC 2009



Les premiers enseignements tirés de cet exercice de monitoring du terrain d'enquête sont les suivants :

- Le démarrage d'un paquet d'adresse semble essentiel. Un traitement trop lent d'un nouveau paquet d'adresses est très difficile à récupérer et pénalise ses taux finaux de réponse. Nous expliquons cela par le fait qu'il est préférable que les interviews se réalisent dans un délai court après la réception de l'annonce officielle de l'enquête envoyé par courrier

aux ménages sélectionnés et qu'il est difficile pour l'institut de gérer par la suite en parallèle les adresses pas encore traitées des anciens paquets et les adresses fraîches des nouveaux paquets.

- La taille des paquets d'adresses et le rythme d'activation sont des enjeux importants pour éviter les risques d'engorgement des centrales d'appel et permettre à celles-ci de tirer le maximum des échantillons bruts et de maximiser ainsi les taux de réponse.
- La mise en adéquation efficace des ressources investies en centrale d'appel et le volume d'adresses à contacter, d'interviews ou de rappels de refus à réaliser est plus difficile à optimiser dans les petites centrales d'appels.
- Dans un panel rotatif, l'optimisation du terrain d'enquête semble plus difficile pour les ménages en première vague. Lors des vagues suivantes, les ménages les moins coopératifs ont en grande partie déjà été perdus lors des vagues précédentes. Alors que la performance de la première vague est particulièrement importante, il s'agit de celle qui est la plus difficile à optimiser. C'est un enjeu important puisque la non-réponse totale en vague 1 due à des défauts du terrain d'enquête pénalise la qualité statistique de la cohorte durant toute son existence dans le panel.
- Le facteur humain lié soit aux responsables de centrale et superviseurs soit aux enquêtrices ou aux enquêteurs ne doit pas être négligé pour expliquer des différences de performance entre les centrales d'appels.

3.3 Anticipation du degré de coopération d'un ménage face à la statistique

Les probabilités de réponse peuvent être maximisées lors de la première vague d'un panel rotatif au travers notamment d'une optimisation du terrain d'enquête, de dédommagements proposés aux répondants ou encore en ciblant les courriers envoyés aux ménages sélectionnés. Si grâce à ces différentes stratégies, ces ménages peu coopératifs face à la statistique participent à l'enquête, cela signifie que les mesures prises ont aidé à obtenir leur participation à l'enquête. Pour une enquête longitudinale, cette optimisation lors de la première vague augmente le risque d'érosion lors des vagues suivantes de part l'entrée de ménages peu coopératifs face à la statistique dans la phase longitudinale de l'enquête. Pour minimiser ce problème, nous avons cherché à estimer le degré de coopération des ménages et tenté de mettre en place des mesures spéciales pour les ménages qui semblaient présenter des risques de refus particulièrement élevés pour la vague suivante.

Plusieurs options sont possibles pour estimer ce degré de coopération à partir de la deuxième vague :

1. Utiliser le nombre d'appels cumulés nécessaire lors de la vague précédente pour obtenir les interviews du ménage. Un nombre d'appels élevé peut être synonyme de refus déguisé, mais aussi d'une faible disponibilité du ménage ou encore d'une mauvaise gestion des appels en centrales. N'étant pas assez univoque, cette information n'a pas été retenue.
2. Exploiter les évaluations subjectives des enquêtrices et enquêteurs quant au degré de coopération de la personne interrogée.
3. Utiliser le pourcentage de non-réponse partielle en vague précédente. Les ménages qui présentent les pourcentages de non-réponse partielle les plus élevés seraient alors considérés comme des ménages peu coopératifs face à la statistique et donc plus susceptibles de refuser de répondre à la vague suivante. Nous avons distingué deux types de ménages peu coopératifs selon ce critère de non-réponse partielle. Le premier regroupait les ménages valides avec plus de 50% des individus éligibles au questionnaire individuel qui ont refusé d'y répondre (critère de la non-réponse totale unitaire). Le second regroupe les ménages avec un volume élevé de non-réponse partielle.

La comparaison entre les options 2 et 3 indique que les évaluations subjectives des enquêteurs-trices varient peu. Ces derniers-ères étant généralement trop positifs-ves dans leur évaluation subjective du degré de coopération de l'individu qu'ils viennent d'interroger. Nous avons donc retenu la troisième option pour estimer le degré de coopération. Nous n'avons pas priorisé les variables selon leur importance ni selon leur degré de difficulté ou d'intimité. Le nombre total de non-réponses partielles cumulées sur le questionnaire ménage et les questionnaires individus a été rapporté sur le nombre total de questions effectivement posées lors de ces questionnaires. Ce nombre total varie selon la situation des ménages, plus concrètement selon les filtres qui gèrent l'enchaînement logique des questions.

Les ménages qui présentaient plus de 50% de non-réponse totale unitaire et ceux qui comportaient les plus grands pourcentages de non-réponse partielle en 2007 ont été livrés aux centrales d'appel avec une indicatrice, une forme de drapeau rouge pour la vague suivante. Ces ménages supposés peu coopératifs ont été gérés séparément par une équipe d'enquêteurs-trices spécialisés, ceci dès les premières tentatives de contact en 2008. Si lors du premier contact ces ménages ont refusé de participer, aucun rappel de refus n'a été organisé.

Tableau 3.3-1
Taux de participation des ménages en 2008 selon le degré de participation estimé en 2007

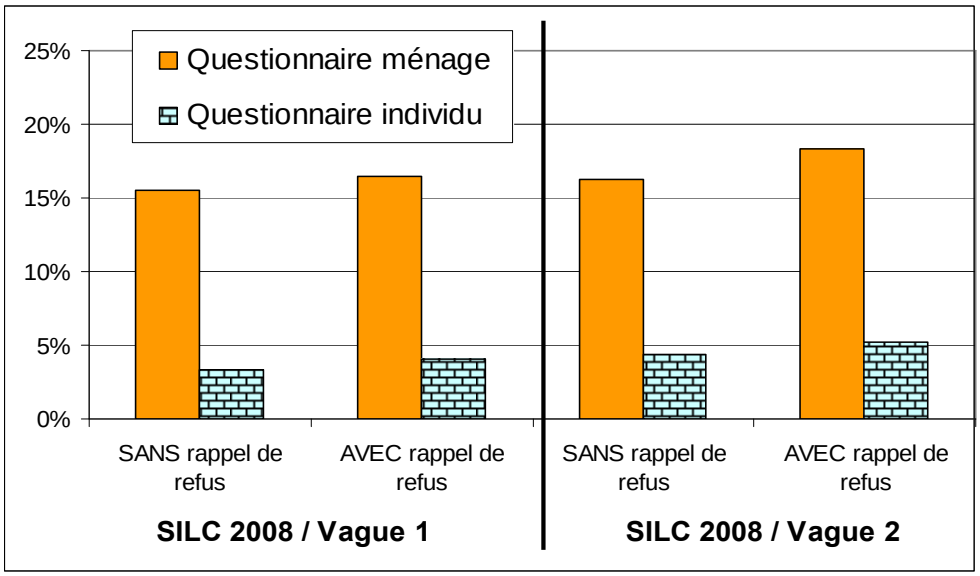
Niveau de coopération estimé en 2007 selon NR partielle (V1)	Taux de participation en 2008 (V2)
Ménages coopératifs	84%
Non coopératifs / NR partielle	74%
Non coopératifs / NR totale unitaire	56%

Ces résultats montrent que la variable de coopération construite sur la base de la non-réponse partielle à la vague T-1 est performante pour estimer les risques de refus en vague T. A notre surprise, la présence importante de personnes ayant refusé le questionnaire individu au sein des ménages valides augmente d'avantage l'érosion longitudinale que la forte présence de non-réponse partielle. La contamination des membres coopératifs du ménage par les non coopératifs semble être un facteur important pour expliquer l'érosion longitudinale. Toutefois, sans groupe de contrôle avec une gestion similaire des ménages coopératifs et non coopératifs, il est très difficile de juger de la performance des mesures prises en 2008 pour les non coopératifs.

3.4 Coût de la minimisation de la non-réponse totale en termes de non-réponse partielle

Les investissements consentis pour améliorer l'organisation du terrain d'enquête via un monitoring d'enquête permettent de réduire la non-réponse totale. A notre sens, cela s'explique par une optimisation de la relation entre les caractéristiques des répondants et leur probabilité de réponse. Dans un contexte où la non-réponse totale est non ignorable et imparfaitement modélisée, toute réduction de la non-réponse totale se traduit par une baisse de la variance et du biais de nos estimateurs. Néanmoins, ce gain pourrait être en partie reperdu si les ménages acceptant de participer suite à la prise de ces mesures accumulaient d'avantage de non-réponse partielle que les autres, augmentant ainsi le nombre de valeurs à imputer. Evaluer cette question en utilisant le degré de coopération construit sur la base du pourcentage de non-réponse partielle en vague précédente n'aurait pas été suffisant et en partie tautologique. Cela aurait plutôt permis de tester si le processus de non-réponse partielle reste stable d'une vague à l'autre. Nous proposons donc d'utiliser la présence de rappel de refus comme indicateur de non coopération pour évaluer cette question. La figure 3.3-1 présente, pour les ménages valides, le taux moyen de questions non répondues par rapport au nombre de questions effectivement posées, ceci en fonction de l'existence ou non d'un rappel de refus. Cette figure montre qu'en moyenne la non-réponse partielle est légèrement plus importante dans les cas où l'interview a été réalisée suite à un rappel de refus, en particulier pour le questionnaire ménage. La différence n'est toutefois pas très marquée.

Figure 3.3-1
Pourcentage de non-réponse partielle selon la présence ou non de rappel de refus, SILC2009



4. Conclusion

La machinerie de production des interviews est difficile à optimiser. Les défauts de réglage possibles sont multiples et susceptibles d'introduire de nouvelles dimensions structurelles à la non-réponse totale non liées aux caractéristiques des répondants. Un monitoring journalier du terrain d'enquête par centrales d'appel et paquet d'activation d'adresses est utile à la fois pour visualiser les problèmes et pour tenter de modifier l'organisation du terrain d'enquête afin de maximiser sa performance générale.

Pour les enquêtes longitudinales, l'estimation du degré de coopération des ménages au travers du taux de non-réponse partielle observé lors de la vague précédente est pertinente. L'identification des ménages les moins coopératifs est fortement appréciée par les enquêteurs-trices qui peuvent mieux se préparer avant de les contacter lors de la vague suivante. Toutefois, en l'absence de groupe test, il est difficile d'évaluer l'impact des mesures mises en place pour maximiser les taux de réponse des ménages supposés non coopératifs. Il est également possible que la gestion séparée de ces ménages par des enquêteurs-trices spécialisés ait des effets pervers sur le taux de participation. En sachant à l'avance qu'il s'agit d'un ménage avec un risque élevé de refus, les enquêteurs-trices sont peut-être trop conditionnés à l'idée d'essuyer un refus et peuvent ainsi les guider inconsciemment vers un tel refus. D'autres mesures sont en cours d'évaluation pour minimiser l'érosion plus élevée pour ces ménages, à commencer par le fait de ne plus les identifier ni les gérer séparément dans les centrales d'appel. Se focaliser uniquement sur les ménages les moins coopératifs n'est peut-être pas suffisant. Il est possible que des mesures visant les ménages moyennement coopératifs soient plus susceptibles d'influencer leur probabilité de réponse.

Si l'érosion longitudinale est influencée par la présence importante de non-réponse partielle en vague précédente, l'effet de contamination des membres coopératifs d'un ménage par les non coopératifs est un facteur nettement plus important pour expliquer l'érosion longitudinale.

L'hypothèse d'une non-réponse totale non ignorable imparfaitement modélisée étant très probable, les efforts consentis pour optimiser l'organisation du terrain d'enquête sont légitimes, ceci d'autant plus que la minimisation de la non-réponse totale n'augmente pas significativement le problème de la non-réponse partielle et des imputations. Ainsi, la diminution du biais et de la variance obtenue par une minimisation de la non-réponse totale ne se repaie pas du côté des imputations.

Une stratégie de collecte dynamique pour l'Enquête sur la dynamique du travail et du revenu

Tracy Tabuchi, François Laflamme, Owen Phillips, Milana Karaganis, et Amélie Villeneuve[1]

Résumé

L'Enquête sur la dynamique du travail et du revenu (EDTR) est une enquête longitudinale qui mesure les variations du bien-être économique des Canadiens et les facteurs qui pourraient les influencer. Les interviews prennent la forme d'interviews téléphoniques assistées par ordinateur (ITAO). Au fil des ans, on a lancé un certain nombre d'initiatives pour mieux gérer les ressources et les efforts en matière de collecte des données. Malgré les efforts déployés pour mieux gérer la collecte, les taux de réponse à l'EDTR ont baissé régulièrement au cours des dernières années.

Dans le cadre de l'Enquête sur les ménages et l'environnement (EME), on a procédé à un essai pilote d'une stratégie de collecte dynamique pour 2009, la stratégie de collecte ayant été modifiée à partir d'une analyse des paradonnées d'entrée. Dans le cadre de l'EDTR, on mettra en œuvre une stratégie similaire pour la collecte de 2010. La présente communication décrit comment on peut mettre en œuvre la stratégie de collecte dynamique dans le cadre de l'EDTR, en fonction de ce qui a déjà été développé.

Mots-clés : stratégie de collecte dynamique, enquête longitudinale, paradonnées, modèle de propension à répondre.

1. Introduction

1.1 Aperçu de l'EDTR

L'Enquête sur la dynamique du travail et du revenu (EDTR) est une enquête longitudinale. Les données servent à mesurer les variations du bien-être économique des Canadiens et les facteurs susceptibles de les influencer. À cette fin, l'EDTR recueille des données sur l'emploi et la situation familiale, les périodes sans travail, ainsi que les salaires et le revenu.

Des interviews sont menées chaque année sous forme d'interviews téléphoniques assistées par ordinateur (ITAO). Pour chaque cycle de collecte, les données sont recueillies auprès de deux panels longitudinaux simultanément. Pour déterminer qui est sélectionné pour faire partie du panel longitudinal de l'EDTR, on choisit d'abord un sous-échantillon d'environ 17 000 ménages qui répondent à l'Enquête sur la population active (EPA). Puis, toutes les personnes membres de ces ménages deviennent membres du panel longitudinal de l'EDTR. Les membres d'un panel longitudinal sont suivis pendant six cycles de collecte. Ils sont interviewés même s'ils n'appartiennent plus aux ménages sélectionnés à l'origine dans l'EPA. Les panels comportent des dates de début échelonnées; tous les trois ans, un nouveau panel est sélectionné, et l'ancien panel est éliminé.

1.2 Réponse à l'enquête et effort de collecte

Les taux de réponse à l'EDTR ont diminué constamment depuis la première collecte en 1994. Non seulement assiste-t-on à une tendance à la baisse dans les taux de réponse pour un panel donné au fil du temps, mais on note un effet de « nouveau panel ». Le taux de réponse pour le premier cycle de collecte d'un panel est plus faible que pour le panel précédent (voir le tableau 1.2-1).

Dans le cadre des enquêtes longitudinales de Statistique Canada, on a noté des augmentations de l'effort requis pour résoudre les cas. Pour ce qui est de l'EDTR, un cas est considéré comme résolu si l'interview est complète ou partiellement complète, s'il y a un refus catégorique, ou si le ménage est identifié comme hors du champ de l'enquête pour le cycle de collecte. L'augmentation de l'effort requis correspond à une hausse du temps moyen consacré à chaque cas. Dufour (2009) a déterminé que cela est attribuable à trois facteurs : le nombre de cas où il n'y a pas de contact avec un ménage; le nombre de cas nécessitant la conversion d'un refus; et le nombre de tentatives d'appels requises pour résoudre un cas. Étant donné que les

[1]Tracy Tabuchi, François Laflamme, Owen Phillips, Milana Karaganis, et Amélie Villeneuve, Statistique Canada, Ottawa, Ontario, Canada, K1A 0T6 (email contact: tracy.tabuchi@statcan.gc.ca)

budgets d'enquête sont limités et qu'il y a eu augmentation de l'effort requis pour obtenir les taux de réponse visés, il est clair qu'il faut trouver des façons de recueillir les données de façon plus efficace, tout en maintenant leur qualité.

Tableau 1.2-1
Taux de réponse longitudinaux selon le panel et le cycle de collecte

	Cycle de collecte					
Panel (année du début du panel)	1	2	3	4	5	6
Panel 1 (1994)	93,3	89,6	86,5	83,9	82,6	81,5
Panel 2 (1997)	89,5	86,8	85,2	82,7	78,5	77,4
Panel 3 (2000)	83,9	83,0	83,0	79,6	76,4	73,7
Panel 4 (2003)	81,2	83,2	78,3	75,0	71,6	68,9
Panel 5 (2006)	78,8	80,6	77,3	72,8	…	…
Panel 6 (2009)	71,0	...	...	...	...	...

... sans objet

À l'heure actuelle, pour l'EDTR et d'autres enquêtes de Statistique Canada, la stratégie de collecte est établie avant la collecte. Chaque bureau chargé de la collecte des données convient de fournir un certain niveau d'effort de collecte, du point de vue du nombre d'heures des intervieweurs et des dépenses auxiliaires. Puis, chacun est chargé d'obtenir les taux de réponse visés en répartissant cet effort le plus efficacement possible. Groves et Heeringa (2006) ont démontré qu'il est possible de réduire les coûts par unité liés à la collecte dans le cas des interviews sur place assistées par ordinateur grâce à une approche adaptative en ce qui a trait à la collecte des données. Une telle approche utilise des paradonnées, c'est-à-dire des données relatives au processus d'enquête, pour modifier la stratégie de collecte. Les répercussions d'une telle approche sont les plus grandes aux dernières étapes de la collecte, lorsque plus d'efforts sont requis pour finaliser les cas.

Mohl et Laflamme (2007) ont proposé plusieurs options pour intégrer des approches adaptatives à la collecte des données à Statistique Canada. Celles-ci ont été intégrées à la stratégie de collecte dynamique qui a fait l'objet d'un essai pilote pour la collecte des données de 2009 d'une autre enquête par ITAO, l'Enquête sur les ménages et l'environnement (EME) (Laflamme et Karaganis, 2010). La stratégie de collecte dynamique de l'EDTR repose dans une large mesure sur ce qui a été fait dans le cadre de l'EME.

Dans la présente communication, la section 2 résume les outils de gestion de la collecte utilisés actuellement pour l'EDTR. Puis, la section 3 donne un aperçu de la stratégie de collecte dynamique proposée. Enfin, la section 4 explique comment cette stratégie fonctionnera pour l'EDTR et ce qui a été fait jusqu'à maintenant pour se préparer en vue de sa mise en œuvre pour la collecte des données de 2010.

2. Outils actuels de gestion de la collecte

2.1 Application d'une limite du nombre d'appels et tranches de temps

Plusieurs pratiques et initiatives ont été mises en place pour mieux gérer les ressources dans les bureaux de collecte des données. À Statistique Canada, on impose une limite au nombre de tentatives d'appels permises pour les unités échantillonnées. Dans le cas de l'EDTR et d'autres enquêtes longitudinales, cette limite du nombre d'appels a été fixée à 40. Chaque fois qu'un numéro de téléphone est composé, peu importe les résultats, il est compté comme une tentative d'appel. Par exemple, si un intervieweur appelle un ménage et obtient un signal occupé, cela est compté comme une tentative d'appel. Toutefois, les tentatives de dépistage, c'est-à-dire les tentatives d'appel effectuées pour dépister un ménage, ne sont pas comptées. Cette pratique aide à gérer les efforts d'interview plus efficacement et à faire en sorte que les répondants ne soient pas inondés d'appels des bureaux chargés de la collecte des données.

Les tranches de temps et groupes de tranches de temps servent à optimiser les chances de contacter des membres du ménage. Les tranches de temps permettent de répartir la semaine en plages, et les tentatives d'appel sont réparties entre ces plages selon les caractéristiques démographiques des ménages, qui servent à définir les groupes de tranches de temps. On utilise quatre groupes de tranches de temps pour l'EDTR : 1) ménages composés de répondants âgés; 2) ménages composés uniquement de jeunes répondants célibataires; 3) ménages qui n'ont pas été résolus, ou refus à la fin de la dernière collecte; et 4) tous les autres ménages. La répartition visée de 40 tentatives d'appel pour chaque groupe de tranches de temps est présentée dans le

tableau 2.1-1. Si une tentative d'appel aboutit à un cas sans réponse, les tranches de temps auront un effet sur la prochaine tentative d'appel.

Tableau 2.1-1
Répartition des tentatives d'appel par groupe de tranche de temps

	Lundi - vendredi				Samedi		Dimanche	
Groupe de tranches de temps	Début - 12 h	12 h - 16 h	16 h - 19 h	19 h - Fin	Début - 12 h	12 h - Fin	Début - 16 h	16 h - Fin
Personnes âgées	14	6	12	2	2		2	2
Jeunes célibataires	5	2	10	13	2	3	2	3
Dépistage/refus	-	-	-	-	-	-	-	-
Autres	4	5	13	11	1	2	2	2

2.2 Groupes prioritaires

L'application d'ITAO permet aux gestionnaires d'enquête de déterminer les groupes prioritaires dans les cas en cours, afin de s'assurer d'obtenir des taux de réponse appropriés pour des domaines d'intérêt déterminés au préalable. Si le taux de réponse pour un domaine particulier est plus faible que prévu, les efforts peuvent être concentrés sur les ménages compris dans ce domaine. Les groupes prioritaires sont utilisés et gérés de concert avec la limite du nombre d'appels. Les domaines d'intérêt dans le cas de l'EDTR comprennent le panel, le groupe de tranches de temps et la province de résidence.

2.3 Ordonnanceur d'appels

Un ordonnanceur d'appels permet d'automatiser la façon dont les cas sont attribués aux intervieweurs, selon de nombreux paramètres, y compris le profil de l'intervieweur. Les facteurs qui déterminent les cas qui seront envoyés au prochain intervieweur disponible comprennent les suivants : la situation actuelle du cas, le nombre de tentatives déjà effectuées, le groupe de tranches de temps, la date et l'heure du dernier appel, la répartition dans la tranche de temps, etc.

3. Aperçu de la stratégie de collecte dynamique

La stratégie de collecte dynamique proposée permet de répartir le processus de collecte des données en quatre étapes : planification, collecte initiale, adoption échelonnée de la stratégie de collecte dynamique 1 et adoption échelonnée de la stratégie de collecte dynamique 2. La première étape (planification) a lieu avant le début de la collecte des données. À ce moment-là, on planifie les activités et les stratégies de collecte, on les élabore et on les met à l'essai pour les trois autres étapes de la collecte. Les principales activités comprennent l'analyse du ou des cycles précédents de collecte des données, afin de déterminer les possibilités d'amélioration, d'évaluer et de valider la base de sondage et l'échantillon, d'élaborer des outils de gestion active et d'établir des plans de dotation.

La deuxième étape (collecte initiale) comprend la première partie du processus de collecte des données; cela va de la date de début de la collecte jusqu'à ce qu'il soit déterminé que l'adoption échelonnée de la stratégie de collecte dynamique 1 doit être entreprise. Au cours de la deuxième étape, de nouvelles caractéristiques peuvent être intégrées au processus de collecte (p. ex., une limite intermédiaire aux tentatives d'appel, afin d'éviter les cas où la limite s'applique avant les dernières étapes de la collecte des données). Au cours de cette étape initiale de collecte, de nombreux indicateurs clés de la qualité, de la productivité, du coût et du potentiel de réponse pour les cas en cours font l'objet d'un contrôle étroit.

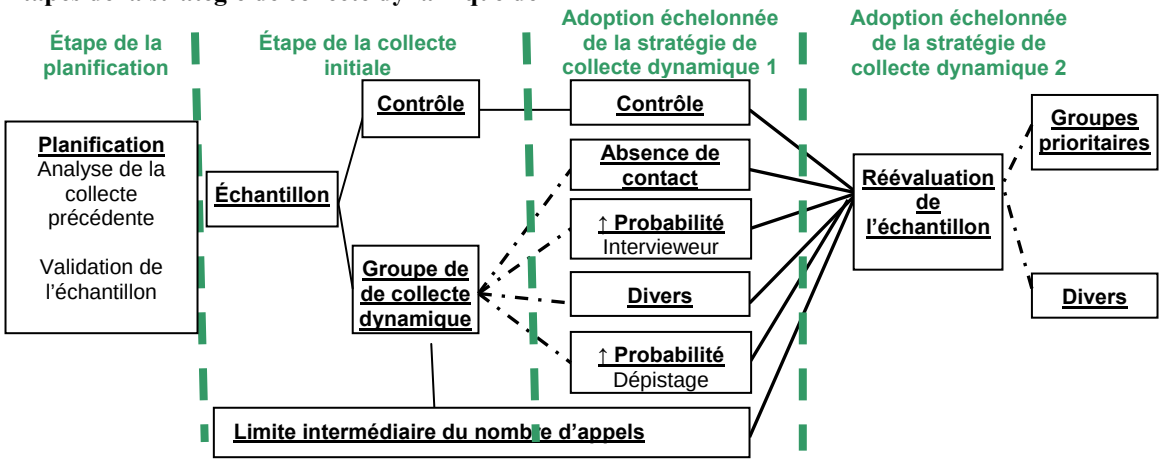
La troisième étape (adoption échelonnée de la stratégie de collecte dynamique 1) permet de répartir les cas en cours en catégories et d'établir une priorité à partir des données disponibles avant le début de la collecte et des paradonnées accumulées pendant la collecte, avec comme objectif d'améliorer les taux de réponse globaux. De façon plus particulière, un modèle de propension (régression logistique) sert à évaluer la probabilité d'être interviewée de chaque unité, ainsi qu'à catégoriser chaque cas en cours et à lui donner une priorité. Au cours de cette étape, on continue de suivre des indicateurs clés. Schouten, Cobben et Bethlehem (2009) ont déterminé la nécessité d'indicateurs de qualité, qui permettent de suivre non seulement les taux de réponse, mais aussi la représentativité de l'échantillon. Ces indicateurs appelés indicateurs R serviront à fournir des données concernant la variabilité des taux de réponse entre les domaines d'intérêt, afin de déterminer quand la dernière étape devrait commencer.

La dernière étape (adoption échelonnée de la stratégie de collecte dynamique 2) permet de réduire la variation des taux de réponse entre les domaines d'intérêt, en ciblant les cas qui appartiennent aux domaines obtenant des taux de réponse plus faibles, ce qui permet d'améliorer la représentativité de l'échantillon.

4. Stratégie de collecte dynamique pour l'EDTR

La figure 4-1 présente un sommaire de la stratégie de collecte dynamique pour l'EDTR. Plus de détails concernant les étapes de cette stratégie sont fournis dans les sous-sections suivantes.

Figure 4-1
Étapes de la stratégie de collecte dynamique de l'EDTR



Afin de permettre une évaluation de l'efficacité de la stratégie de collecte dynamique, l'échantillon est réparti entre un groupe de contrôle et un groupe de collecte dynamique avant le début de la collecte. Le groupe de contrôle et l'autre groupe doivent avoir à peu près la même taille. Afin d'assurer la représentativité à l'intérieur de chacun de ces groupes, un échantillon systématique sera tiré, une fois l'échantillon stratifié par panel, bureau de collecte des données, groupe de tranches de temps, langue d'interview, résultats du cas au cycle précédent, province de résidence et revenu total du ménage.

4.1 Étape de la planification

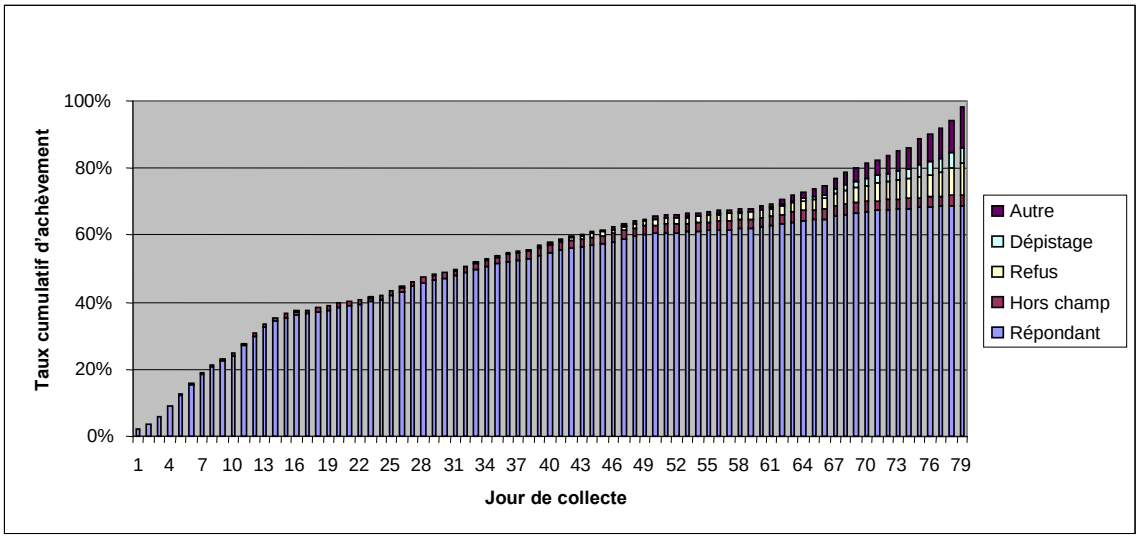
La collecte des données pour le cinquième cycle du panel 5 et du deuxième cycle du panel 6 devrait avoir lieu de janvier à mars 2010. La première étape de la mise en œuvre de la stratégie de collecte dynamique pour cette collecte consiste à examiner les paradonnées de collecte de 2009. Lorsque l'on comprend la façon dont la collecte a progressé auparavant, on peut établir les paramètres de la stratégie de collecte dynamique.

Les paradonnées proviennent du fichier historique des transactions Blaise (BTH). Ce fichier comprend un enregistrement pour chaque accès à un cas au moyen de l'application d'ITAO. Les variables de paradonnées disponibles dans le fichier BTH comprennent par exemple la date, le début et la fin des tentatives d'appel et le résultat de la tentative d'appel, qui sont inscrits sous forme de codes de résultat numériques.

4.1.1 Date de mise en oeuvre

L'examen des taux d'achèvement quotidiens dans le cadre de la collecte de 2009 (voir la figure 4.1.1-1) a révélé deux périodes où ces taux atteignent un plateau, à savoir les jours de collecte 17 à 23 inclusivement et 52 à 58 inclusivement. Ces plateaux correspondent aux semaines pendant lesquelles les bureaux de collecte des données effectuent la collecte de janvier et février pour l'EPA. L'EPA est une enquête importante et vaste; souvent, la collecte des données pour les enquêtes autres que l'EPA est réduite au minimum ou suspendue pendant les trois ou quatre premiers jours de collecte de l'EPA. Compte tenu de la réduction du temps qui sera consacré aux interviews de l'EDTR, la semaine de février de l'EPA représente un moment approprié pour analyser les paradonnées de collecte, calculer les probabilités de réponse, répartir l'échantillon en cours en groupes et entreprendre l'adoption échelonnée du stratégie de collecte dynamique 1. En 2009, la semaine de février de l'EPA a commencé le 22.

Figure 4.1.1-1
EDTR 2009 – Taux cumulatif d'achèvement selon le jour de collecte



4.1.2 Limite intermédiaire du nombre d'appels

Afin de s'assurer que les cas du groupe de collecte dynamique ne se voient pas imposer une limite d'appels trop tôt, une limite intermédiaire du nombre d'appels sera adoptée. Tous les cas qui atteignent cette limite intermédiaire seront mis de côté jusqu'au début de l'adoption échelonnée de la stratégie de collecte dynamique 1. Parmi les considérations importantes pour la détermination de la limite intermédiaire du nombre d'appels figure la façon de supprimer les cas de la charge de travail régulière et ses répercussions sur les activités des bureaux de collecte. Si de trop nombreux cas sont inaccessibles pendant l'étape initiale de la collecte, les activités des bureaux de collecte des données pourraient être interrompues. Toutefois, si on ne réserve pas suffisamment de tentatives d'appel pour les étapes de la stratégie de collecte dynamique, il pourrait être difficile de voir les avantages d'une telle stratégie. Il faut établir un équilibre entre ces deux considérations.

Étant donné que la limite imposée au nombre d'appels de l'EDTR est de 40, on a évalué des limites intermédiaires de 25 et 30. Le tableau 4.1.2-1 présente certaines statistiques sur la façon dont les limites intermédiaires auraient influencé la collecte de 2009, si elles avaient été en place. Par exemple, presque 20 % des tentatives d'appels effectuées entre la date du premier cas assujetti à la limite du nombre d'appels et le début de l'EPA de février correspondaient à des cas qui avaient atteint la limite des 25 tentatives d'appel.

Auparavant, dans le cadre de l'EDTR, on définissait les cas exigeant un effort intensif comme ceux nécessitant 25 tentatives d'appel ou plus. Initialement, on croyait que l'établissement d'une limite intermédiaire de 25 appels, pour correspondre à la définition d'un cas nécessitant un effort soutenu, serait approprié. Toutefois, compte tenu du nombre élevé de cas qui auraient été supprimés de la charge de travail régulière, on a décidé que cela entraînerait plus de perturbations pour les bureaux de collecte des données. Dans ce contexte, la limite intermédiaire a été fixée à 30 tentatives d'appel.

Tableau 4.1.2-1
EDTR de 2009 – Statistiques comparatives des limites intermédiaires de 25 et de 30 appels

	Limite de 25 appels	Limite de 30 appels
Jour de la collecte où le premier cas atteint la limite intermédiaire	Jour 24	Jour 26
Nombre de cas atteignant la limite au début de l'EPA de février	3 443	1 681
Pourcentage de tentatives totales d'appel pour les cas ayant atteint la limite*	19,6%	11,3%
Pourcentage du temps total consacré aux cas ayant atteint la limite*	14,0%	7,9%
Nombre de cas ayant atteint la limite résolus*	256	144

* Entre la date du premier cas ayant atteint la limite et le début de la semaine de l'EPA, le 22 février 2009.

4.1.3 Modèle de propension

Un modèle de propension servira à évaluer la probabilité qu'un ménage soit interviewé. À partir d'un modèle logistique, les cas comportant les probabilités prédites les plus élevées auront la priorité pendant l'adoption échelonnée de la stratégie de collecte dynamique 1. Plusieurs modèles ont été comparés, afin de choisir celui convenant le mieux à cette étape.

Comme dans le cas de la plupart des autres enquêtes, dans l'EDTR, on s'intéresse à la situation de réponse des ménages sélectionnés dans l'échantillon. C'est donc dire que, naturellement, une des variables de réponse prises en compte est le fait ou non que le ménage était un ménage répondant à la date de clôture de l'EPA de février. Dans le cadre de l'EDTR, on se demande aussi si le cas du ménage a été résolu ou non pendant la période de collecte. C'est donc dire qu'une autre variable de réponse prise en compte est la résolution du cas du ménage à la date de clôture de l'EPA.

Les variables explicatives prises en compte dans les modèles proviennent de trois sources différentes. Pour les deux cycles, on disposait des données BTH de 2009 et des données démographiques du fichier de l'échantillon. Étant donné que 2009 était le premier cycle de collecte du panel 6, seul le panel 5 comportait des données BTH de la collecte de 2008. À partir des paradonnées de collecte, des variables catégoriques ont été créées à partir de variables discrètes (p. ex., le nombre de tentatives d'appel, le nombre de contacts), de même que des variables d'indicateur. Une liste complète des variables prises en compte se trouve dans les tableaux 4.1.3-1.

Tableau 4.1.3-1
Variables explicatives prises en compte dans les modèles de propension

Source des données	Liste des variables
Fichier de l'échantillon de 2009	Province de résidence, langue de l'interview précédente, indicateur de séparation du ménage, groupe de tranches de temps
BTH 2009* – de base	Nombre de tentatives d'appel, date de la dernière tentative, nombre de tentatives de dépistage, indicateur de refus, indicateur de dépistage, nombre de contacts
BTH 2009* – détaillé	Nombre de rendez-vous, nombre total de tentatives (régulières et de dépistage combinées), contact pendant une tranche de temps, variables d'indicateur : tentatives la fin de semaine, tentatives la semaine, envoi à un intervieweur principal, indicateurs de résultat : mauvais numéro, absent pendant la durée de l'enquête, problèmes linguistiques, interview interrompue, circonstances spéciales, appel suivant un premier refus, appel suivant une première tentative de dépistage, appel suivant une piste de dépistage, appel suivant un premier renvoi à l'intervieweur principal
BTH 2008**	Nombre de tentatives d'appel, date de la dernière tentative, nombre de tentatives de dépistage, indicateur de refus, indicateur de dépistage, nombre de contacts, situation de réponse, résultat du cas

* Variables dérivées à la date de clôture de l'EPA
** Variables dérivées à la fin de la collecte de 2008, disponibles uniquement pour le panel 5

Afin de simuler le lancement de l'adoption échelonnée de la stratégie de collecte dynamique 1, on a imposé des contraintes aux paradonnées de collecte de 2009. Tout d'abord, afin de simuler la limite intermédiaire de tentatives d'appel, seuls les enregistrements BTH comportant 30 tentatives d'appel ou moins ont été conservés. Par la suite, étant donné que le modèle ne devait pas être exécuté pendant la semaine de l'EPA de février, seuls les enregistrements BTH pour les appels effectués avant la date de clôture de l'EPA, soit le 22 février 2009, ont été conservés.

Le taux de réponse et le taux de cas résolus au moment de la clôture de l'EPA dépendaient du panel, ce qui fait que des modèles logistiques ont été créés séparément pour chaque panel. Chaque modèle a été élaboré à partir de la situation de réponse ou de la résolution comme variable de réponse. Dans le cas des variables explicatives, tous les modèles ont utilisé les variables du fichier de l'échantillon et les variables BTH de 2009 de base, tandis que seulement certains modèles ont utilisé les variables BTH détaillées de 2009. Certains des modèles logistiques du panel 5 sont aussi inclus dans les variables BTH de 2008, outre le fichier de l'échantillon et les variables BTH de base de 2009.

Chaque modèle a été élaboré en SAS, au moyen d'un processus de sélection pas à pas ascendante, avec un maximum de 20 variables explicatives. Les probabilités prédites ont été calculées pour toutes les unités de l'échantillon, tant pour les cas en cours que pour les cas finalisés. Afin d'évaluer le rendement du modèle, seuls les cas en cours ont été conservés et regroupés en déciles selon les probabilités prédites.

Afin de déterminer dans quelle mesure le modèle de propension permettait de déterminer les cas les plus susceptibles d'être complétés, deux critères ont été pris en compte : les taux de réponse des cas comportant des probabilités prédites comprises

dans le groupe du décile supérieur, et la répartition des taux de réponse entre les groupes des déciles supérieur et inférieur. La comparaison des caractéristiques des modèles de cas résolus par rapport aux modèles de situation de réponse pour les groupes de déciles les plus élevés a révélé que les taux de réponse étaient uniformément plus élevés que le taux de cas résolus. La fourchette dans les taux de réponse moyens entre le décile supérieur et le décile inférieur était plus large que la fourchette correspondante des taux moyens de cas résolus. Cela a permis de démontrer que l'utilisation de la situation de réponse était plus appropriée pour déterminer les cas ayant pu être complétés à la fin de la collecte. Par conséquent, on a conservé uniquement les modèles utilisant la situation de réponse.

Dans le cas du panel 5, le modèle le plus simple, qui comprenait uniquement le fichier d'échantillon et les variables BTH de base de 2009, a obtenu les meilleurs résultats. Pour mesurer cela, on a comparé les taux de réponse à la fin de la collecte pour le décile supérieur des cas en cours entre les modèles. Dans le cas du panel 6, le modèle plus complexe a permis de prédire plus précisément les cas en cours les plus susceptibles d'être répondants à la fin de la collecte. Le tableau 4.1.3-2 résume les variables qui ont été intégrées dans chaque modèle. Il convient de souligner que les modèles de propension de l'EDTR utilisent des variables différentes du modèle de propension de l'EME.

Tableau 4.1.3-2
Liste des variables explicatives des modèles finaux

	Liste des variables explicatives du modèle de propension
Panel 5	Date de la dernière tentative, indicateur de séparation du ménage, nombre de tentatives d'appel, indicateur de refus, nombre de tentatives de dépistage, province de résidence, nombre de contacts, groupe de tranches de temps et langue de l'interview précédente
Panel 6	Date de la dernière tentative; appel suivant le premier refus; contact pendant la tranche de temps du soir, nombre de tentatives de dépistage, nombre de tentatives d'appel, circonstances spéciales, contact pendant la tranche de temps de la fin de l'après-midi, problèmes linguistiques, appel suivant le premier renvoi à l'intervieweur principal, appel suivant le premier refus, contact pendant la tranche de temps du matin, nombre de rendez-vous, contact pendant la tranche de temps du début de l'après-midi, absent pour la durée de l'enquête, appel suivant la première tentative de dépistage, nombre de tentatives totales, envoi à l'intervieweur principal, nombre de contacts, interview interrompue, appel suivant une piste de dépistage

Lorsque l'on examine les cas qui étaient en cours à la date de clôture, les taux de réponse à la fin de la collecte pour le panel 5 et le panel 6 se situaient à 25,7 % et 26,7 % respectivement. Par contre, lorsque l'on examine les taux de réponse des cas en cours du décile supérieur, ceux-ci passent à 54,5 % pour le panel 5 et à 72,9 % pour le panel 6.

4.2 Étapes de la collecte

Pendant l'étape initiale de la collecte, celle-ci sera effectuée de la même façon qu'aux cycles précédents. Toutefois, les cas qui font partie du groupe de collecte dynamique qui atteignent la limite intermédiaire d'appels de 30 seront mis de côté. Les travaux seront suspendus pour ces cas jusqu'au début de l'adoption échelonnée de la stratégie de collecte dynamique 1.

Une fois l'adoption échelonnée de la stratégie de collecte dynamique 1 amorcée pour le groupe visé, les cas qui sont passés au groupe de la limite intermédiaire du nombre d'appels pendant l'étape de la collecte initiale seront alors accessibles. Après l'exécution du modèle de propension, les cas en cours réguliers seront répartis en trois groupes : ménages qui n'ont pas encore été contactés; ménages qui ont été contactés et, selon le modèle de propension, qui sont les plus susceptibles de terminer l'interview (dans le décile supérieur); et tous les autres ménages qui ont été contactés. Par ailleurs, de nombreux ménages doivent faire l'objet d'un dépistage pendant la collecte. Les cas de dépistage dont les probabilités prédites de réponse se situent dans les 10 % supérieurs de toutes les probabilités prédites parmi les cas de dépistage en cours seront identifiés et intégrés à un autre groupe. Au cours de l'adoption échelonnée de la stratégie de collecte dynamique 1, tous les cas de non-dépistage en cours liés à la stratégie de collecte dynamique qui atteignent la limite de 30 tentatives d'appel seront transférés au groupe de la limite intermédiaire du nombre d'appels. On souhaite que la répartition de l'échantillon en cours entre des groupes facilement identifiables permette aux gestionnaires de la collecte de doter plus efficacement leurs postes d'intervieweurs.

Enfin, lorsqu'il sera temps de passer à l'adoption échelonnée de la stratégie de collecte dynamique 2, les cas recevront un ordre de priorité à partir des groupes prioritaires qui ont été utilisés dans les cycles précédents de collecte. Laflamme et Karaganis (2010) décrivent en détail les critères qui serviront à déterminer le moment approprié pour passer d'une étape de la collecte à la suivante. Il n'y aura pas de critère unique signalant le début ou la fin d'une étape de la stratégie de collecte dynamique, mais plutôt un ensemble d'indicateurs qui comprennent notamment : le taux de réponse, les niveaux de productivité et le pourcentage du budget dépensé.

5. Conclusions et travaux à venir

Du fait de la diminution des taux de réponse et de l'augmentation des coûts des enquêtes, l'adoption d'une stratégie de collecte dynamique pour la collecte des données de l'EDTR de 2010 vise à augmenter la représentativité de l'échantillon, tout en maintenant les coûts de la collecte constants. L'utilisation d'un modèle de propension pour déterminer les cas qui comportent une probabilité plus grande d'être interviewés devrait réduire la tendance à la baisse des taux de réponse qui a été notée dans le cadre de l'EDTR au fil des ans. Par ailleurs, il est à souhaiter que la dotation plus efficace des postes d'intervieweur, le ciblage des efforts et la réduction des coûts des enquêtes soient utiles à cette fin. L'utilisation d'un indicateur R pour évaluer la représentativité de l'échantillon contribuera à maintenir et même à améliorer la qualité des données.

La prochaine collecte pour l'EDTR commence en janvier 2010. À ce moment-là, on pourra profiter de l'expérience d'une stratégie de collecte dynamique de l'EME, dont la collecte prend fin en novembre 2009. Toutes les leçons tirées de l'EME peuvent aider à améliorer la stratégie de collecte dynamique proposée dans la présente communication.

Après la fin de la collecte de l'EDTR de 2010, il sera nécessaire d'évaluer l'efficacité de la stratégie de collecte dynamique, grâce à l'analyse des paradonnées et à la comparaison des résultats entre le groupe de contrôle et le groupe de collecte dynamique. Par ailleurs, la stratégie de collecte dynamique devra être prise en compte au moment l'ajustement de la non-réponse, afin de tenir compte de tous les biais qui pourraient découler de l'utilisation de ce type d'approche adaptative pour la collecte des données.

Bibliographie

Dufour, S. (2009). A Review of Response Rates for Statistics Canada Surveys, unpublished report, Ottawa, Canada: Statistics Canada.

Groves, R. M. et Heeringa, S. G. (2006). Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs, *Journal of the American Statistical Association*, 169, pp. 439-457.

Laflamme, F.et Karaganis, M. (2010). Development and Implementation of Responsive Design for CATI Surveys at Statistics Canada, to be presented at the European Quality Conference, Helsinki, Finland.

Laflamme, F. et Mohl, D. (2007). Research and Responsive Design Options for Survey Data Collection, paper presented at the Joint Statistical Meeting, Salt Lake City, USA.

Schouten, B., Cobben, F. et Bethlehem, J. (2009), Indicateurs de la représentativité de la réponse aux enquêtes, *Techniques d'enquête*, 35, pp. 107-121.

SESSION 11 – PONDÉRATION ET ESTIMATION

L'ajustement de la pondération à l'aide de scores de propension dans un plan à base de sondage double

C. Boudreau, M.E. Thompson et M. Iraniparast [1]

Résumé

La International Tobacco Control Policy (ITC) Netherlands Survey est une enquête longitudinale permanente menée auprès de plus de 2 200 fumeurs depuis 2008. Elle vise à examiner les comportements d'usage du tabac et les répercussions des politiques antitabac au niveau national. Il s'agit de la première enquête menée dans le cadre du projet ITC (qui comprend maintenant des enquêtes réalisées dans plus de 18 pays) au moyen d'une base de sondage double en ligne et à composition aléatoire. La base de sondage en ligne est une base de données sur plus de 200 000 répondants hollandais qui ont convenu de participer régulièrement à des études de recherche menées par la firme de sondage internationale TNS NIPO. Malgré tous les soins pris par TNS NIPO pour assurer la représentation exacte de la population hollandaise dans sa base de données, certains groupes demeurent sous-représentés et d'autres, surreprésentés. Même si les poids d'échantillonnage ont été construits pour en tenir compte dans toute la mesure du possible, un certain biais de sélection demeure.

Dans cette communication, nous présentons une méthode qui utilise l'échantillon sélectionné par composition aléatoire, qui est plus petit, de l'ITC Netherlands Survey de concert avec un score de propension pour corriger les poids de l'échantillon en ligne, qui est plus grand. La méthode consiste à ajuster un modèle de score de propension dans lequel la réponse correspond au mode d'enquête, puis à procéder à un ajustement poststratification des poids de l'échantillon en ligne en utilisant les scores de propension estimés. Il est facile de modifier la méthode dans le cas de valeurs manquantes. Nous l'illustrons au moyen de données de la première vague de l'ITC Netherlands Survey, mais la méthode peut également être utilisée dans le cas d'autres enquêtes à deux bases de sondage dont l'une n'est pas entachée d'un biais.

[1] C. Boudreau (cboudreau@math.uwaterloo.ca), M.E. Thompson et M. Iraniparast, University of Waterloo, Canada

L'estimation longitudinale dans l'enquête européenne sur le revenu et les conditions de vie

Ralf Münnich et Stefan Zins[1]

Résumé

La pauvreté et la cohésion sociale au sein de l'Union européenne sont mesurées au moyen des indicateurs de Laeken. La plupart de ces indicateurs sont des mesures transversales. L'Enquête européenne sur le revenu et les conditions de vie (EU-SILC, de l'anglais *European Survey of Income and Living Conditions*) est une enquête par panel rotatif qui a été lancée afin de mesurer l'évolution des indicateurs de Laeken au cours du temps.

Le présent article décrit l'estimation de la variation de certaines mesures de la pauvreté et de l'inégalité, ainsi que l'évaluation de l'exactitude de leurs estimations au moyen de méthodes de linéarisation. Les mesures étudiées comprennent le coefficient de GINI, le rapport interquintile de revenu et le taux de risque de pauvreté. En outre, une attention particulière est accordée au plan de sondage à panel rotatif. L'étude comprend les résultats d'une étude par simulation de Monte Carlo.

Mots-clés : indicateurs de la pauvreté, mesures de l'inégalité, estimation de la variance pour l'estimation du changement, panel rotatif.

1. Introduction

En mars 2000, le Conseil européen de Lisbonne a demandé aux États membres de l'Union européenne de prendre des mesures pour améliorer la cohésion sociale et éliminer la pauvreté de là à 2010. Afin de bien mesurer la pauvreté et la cohésion sociale, il a été convenu qu'un ensemble d'indicateurs appelés indicateurs de Laeken seraient mesurés et publiés chaque année par tous les États membres. Une description détaillée de ces indicateurs peut être consultée dans Eurostat (2008).

Toutefois, si la simple publication des indicateurs de Laeken par chaque État membre de l'Union européenne est nécessaire, elle n'est pas suffisante pour évaluer les progrès réalisés en regard des objectifs de l'Union européenne. Il faudrait pour cela mesurer l'évolution des indicateurs de Laeken au cours du temps, c'est-à-dire comparer les estimations obtenues pour deux années consécutives. Cette mesure fournirait un outil précieux pour évaluer l'efficacité des politiques mises en œuvre par les États membres de l'Union européenne en vue d'éradiquer la pauvreté et d'améliorer la cohésion sociale (voir Eurostat, 2005, section 2).

L'Enquête européenne sur le revenu et les conditions de vie (EU-SILC) est une enquête par panel rotatif qui a été lancée afin de pouvoir mesurer adéquatement les indicateurs de Laeken transversalement, ainsi qu'au cours du temps. Aux termes du Code de bonnes pratiques de la statistique européenne, les estimations des indicateurs de Laeken doivent être accompagnées de mesures d'exactitude appropriées qui requièrent généralement l'application de méthodes d'estimation de la variance.

Le but du présent article est de décrire des méthodes d'estimation de la variance de l'estimation du changement de certains indicateurs de Laeken au cours du temps. Des différences ainsi que des rapports sont examinés en tant que mesures du changement. L'étude met principalement l'accent sur le plan d'échantillonnage rotatif de l'EU-SILC, qui comprend quatre panels rotatifs dont un est remplacé chaque année. Ainsi, pour mesurer le changement d'une année à la suivante, on dispose d'un échantillon bidimensionnel : trois des panels se chevauchent chaque année et l'un d'eux est indépendant aux deux points dans le temps.

À la section suivante, nous donnons un aperçu des indicateurs de Laeken choisis pour l'étude et de leur estimation. Ensuite, nous présentons les méthodes de linéarisation pour l'estimation de la variance, d'abord pour des estimations transversales à la section 3, puis pour des estimations du changement à la section 4. Enfin, à la section 5, nous donnons un aperçu des résultats d'une étude par simulation de Monte Carlo basée sur les données de l'EU-SILC afin d'évaluer les estimateurs de variance proposés dans un cadre s'approchant des conditions réelles.

[1]Ralf Münnich et Stefan Zins, University of Trier, Faculty IV, Economic and Social Statistics Department. D-54296 Trier, Germany (Muennich@uni-trier.de; Zins@uni-trier.de).

2. Les indicateurs de Laeken et leur estimation

Les indicateurs de Laeken étudiés ici sont le coefficient de Gini, le taux de risque de pauvreté et le rapport interquintile de revenu. La portée des indicateurs de Laeken est considérablement plus étendue que celle des statistiques que nous avons choisies, mais ces dernières permettent de couvrir la méthodologie d'intérêt qu'il convient d'appliquer pour la plupart des autres mesures de la pauvreté financière et de l'inégalité du revenu. Pour l'ensemble complet d'indicateurs de Laeken, consulter Eurostat (2008) ou Osier (2009).

Le *taux de risque de pauvreté* (TRP) est la proportion de personnes dont le revenu disponible équivalent est inférieur à ce qu'il est convenu d'appeler le seuil de pauvreté, fixé à 60 % du revenu médian national équivalent. Donc, le TRP peut s'écrire sous la forme

$$\text{TRP} = \frac{1}{N}\sum_{j\in U}(y_j < 0{,}6 \cdot Q_{0.5}), \qquad [1]$$

où N est le nombre total de personnes dans la population finie $U=\{1, ..., N\}$ et y_j est le revenu disponible équivalent de la j^e personne dans U. $Q_{0.5}$ est la médiane de la fonction de répartition $F(y)$, avec $Q_{0.5} = \inf_{y\in\square}(F(y) > 0{,}5)$ et $F(y) = \frac{1}{N}\cdot\sum_{j\in U} 1(y_j \le y)$. $1(.)$ désigne la fonction indicatrice qui est égale à 1 si la condition est satisfaite et à 0 autrement. Le TRP est estimé par

$$\hat{\text{TRP}} = \hat{F}(0{,}6 \cdot \hat{Q}_{0.5}), \qquad [2]$$

où $\hat{F}$ est une estimation de la fonction de répartition obtenue en utilisant les poids de sondage et $\hat{Q}_{0.5}$ est une estimation de la médiane. Pour obtenir plus de détails, consulter Osier (2009). Le *rapport interquintile de revenu S80/S20* (R8020) est le rapport entre le revenu total des 20 % de la population du pays ayant le revenu le plus élevé (quintile supérieur) et le revenu total des 20 % ayant le revenu le plus faible (quintile inférieur).

$$\text{R8020} = \frac{\sum_{j\in U} y_j \cdot 1(y_j > Q_{0.8})}{\sum_{j\in U} y_i \cdot 1(y_i \le Q_{0.2})}$$

où $Q_{0.2}$ et $Q_{0.8}$ désignent le quintile inférieur et le quintile supérieur, respectivement. L'indicateur R8020 peut être estimé comme il suit :

$$\hat{\text{R8020}} = \frac{\sum_{j\in S} w_j \cdot (y_j - y_j \cdot 1(y_j \le \hat{Q}_{0,8})) / \sum_{j\in S} w_j \cdot (1 - 0{,}8)}{\sum_{j\in S} w_j \cdot y_j \cdot 1(y_j \le \hat{Q}_{0,2}) / \sum_{j\in S} w_j \cdot 0{,}2},$$

où w_j est le poids de sondage correspondant à la j^e personne dans un échantillon S. Le *coefficient de Gini* (GINI), qui est donné pour une population finie sous la forme

$$\text{GINI} = \frac{1}{N}\cdot\sum_{j\in U}(2\cdot F(y_j) - 1)\cdot\frac{y_j}{\mu},$$

est estimé par

$$\text{GINI} = \frac{1}{\hat{N}} \cdot \hat{\mu} \cdot \sum_{i \in S} w_i \cdot (2 \cdot \underbrace{\frac{1}{\hat{N}} \cdot \sum_{j \in S} w_i \cdot 1(y_j \le y_i)}_{\hat{F}(y_i)} - 1) \cdot y_i .$$

Ici, la population totale est estimée par $\hat{N} = \sum_{j \in S} w_j$ en utilisant les poids de sondage w_j.

3. Estimation de la variance par des méthodes de linéarisation

L'estimation de la variance de statistiques non linéaires peut être approchée de deux façons, par les méthodes de rééchantillonnage et par les méthodes de linéarisation. Un aperçu des méthodes de rééchantillonnage pour les mesures d'inégalité est donné dans Kovačević et Yung (1997). Puisque le présent article est axé sur les méthodes de linéarisation, pour un aperçu des méthodes de rééchantillonnage, nous renvoyons le lecteur à Davison et Sardy (2007) ou à Münnich (2007) et aux références mentionnées par ces auteurs.

Les méthodes de linéarisation englobent la linéarisation de Taylor et la linéarisation de Woodruff (voir Woodruff, 1971, Andersson et Nordberg, 1994) ou des méthodes plus générales, telles que l'application d'équations d'estimation (voir Kovačević et Binder, 1997) et de fonctions d'influence (voir Deville, 1999). Demnati et Rao (2004) offrent une méthode de rechange pour calculer les estimations de variance par linéarisation de Taylor.

Deville (1999) a présenté un cadre général fondé sur le concept des fonctions d'influence qui avait d'abord été introduit dans le domaine des statistiques robustes par Hampel, Ronchetti, Rousseeuw et Stahel, (1986). Deville (1999) emploie une classe générale de statistiques non linéaires pour une population finie U fondée sur le concept d'une mesure fonctionnelle. Un paramètre de population d'intérêt θ peut alors être décrit comme une fonctionnelle $T(M)$ d'une mesure discrète finie M qui répartit un poids de 1 entre toutes les unités $k \in U$. Si θ doit être estimé d'après un échantillon S, soit $\hat{M}$ l'estimateur de M, qui désigne la mesure qui attribue un poids w_k à n'importe quelle observation $k \in S$ et un poids nul à toutes les observations $k \notin S$, où w_k peut être un poids de sondage de n'importe quel type associé à la k^e unité dans l'échantillon. L'estimateur de $T(M)$ s'obtient en remplaçant M par $\hat{M}$ et nous pouvons donc écrire $\hat{\theta} = T(\hat{M})$. La variance de cet estimateur peut être approximée par celle d'une statistique linéaire

$$V(T(\hat{M})) = V\left(\sum_{k \in S} w_k \cdot u_k\right), \qquad [3]$$

où u_k est la valeur linéarisée pour l'unité k donnée par $u_k = IT(M; k)$ où IT correspond à la *fonction d'influence de T* dans M. La fonction d'influence d'une fonctionnelle T est donnée par

$$IT(M;k) = \lim_{t \to \infty} \frac{1}{t} \cdot \left[T\left(M + t\delta_k\right) - T\left(M\right)\right], \qquad [4]$$

où δ_k est la mesure de Dirac de l'observation k désignant la masse unitaire supposée à l'observation k.

Si l'expression [4] doit être estimée d'après un échantillon observé S, la mesure d'échantillon $\hat{M}$ doit remplacer M dans [4]. Les règles de construction d'une fonction d'influence sont essentiellement celles utilisées pour dériver les fonctions en calcul différentiel. Pour un ensemble de règles, voir Deville (1999) et pour les hypothèses qui doivent être satisfaites afin d'appliquer cette approximation, voir Deville (1999) ou directement pour les indicateurs de Laeken, voir Osier (2009).

La façon de calculer les valeurs linéarisées du TRP est décrite ci-après. Premièrement, nous avons besoin de la fonction d'influence pour $F(y)$, désignée par $IF(y; k)$. La fonction d'influence pour $F(y)$ à l'observation k peut être dérivée comme il suit

$$IF(y;k) = \frac{1}{N} \cdot \left[1(y_k \le y) - F(y)\right] \qquad [5]$$

(voir Osier, 2009, comme un résultat tiré de Deville, 1999). Partant de l'équation [2], la fonction d'influence du TRP peut s'écrire sous la forme (Deville, 1999) :

$$ITRP(k) = \frac{1}{N} IF(0,6 \cdot Q_{0,5}; k) + f(0,6 \cdot Q_{0,5}) \cdot 0,6 \cdot IQ_{0,5}, \quad [6]$$

où *f(y)* est la dérivée d'une fonction lissée de *F(y)* et $IQ_{0.5}$ est la fonction d'influence de la médiane, qui est donnée par

$$IQ_{0,5} = -\frac{1}{f(Q_{0,5}) \cdot N} \cdot \left[1(y_k \le Q_{0,5}) - 0,5\right]. \quad [7]$$

Enfin, en insérant [5] et [7] dans [6], nous pouvons écrire la fonction d'influence du TRP à l'observation *k* sous la forme

$$ITRP(k) = \frac{1}{N}\left[1\left(y_k \le 0,6 \cdot Q_{0,5}\right) - F(0,6 \cdot Q_{0,5})\right] - \frac{1}{N}\left(\frac{f(0,6 \cdot Q_{0,5})}{f(Q_{0,5})}\right) \cdot \left[1(y_k \le Q_{0,5}) - 0,5\right].$$

Puisque *F* est une fonction non lisse, sa dérivée est toujours nulle ou non définie au sens classique. Afin de lisser *F*, nous pouvons appliquer des estimateurs de densité par noyau (voir Berger et Skinner, 2003). Une attention particulière pourrait être accordée à la fonction de lissage, car la distribution sous-jacente du revenu est parfois très asymétrique. Une autre solution consiste à appliquer des fonctions splines.

Des valeurs linéarisées u_k des indicateurs GINI et R8020 peuvent également être calculées en utilisant le concept de fonctions d'influence. Pour savoir précisément comment établir ces fonctions pour le GINI, voir Deville (1999) ou Osier (2009), et pour l'approche fondée sur des équations d'estimation, voir Kovačević et Binder (1997). Pour le rapport interquintile de revenu R8020, les valeurs d'influence peuvent être tirées de Hulliger et Münnich (2007) ainsi que de Langel et Tillé (2009), ou Osier (2009). Une comparaison empirique des méthodes de linéarisation et de rééchantillonnage pour les mesures transversales est donnée dans Münnich et coll. (2010).

4. Estimation de la variance pour la différence et le rapport des estimateurs des indicateurs de Laeken

La différence $\theta_{\Delta:0,1} = \theta_1 - \theta_0$ et le rapport $\theta_{R:0,1} = \theta_1 / \theta_0$ de deux valeurs d'indicateur aux temps t=0, 1 peuvent être estimés simplement. Les estimations de la variance sont calculées conformément aux règles données à la section 3 en utilisant les valeurs d'influence des mesures transversales. Cela donne

$$V\left(\hat{\theta}_{R:0,1}\right) = V\left(\frac{\hat{\theta}_1}{\hat{\theta}_0}\right) \cong V\left(\frac{1}{\hat{\theta}_{t_0}}\left[u_1 + \hat{\theta}_{R:0,1} \cdot u_0\right]\right) \cong \frac{1}{\hat{\theta}_0^2}\left[V\left(\hat{\theta}_1\right) + \hat{\theta}^2_{R:0,1} \cdot V\left(\hat{\theta}_0\right) - 2 \cdot \hat{\theta}^2_{R:0,1} \cdot Cov\left(u_0, u_1\right)\right], \quad [8]$$

où u_0 et u_1 sont les valeurs linéarisées correspondant aux valeurs d'indicateur θ_0 et θ_1 respectivement.

La variance de la différence est calculée de la même façon avec $V\left(\hat{\theta}_{\Delta:0,1}\right) = V\left(\hat{\theta}_1\right) + V\left(\hat{\theta}_0\right) - 2 \cdot Cov\left(u_0, u_1\right)$. Dans l'équation [8], les variances sont simplement les estimations transversales selon [3] et ne nécessitent aucune autre étude. La partie cruciale est l'estimation de la covariance temporelle des valeurs d'influence. Cette covariance doit être estimée en tenant compte du plan d'échantillonnage rotatif. La justification théorique est donnée en utilisant les observations bivariées à deux points dans le temps où un panel possède des observations indépendantes. Une approche plus générale comportant deux échantillons est exposée dans Goga, Deville et Ruiz-Gazen (2009), qui ont d'abord appliqué la théorie à l'estimation des variations de l'indice GINI.

5. L'étude de Monte Carlo

Le principal objectif de l'étude de Monte Carlo est d'évaluer empiriquement l'exactitude des estimateurs de la variance proposés pour les estimations du changement dans le contexte de l'EU-SILC. Le volet transversal de l'enquête comporte quatre panels, sélectionnés chacun durant quatre années successives afin d'ajouter une dimension longitudinale à l'ensemble de données. Après qu'un panel a participé à l'enquête pendant quatre ans, il est éliminé et remplacé par un panel nouvellement sélectionné. Cette procédure produit un ensemble de données transversales composé de quatre panels par année, tandis que les données longitudinales couvrant deux années consécutives n'englobent que trois panels.

Pour l'étude par simulation, nous avons choisi l'ensemble de données longitudinales de l'EU-SILC de 2006 comme population finie. Nous nous sommes servis du revenu disponible équivalent annuel des ménages observés pour 2005 et 2006 pour calculer les indicateurs de Laeken. Les données comprennent les observations provenant de tous les États membres de l'Union européenne, sauf la Bulgarie, la Roumanie et Malte, ainsi que l'Islande et la Norvège.

Deux distributions du revenu ont été prises en considération dans l'étude, à savoir la distribution du revenu réel de l'EU-SILC en excluant les revenus négatifs (désignée EUS-INC) et une distribution tronquée de l'EU-INC, dans laquelle les revenus supérieurs à 150 000 euros ont été supprimés (désignés EUS-INC-ST). La taille N de la population de EUS-INC est de 132 616 et celle de EUS-INC-ST, de 132 443. Plusieurs distributions du revenu ont été utilisées pour tenir compte de la non-robustesse de l'indicateur R8020 aux fortes inégalités vers la queue supérieure d'une distribution. Les valeurs de revenu négatives ont dû être exclues de l'ensemble de données parce que le coefficient de Gini est indéfini pour les revenus négatifs.

Afin de mettre en œuvre un plan d'échantillonnage rotatif, la population a été divisée aléatoirement en panels (quarts) rotatifs. Le plan d'échantillonnage peut maintenant être décrit comme il suit :

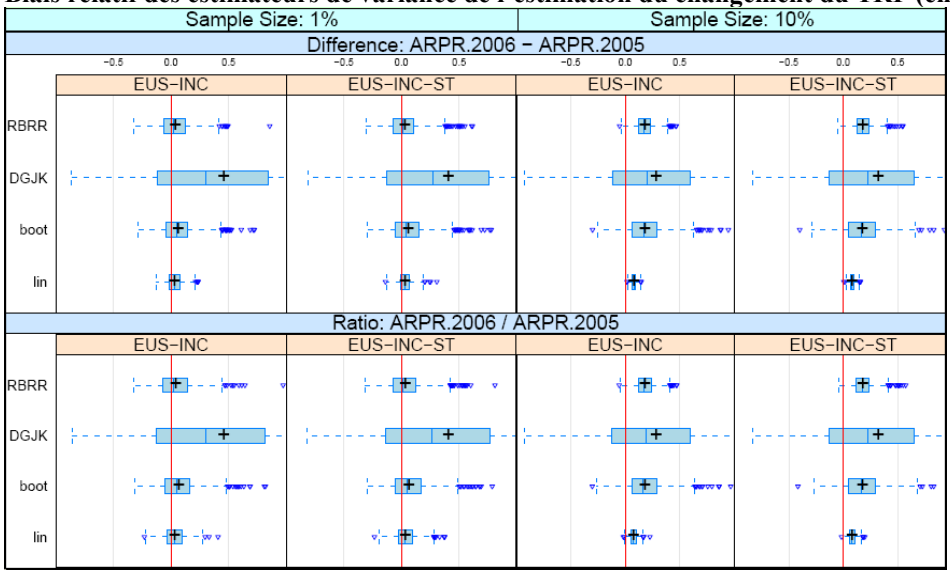
1. un premier échantillon de taille n est tiré indépendamment de tous les panels;
2. un deuxième échantillon de taille $n/4$ est tiré d'un des panels seulement.

Les échantillons des étapes 1 et 2 ont été sélectionnés selon un plan d'échantillonnage stratifié avec répartition proportionnelle, où les pays ont été utilisés comme variables de stratification.

Les variances qui figurent dans l'équation [8] ont été calculées en utilisant un estimateur de variance classique pour l'échantillonnage aléatoire stratifié (voir Andersson et Nordberg, 1994, p. 403). La covariance est également estimée en utilisant son estimateur de variance classique pour l'échantillonnage aléatoire stratifié et en l'appliquant aux valeurs d'influence aux deux périodes en omettant les valeurs linéarisées provenant du panel indépendant.

Pour certains cas particuliers de différents types de chevauchement des échantillons, Qualité et Tillé (2008) donnent un aperçu des estimateurs de variance des estimations du changement basés sur des estimateurs de type Horvitz-Thompson. Voir également Tam (1984) pour l'estimation de la variance pour des échantillons chevauchants.

Figure 5-1
Biais relatif des estimateurs de variance de l'estimation du changement du TRP (en anglais, ARPR)



La figure 5-1 donne les boîtes à moustaches pour le biais relatif de plusieurs estimateurs de variance des estimations du changement pour le TRP. La figure est subdivisée en quatre segments, chacun ayant trait à un objectif différent. La moitié supérieure de la figure a trait à l'estimateur de la différence et la moitié inférieure, à l'estimateur du ratio. La partie gauche de la figure donne les résultats pour une fraction d'échantillonnage de 1 % et la partie de droite, pour une fraction de 10 %. Chaque segment contient les boîtes à moustaches pour les deux distributions du revenu étudiées EUS-INC (à gauche) et EUS-INC-ST (à droite). Quatre estimateurs de variance différents ont été examinés, trois fondés sur des méthodes de rééchantillonnage et *lin*, l'estimateur de variance linéarisé donné en [8]. Les trois méthodes de rééchantillonnage utilisées sont *RBRR*, un estimateur par réplique répétée équilibrée (Kovačević et Yung, 1997), *DGJK*, un estimateur jackknife avec suppression d'un groupe (Kott, 2001) et *boot*, un estimateur bootstrap Monte Carlo non paramétrique. L'estimateur de la variance linéarisée est celui qui donne les meilleurs résultats pour tous les objectifs. Parmi les méthodes de rééchantillonnage, *RBRR* et *boot* produisent un biais moins important que *DGJK*, qui donne lieu à un biais positif important. Les estimations de variance pour le TRP semblent être entachées d'un faible biais. Cependant, la variabilité diminue quand la taille de l'échantillon augmente. Les résultats obtenus pour les indicateurs GINI et R8020 sont similaires, mais le biais est nettement plus faible et les estimateurs semblent plus stables. Cependant, dans ces cas, la combinaison de la petite taille d'échantillon et de la présence de valeurs extrêmes dans la distribution du revenu produit des distributions des estimations de variance fortement asymétriques, comme prévu. Les taux de couverture des intervalles de confiance sont bons pour EUS-INC-ST, mais peuvent poser certains problèmes pour EUS-INC dans le cas des indicateurs GINI et R8020.

7. Résumé et perspective

Dans le cadre du présent article, nous avons montré que les méthodes de linéarisation produisent des estimations de variance exactes pour les statistiques fortement non linéaires des indicateurs de Laeken. Certains problèmes peuvent se poser quand les distributions du revenu contiennent des valeurs aberrantes. Le TRP est moins sensible à l'asymétrie de la distribution, mais a tendance à présenter un biais plus important en ce qui concerne l'estimation de la variance. Pour les indicateurs GINI et R8020, les estimations semblent être relativement non robustes aux valeurs aberrantes, mais très efficaces dans les cas lissés.

Dans l'avenir, les travaux de recherche pourraient porter sur les méthodes de calage afin de stabiliser les résultats et d'obtenir des estimations plus exactes pour tous les indicateurs de Laeken. En ce qui concerne le biais observé dans le cas du TRP, le lissage de la fonction de répartition peut être amélioré en utilisant la régression spline au lieu d'estimations de la densité par noyau.

Remerciements

L'étude a été menée dans le cadre du projet de recherche AMELI (Advanced Methodology for European Laeken Indicators; http://ameli.surveystatistics.net) qui est financé par la Commission européenne aux termes de son septième programme-cadre.

Bibliographie

Andersson, C. et Nordberg, L. (1994). A Method for Variance Estimation of Non-Linear Functions of Totals in Surveys - Theory and Software Implementation, *Journal of Official Statistics*, 10, pp. 395-405.

Berger, Y. et Skinner, C. (2003). Variance estimation for a low income proportion, *Journal of Applied Statistics*, 52, pp. 457-468.

Davison, A. C. et Sardy, S. (2007). Resampling Variance Estimation in Surveys with Missing Data, *Journal of Official Statistics*, 23, pp. 371-386.

Demnati, A. et Rao, J. N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête, *Techniques d'enquêt*e, 30, pp. 17-27.

Deville, J. C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus, *Techniques d'enquête*, 25, pp. 219-230.

Eurostat (2005). Measuring progress towards a more sustainable Europe, European Commission, Eurostat, DOC KS-68-05-551-EN.

Eurostat (2008). Algorithms to compute Overarching Indicators based on EU-SILC and adopted under the Open Method of Coordination (OMC), Working group Statistics on Living Conditions, European Commission, Eurostat, DOC LC-ILC/11/08/EN – REV.

Goga, C. et Deville, J. C., and Ruiz-Gazen, A. (2009). Use of functionals in linearization and composite estimation with application to two-sample survey data, *Biometrika*, 96, pp. 691-709.

Hulliger et Münnich (2007). Variance Estimation for Complex Surveys in the Presence of Outliers, *Proceedings of the Survey Research Methods Section*, *American Statistical Association*, pp. 3153-3161.

Kott, P. S. (2001). The Delete-a-Group Jackknife, *Journal of Official Statistics*, 17, pp. 521-526.

Kovačević, M. C. et Binder, D. A. (1997). Variance Estimatin for Measures of Income Inequality and Polarization – The Estimating Equations Approach, *Journal of Official Statistics*, 13, pp. 41–58.

Kovačević, M. C. et Yung, W. (1997). Estimation de la variance des mesures de l'inégalité et de la polarisation du revenu : étude empirique, Techniques d'enquête, 23, pp. 47-59.

Langel, M. et Tillé, Y. (2009). Statistical Inference for the Quintile Share Ratio, unpublished report, Neuchâtel, Swiss, University of Neuchâtel.

Münnich, R. (2008). Varianzschätzung in komplexen Erhebungen, *Austrian Journal of Statistics*, 37, pp. 319-334.

Münnich, R., Zins, S. et Kolb, J.-P. (2010). Accuracy of Poverty and Inequality Indicators for the German EU-SILC survey, *en cours de soumission*.

Osier, G. (2009). Variance estimation for Complex Indicators of poverty and inequality using linearization Techniques, *Survey Research Methods*, 3, pp. 167-195.

Qualité, L. et Tillé, Y. (2008). Estimation de la précision d'évolutions dans les enquêtes répétées, application à l'enquête suisse sur la valeur ajoutée, *Techniques d'enquêtes*, 34, pp. 193-201.

Tam, S. M. (1984). On Covariances from Overlapping Samples, *The American Statistician*, 38, pp. 288-289.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate, *Journal of the American Statistical Association*, 66, pp. 411-414.

La pondération et l'estimation de la variance dans l'enquête par panel à base de sondage double menée auprès des ménages allemands (PASS)

Hans Kiesl [1]

Résumé

L'Institut allemand pour la recherche sur l'emploi vient d'achever la deuxième vague d'une nouvelle enquête par panel annuelle portant sur les ménages à faible revenu, qui est conçue comme enquête à deux bases de sondage. La première base est un registre des ménages qui touchent actuellement des prestations de chômage; la deuxième base se compose d'un registre des adresses de l'ensemble de la population. Au début, 6 000 ménages ont été sélectionnés dans chaque base de sondage (avec échantillonnage PPT des codes postaux à la première étape), d'où la forte variance des poids de sondage entre les deux échantillons. Les poids pour les ménages et pour les personnes sont fournis pour les deux sous-échantillons ainsi que pour l'échantillon combiné, ce qui donne six ensembles différents de poids transversaux pour chaque vague.

Dans cette communication, nous décrivons les défis que présentent la pondération et l'estimation de la variance pour les deux premières vagues de notre enquête, y compris l'élagage des poids extrêmes, la coordination des poids des ménages et de ceux des personnes, les stratégies de correction de la non-réponse pour les ménages et les personnes, une comparaison de différentes méthodes de partage des poids pouvant être utilisées pour tenir compte de la variation de la composition des ménages au fil du temps et l'utilisation d'une pondération convexe afin d'intégrer un échantillon de naissances de deuxième vague tiré de la base de sondage plus petite (c.-à-d. de nouveaux ménages ayant besoin de prestations).

[1] Hans Kiesl, Institute for Employment Research (IAB), Allemagne (Hans.Kiesl@iab.de)

SESSION 12 – TENIR COMPTE DES DONNÉES MANQUANTES DANS L'ANALYSE DE DONNÉES D'ENQUÊTES LONGITUDINALES

Modélisation et analyse des durées d'après des données d'enquêtes longitudinales

J.F. Lawless et D. Mariaca Hajducek[1]

Résumé

La durée des périodes vécues dans certains états présente un intérêt dans le contexte des processus du cycle de vie pour lesquels les enquêtes longitudinales fournissent des données. Malgré les progrès réalisés ces 20 dernières années en analyse des données sur les événements de la vie, de nombreuses difficultés persistent lorsque l'on se sert de données provenant d'enquêtes longitudinales. Le présent article traite de la modélisation et de l'analyse des durées de période fondées sur ce genre de données. On y examine les difficultés dues aux données manquantes, à la perte de vue des membres du panel au suivi, à la complexité de la population étudiée et à celle du plan de sondage. Certaines méthodes d'analyse sont décrites et illustrées brièvement au moyen de données sur les périodes de chômage provenant de l'Enquête sur la dynamique du travail et du revenu réalisée par Statistique Canada.

Mots clés : pondération par l'inverse de la probabilité de censure, données manquantes, pseudo-vraisemblance, poids de sondage, analyse de survie.

1. Introduction

Les enquêtes longitudinales permettent de recueillir des données sur les études, l'emploi, la santé et d'autres aspects du cycle de vie d'un individu. Ce genre de données rendent possible l'analyse biographique, dans le cadre de laquelle sont étudiés et modélisés des processus faisant intervenir des événements et d'autres variables associés aux individus au cours de leur vie. Un cadre utilisable à cette fin est celui de la modélisation à états multiples, dans lequel les individus évoluent entre certains états au cours d'une certaine période. Les modèles de ce genre sont, par exemple, utilisés pour étudier l'emploi, les partenariats conjugaux et la maternité. Ces modèles sont spécifiés et analysés en tenant compte de la durée des périodes passées dans des états particuliers et des transitions d'un état à un autre. Le but de la présente étude est d'examiner des moyens d'analyser les durées de ce type de périodes en se basant sur des données d'enquêtes longitudinales. L'analyse des durées et l'analyse biographique sont des domaines d'étude bien établis qui bénéficient d'un appui logiciel considérable (p. ex., Andersen et coll., 1993, Kalbfleisch et Prentice, 2002, Cook et Lawless, 2007). Cependant, les données recueillies au moyen d'enquêtes longitudinales sont assorties de problèmes que les méthodes classiques ne permettent pas de traiter adéquatement. Notre objectif est de discuter de ce genre de problèmes et d'envisager des moyens de les résoudre.

Les populations étudiées dans les enquêtes longitudinales ont tendance à être hétérogènes, et les membres d'un panel sont habituellement sélectionnés conformément à un plan de sondage complexe. Les données sur les membres du panel ne sont recueillies que par intermittence (p. ex., tous les ans ou tous les deux ans), au cours d'une période spécifiée. Dans le cas de l'Enquête sur la dynamique du travail et du revenu (EDTR) de Statistique Canada, dont nous discuterons plus loin, les individus sont interviewés annuellement au cours d'une période de six ans. Il est fréquent que des données manquent pour des variables particulières (non-réponse partielle). En outre, certains membres du panel peuvent être perdus de vue au suivi avant la dernière interview, ou peuvent manquer une interview puis réapparaître à une interview ultérieure. Comme nous en discutons plus loin, la perte de vue au suivi peut-être reliée aux résultats antérieurs ou covariables d'un individu.

Le présent article porte sur la modélisation et l'analyse des durées de période. Les objectifs de ce genre d'analyse peuvent être répartis en deux grandes catégories. La première est descriptive, l'objectif étant d'estimer les caractéristiques d'une distribution dans la population à partir de laquelle est sélectionné le panel. Par exemple, nous pourrions vouloir estimer la distribution marginale de la durée des périodes de chômage chez les résidents de l'Ontario en 2006, ou nous pourrions vouloir relier les périodes de chômage à des covariables associées à l'individu ou à la conjoncture économique. La deuxième catégorie englobe l'analyse explicative des processus du cycle de vie, ou processus biographique des individus. Ici, nous nous intéressons à l'association entre les périodes récurrentes de chômage et d'emploi, et à leurs relations avec les covariables. Les données longitudinales permettent d'estimer plus facilement les distributions des durées des périodes que les données rétrospectives

[1] J.F Lawless, Dept. of Statistics and Actuarial Science, University of Waterloo, Waterloo ON Canada N2L 3G1 (jlawless@uwaterloo.ca); D. Mariaca Hajducek, Dept. of Statistics and Actuarial Science, University of Waterloo, Waterloo ON Canada N2L 3G1 (hajducek@gmail.com)

obtenues durant une seule interview. Néanmoins, les problèmes de données manquantes ou d'erreurs de mesure peuvent rendre les enquêtes longitudinales nettement moins utiles qu'une étude de cohorte longitudinale dans laquelle les individus sont suivis de près.

2. Analyse des durées pour les données d'enquêtes longitudinales

Nous considérons des conditions où les individus occupent des séries d'états au cours du temps. Par exemple, un modèle d'emploi très simple comprend les états Occupé, Chômeur et Inactif. Nous nous intéressons ici à la durée des périodes passées dans des états particuliers. Une période dans un état se termine par une transition à un autre état, et les données types sur un individu observé au cours d'une période comprennent une série d'états E_1, E_2, E_3,... et les durées Y_1, Y_2, Y_3,… des périodes passées dans chacun de ces étapes. Dans certains cas, nous pourrions chercher à modéliser séparément les transitions d'un état vers chacun des autres états dans lesquels un individu peut passer, mais pour simplifier, nous considérons les durées sans tenir compte de l'état auquel passe l'individu.

Supposons que les membres d'un panel doivent être observés aux temps *0,1,2,…, M* sur une période *[0, M]*; pour simplifier, nous supposons que le temps est exprimé en années et que les individus sont vus au début de chaque année t. À chaque temps $t = 1,\ldots,6$, on recueille l'information au sujet de la période d'un an précédente *(t-1,t]*; au temps $t = 0$, on obtient les données repères. Posons que les états occupés par l'individu i au cours de la période *[0, M]* sont (en ordre) $E_{i_1}, E_{i_2}, \ldots E_{im_i}$. Soit u_{ij} et v_{ij} les temps d'entrée et de sortie dans l'état E_{ij}; notons qu'en général, $u_{i1} < 0$ et $v_{im_i} > M$. La durée de la période dans l'état E_{ij} est $Y_{ij} = v_{ij} - u_{ij}$. Si un individu est perdu de vue au suivi avant le temps M, certaines périodes qu'il a vécues peuvent être inobservées ou leur durée peut-être censurée. La durée de la dernière période est habituellement censurée, même si l'individu est suivi jusqu'au temps M; autrement dit, nous savons seulement que $Y_{im_i} > M - u_{im_i}$.

L'analyse d'une seule durée tout à fait définie d'après des données d'enquête est bien établie. Considérons, en particulier, une durée Y_i et un vecteur de covariables x_i pour les individus $i = 1,\ldots,n$ qui sont sélectionnés selon un plan d'échantillonnage complexe. L'estimation sous des modèles de régression à risques proportionnels de Cox a été abordée par Binder (1992), qui a considéré des fonctions d'estimation pondérées par les poids de sondage pour estimer les coefficients de régression et a fourni des estimations de la variance fondée sur le plan de sondage pour les estimateurs des coefficients de régression. Lin (2000), ainsi que Boudreau et Lawless (2006) ont examiné l'estimation de paramètres de population finie et de superpopulation, et fourni les estimations de la variance pour les estimateurs des paramètres de régression et des fonctions de risque de base cumulatif. Ces méthodes sont implémentées dans des progiciels tels que SAS, SUDAAN et R/SPLUS. Lawless et Boudreau (2003) ont passé en revue des méthodes similaires pour les modèles de régression paramétriques.

Les méthodes mentionnées au paragraphe qui précède permettent aussi de traiter les durées multiples Y_{ij} avec les vecteurs de covariables x_{ij} pour les individus $i = 1,\ldots,n$. Kovacevic et Roberts (2007) traitent les périodes multiples en utilisant l'approche répandue de l'« indépendance de travail » basée sur le modèle de Cox et des estimations de variance robustes (p. ex., Lin, 1994). Cependant, cette approche a été conçue pour des durées multiples observées parallèlement et est sujette à un biais dans le cas d'une série de durées corrélées (p. ex., voir Cook et Lawless, 2007, section 4.4). Dans ces conditions, nous devrions normalement inclure l'information sur les durées et les états précédents examinés dans un modèle pour une durée particulière, afin de tenir compte des dépendances entre les durées pour un individu donné. Malgré ce fondement solide de l'inférence au sujet des durées d'après des données d'enquête, la méthodologie existante présente plusieurs lacunes dont nous allons discuter maintenant.

3. Certaines lacunes dans la méthodologie

Par souci de simplicité, considérons des durées de période Y pour un état particulier $E = e$, et désignons la fonction de répartition conditionnelle de Y sachant les covariables x par $(y|x)$. La spécification de la population pour Y doit se faire avec soin, afin que les Y puissent être comparés au fil du temps et entre individus, et que les paramètres des modèles associés à Y soient interprétables. Nous discutons de quatre domaines dans lesquels existent des lacunes et des défis méthodologiques. Ceux-ci résultent de l'incomplétude des données obtenues pour les individus dans une enquête longitudinale et de la nature du plan de sondage suivant lequel sont sélectionnés les individus.

3.1 Perte de vue au suivi dépendante

Dans les enquêtes longitudinales, la perte de vue au suivi (PVS) a lieu quand un individu n'est pas « vu » (c.-à-d. que l'information le concernant n'est pas recueillie) après une interview donnée $C < M$; nous considérons C comme la période de la PVS. Nous supposons ici que tous les individus sont interviewés au moment de la sélection (interview 0) et à la première

interview subséquente (interview 1), mais qu'ils peuvent ensuite être perdus de vue au suivi à n'importe laquelle des interviews *2,…,M-1*. Les individus qui ne sont pas perdus de vue au suivi sont vus pour la dernière fois à l'interview *M*. Il est également possible que les données sur un individu ne soient pas disponibles pour une interview particulière, mais qu'elles le deviennent de nouveau à l'interview suivante; nous examinons ce cas à la section 3.2.

Une durée Y_i possède un temps de censure notionnel C_i et les méthodes classiques d'analyse de survie, y compris celles mentionnées à la section 2, requièrent que Y_i et C_i soient indépendants, étant donné les covariables x_i utilisées dans l'analyse. Cette contrainte est souvent violée parce que a) les covariables ayant une incidence sur Y_i ainsi que C_i sont exclues du modèle ou b) Y_i et C_i sont tous deux liés à la biographie du sujet avant le temps u_i auquel a débuté la période, qui n'est pas incluse dans le modèle. À partir d'ici, nous supposerons, pour la clarté de l'exposé, que les individus sont vus une fois par an, en janvier. Supposons qu'une période débute chez une personne en 2006 après l'interview et qu'elle se termine en 2007. Trois scénarios sont alors possibles : i) si la personne a été perdue de vue en janvier 2007 (c.-à-d. qu'elle a été vue en janvier 2006 mais non en janvier 2007), la période de la PVS est effectivement janvier 2006 et nous n'en serions pas au courant; ii) si la personne a été vue en janvier 2007 mais non en janvier 2008, la période de la PVS est effectivement janvier 2007 et Y_i est censurée, le temps de la censure correspondant à la date de la dernière interview moins u_i; iii) si la personne est vue en janvier 2007 et en janvier 2008, alors la durée Y_i est observée. Il s'ensuit que, si Y_i et la probabilité que la personne soit PVS à l'interview de 2006, de 2007 ou de 2008 dépendent toutes deux des covariables qui ne figurent pas dans le modèle de régression, la censure est dépendante et peut causer un biais important dans les estimations de la distribution des durées.

L'occurrence d'une période, et son temps de début u_i, peut également dépendre de certaines covariables. Si une telle covariable est également liée à la PVS et à la durée et qu'elle n'est pas incluse dans le modèle de durée, un biais peut de nouveau se produire.

3.2 Autres types de données manquantes

Quand se produit une période qui débute au temps u_i et se termine au temps v_i, il se peut que le temps de début, le temps de fin ou les covariables ne soient pas observés, à cause de la PVS ou de la non-réponse (provisoire). Par exemple, chez les résidents de l'Ontario du panel de l'EDTR de 1999, les dates de début sont inconnues pour 29 % des périodes de chômage et, en outre, les valeurs d'une ou de plusieurs covariables manquent pour 18 % des périodes dont la date de début est connue. De plus, environ 30 % des personnes ont été perdues de vue au suivi avant l'interview de la sixième année (janvier 2005). Les raisons pour lesquelles des données manquent comprennent la non-réponse totale, due au fait qu'une personne manque une interview annuelle, mais réapparaît au cours d'une année ultérieure, la non-réponse partielle, due au refus ou à l'incapacité de répondre à certaines questions (p. ex., dans le cas d'un répondant par procuration) et à la perte de vue au suivi, discutée à la section précédente.

3.3 Estimation de la variance

Des méthodes en vue de tenir compte du plan de sondage par pondération ont été élaborées pour l'inférence fondée sur le plan dans les articles mentionnés à la section 2. Ces méthodes tiennent également compte des effets de grappe dans la population. Pour traiter l'intégration du calage ou des ajustements pour la PVS, nous devons permettre l'utilisation de poids dont les composantes sont estimées. Les méthodes bootstrap sont souvent utilisées, quoique leur validité n'ait pas été rigoureusement établie dans de nombreux cas. Les méthodes existantes ne permettent pas de résoudre de façon satisfaisante la question de la pondération dans l'analyse des durées ni celle de l'association entre les durées. Nous examinons ces questions à la section suivante où nous considérons divers moyens d'y faire face.

4. Méthodes pour traite la PVS dépendante et les données manquantes

4.1 Perte de vue au suivi dépendante

Supposons que les individus sont vus aux temps $t= 0,1,…,M$, comme nous l'avons spécifié à la section 2. Au temps t ($t = 1,…,M$), nous recueillons l'information concernant la période $(t-1, t]$. Les données de référence recueillies au temps $t = 0$ ne comprennent habituellement pas de renseignements détaillés sur les séjours dans les états survenus avant cela. Donc, les données complètes pour un individu comprennent les données de référence ainsi que les données couvrant les M années subséquentes. C_i désigne la dernière année durant laquelle des données sont recueillies pour l'individu i; nous supposons que des données sont recueillies pour tous les individus à l'année *1*.

Deux points sont importants dans le développement qui suit : une période peut occuper (des parties de) plus d'une année et la probabilité qu'une personne soit perdue de vue au suivi à l'année t peut dépendre des événements biographiques antérieurs. Les

méthodes que nous décrivons requièrent que les données manquent au hasard dans la terminologie de Little et Rubin (2002) et, dans ces conditions, nous supposons par conséquent que la perte de vue au suivi à l'année *t* peut dépendre seulement de l'information disponible jusqu'à l'année *t-1*. Dans certaines conditions, il est probable que cette perte de vue au suivi dépende en outre des occurrences au cours de l'année *t*, mais il n'existe aucun moyen d'en tenir compte en se basant sur les données observées sans émettre des hypothèses non vérifiables. Nous en discutons plus en détail à la section 6.

Supposons que nous nous intéressons à la durée des périodes passées dans un état particulier et représentons par $D_i(t)$ l'information sur ces durées pour l'individu *i* à l'année *t*. Supposons que nous nous intéressons à une famille de modèles reliant $D_i(t)$ à un ensemble de covariables. Nous notons cela comme il suit

$$P_\theta(D_i(t) \mid Z_i^0(t)) = \Pr(D_i(t) \mid Z_i^0(t); \theta), \tag{1}$$

où $Z_i^0(t)$ comprend l'information au sujet de $\{D_i(1), D_i(2), \ldots, D_i(t-1)\}$ et des covariables externes $x_i(t)$ d'intérêt. Le paramètre θ (qui peut être multidimensionnel, voir même contenir des composantes fonctionnelles) spécifie la famille (1) et notre objectif est d'estimer θ et d'évaluer l'adéquation des modèles ajustés.

Pour l'inférence en superpopulation, on suppose que, pour un paramètre θ_0, l'espérance du score de log-vraisemblance

$$U_{it}(\theta) = \frac{\partial \log P_\theta(D_i(t) \mid Z_i^0(t))}{\partial \theta}, \tag{2}$$

conditionnellement à $Z_i^0(t)$, est nulle pour un membre aléatoire de la population. Cependant, quand un ensemble d'individus $i=1,2,\ldots,n$ provenant du panel d'une enquête longitudinale vivent les périodes en question, cette hypothèse n'est généralement plus vraie. En outre, un membre du panel peut être perdu de vue au suivi avant l'année *t*. Nous spécifions maintenant un cadre pour traiter cet aspect.

Soit $R_{it} = I(C_i \geq t)$ indiquant que l'individu *i* $(i=1,\ldots,n)$ du panel est vu à l'année *t*; nous supposons qu'aucune donnée ne manque dans $D_i(t)$ ni dans $Z_i^0(t)$ dans ce cas. Soit $Z_i(t)$ un vecteur de covariables contenant les variables du plan de sondage Z_i^D et d'autres facteurs temporellement variables possibles $Z_i^C(t)$, tels que a) la sélection dans le panel est indépendante de $D_i(t)$ et $Z_i^0(t)$ sachant Z_i^D et b) R_{it} est indépendant de $D_i(t)$ et de $Z_i^0(t)$ sachant $Z_i^C(t)$. Nous supposons que $Z_i^C(t)$ dépend seulement de l'information observée jusqu'à l'année *t-1*, de même que $Z_i^0(t)$. Nous notons π_i =Pr (individu *i* sélectionné dans le panel $\mid Z_i^D$) et $p_{it}(\alpha)$=Pr(R_{it} =1$\mid Z_i^C(t); \alpha$); soulignons que, tandis que la probabilité d'inclusion π_i est une fonction connue de Z_i^D, la probabilité de la PVS $p_{it}(\alpha)$ est basée sur une famille de modèles indicés par le paramètre α. Nous fondons maintenant l'estimation de θ sur l'estimation de la fonction

$$U^W(\theta) = \sum_{i=1}^{n} \sum_{t=1}^{M} \frac{R_{it}}{\pi_i p_{it}(\hat{\alpha})} U_{it}(\theta), \tag{3}$$

où $\hat{\alpha}$ est une estimation de α.

La fonction d'estimation $U^W(\theta) = 0$ produit un estimateur convergent $\hat{\theta}$ de θ, à condition que le modèle de PVS $p_{it}(\alpha)$ soit correct et que $\hat{\alpha}$ soit un estimateur convergent de α (Robins et coll., 1995). La pondération par l'« inverse de la probabilité de censure » (IPC) $p_{it}(\hat{\alpha})^{-1}$ a été utilisée souvent dans le contexte des enquêtes longitudinales (p. ex., Miller et coll., 2001), mais pas pour l'analyse des durées. Les poids de sondage π_i^{-1} ont été utilisés seuls pour les données sur les durées provenant d'enquêtes (p. ex., Binder 1992, Lin 2000).

Afin d'utiliser (3), nous avons besoin d'un modèle pour le décrochage (arrêt de participation à l'enquête). Une approche très généralement applicable consiste à modéliser la distribution de R_{it} sachant $R_{i,t-1}=1$ par régression logistique sous la forme

$$\text{logit } \lambda_{it}(\alpha) = \text{logit } \Pr(R_{it} = 1 \mid R_{i,t-1} = 1, Z_i^C(t); \alpha) = \alpha' Z_i^C(t). \tag{4}$$

Alors, $p_{it} = \lambda_{i1}\lambda_{i2}\cdots\lambda_{it}$. Les modèles (4) peuvent être ajustés pour $t=1,\ldots,M$ d'après les données observées, en supposant une fois qu'une personne manque à l'année *t* (c.-à-d., $R_{it}=0$), elle n'est plus revue par après. On peut pour cela désigner une personne comme étant perdue de vue au suivi (PVS) la première année où aucune information la concernant n'est obtenue.

L'utilisation de (3) requiert que les données sur les durées de période soient partitionnées en fractions $D_i(t)$. Par exemple, une période qui débute au temps u_i à l'année 1 et se termine au temps v_i à l'année 2 est partitionnée lorsque $D_i(1)$=*{période commencée à* $t=u_i$ *et encore en cours à* $t=1$*}* et $D_i(2)$=*{période ayant la durée courante* $1-u_i$ *à* $t=1$*, et terminée à* $t=v_i$*}*. Dans (3), des poids de PVS différents p_{i1}^{-1} et p_{i2}^{-1} sont appliqués aux deux fractions. Les logiciels standards pour le modèle de Cox,

tels que **coxph** dans R/SPLUS et **phreg** dans SAS, permettent cette partition, mais de nombreux programmes d'analyse paramétrique de la survie ne le permettent pas. Lawless et Mariaca Hajducek (2010) fournissent des renseignements détaillés sur l'application des fonctions d'estimation (3); et une illustration est présentée à la section 5.

4.2 Autres types de données manquantes

En général, les dates de début de période manquantes ou les valeurs manquantes des covariables posent des problèmes. Dans les enquêtes longitudinales, les profils de données manquantes peuvent varier considérablement d'un individu à l'autre. Même dans les cas où des méthodes relativement simples sont applicables, peu de logiciels standards sont disponibles. Nous discutons brièvement ici de certains de ces problèmes.

Les valeurs manquantes peuvent être traitées par des méthodes d'imputation et par des méthodes de pseudo-vraisemblance ou de vraisemblance pondérées. Le remplacement d'une valeur manquante par une valeur imputée unique est peu attrayant quand les données manquantes sont nombreuses. L'imputation multiple (Rubin, 1996) est meilleure, mais nécessite des hypothèses au sujet des processus menant aux valeurs manquantes. Une troisième option, surtout dans le cas de covariables catégoriques, consiste à ajouter une catégorie appelée « valeur manquante » à la covariable. Cette option n'est pas particulièrement satisfaisante quand l'objectif est d'évaluer la relation d'une covariable avec une variable temporelle de durée, mais peut être utile quand on souhaite faire une prévision. L'utilisation de fonctions d'estimation pondérées par l'inverse de la probabilité (Robins et coll., 1995) ou du maximum de vraisemblance (Little et Rubin 2002, Cook et Lawless 2007) en se basant sur des modèles pour les valeurs manquantes permet, en principe, de traiter les cas où la probabilité qu'une valeur manque est associée à des covariables ou à des réponses observées, mais parce que ces modèles de données manquantes sont nécessaires, les méthodes sont principalement utiles quand les valeurs des mêmes variables manquent chez les divers individus. Il est possible d'étendre les méthodes de Robin et coll. (1995), de Lawless et coll. (1999) et celles mentionnées dans les références connexes en vue de les appliquer à des données d'enquête, mais, jusqu'à présent, cela n'a pas été fait dans des conditions comportant l'analyse des durées. Une question importante est celle de savoir ce qu'il faut faire quand les profils de données manquantes ne sont pas suffisamment simples pour utiliser ces méthodes. Une certaine forme d'imputation aléatoire semble être l'approche la plus pratique, mais le grand défi consiste à établir comment exécuter l'imputation et comment obtenir les estimations de la variance; voir, par exemple, Kenward et Carpenter (2007).

4.3 Estimation de la variance

Nous discuterons brièvement de la manière dont peut être abordée l'estimation de la variance dans les cas de perte de vue au suivi qui sont traités par les méthodes décrites à la section 4.1. Des méthodes comparables peuvent être élaborées pour les méthodes de pondération par l'inverse de la probabilité ou celles du maximum de vraisemblance mentionnées à la section précédente.

L'estimation de la variance fondée sur des fonctions d'estimation telles que (3) doit tenir compte de l'estimation des paramètres α dans les poids, ainsi que de l'association entre les individus et des périodes récurrentes dans les individus. Dans le cas de modèles entièrement paramétriques, l'approche peut être celle de la théorie asymptotique pour les fonctions d'estimation. Par exemple, nous considérons pour (3) et (4) deux ensembles de fonctions d'estimation

$$U^W(\theta,\alpha) = \sum_{i=1}^{n}\sum_{t=1}^{M} \frac{R_{it}}{\pi_i p_{it}(\alpha)} U_{it}(\theta) \tag{5}$$

$$U^R(\alpha) = \sum_{i=1}^{n}\sum_{t=1}^{M} R_{i,t-1}\left\{\frac{R_{it}-\lambda_{it}(\alpha)}{\lambda_{it}(\alpha)(1-\lambda_{it}(\alpha))}\right\}\frac{\partial\lambda_{it}(\alpha)}{\partial\alpha}, \tag{6}$$

où (6) est une équation du score de vraisemblance basée sur (4). Nous pouvons obtenir une matrice de covariance estimée pour $\hat{\theta}$ en partant de la matrice de covariance asymptotique pour $\psi = (\theta,\alpha)$. Celle-ci a la forme standard $AB(A^{-1})'$ avec $A = E(-\partial U/\partial\psi')$ et $B = Var(U)$, où $U = U(\psi) = (U^W(\psi)', U^R(\psi)')'$ avec les vecteurs $U, \psi, \theta, \alpha,$ etc. écrits sous forme de vecteurs colonnes. La matrice de covariance B devrait reconnaître les effets de grappe dans l'échantillon et dans la population; voir Lawless et Mariaca Hajducek (2010). Les modèles semi-paramétriques pour lesquels θ possède des composantes fonctionnelles sont plus compliqués. Lawless et Mariaca Hajducek (2010) considèrent les modèles de régression de Cox.

5. Exemple

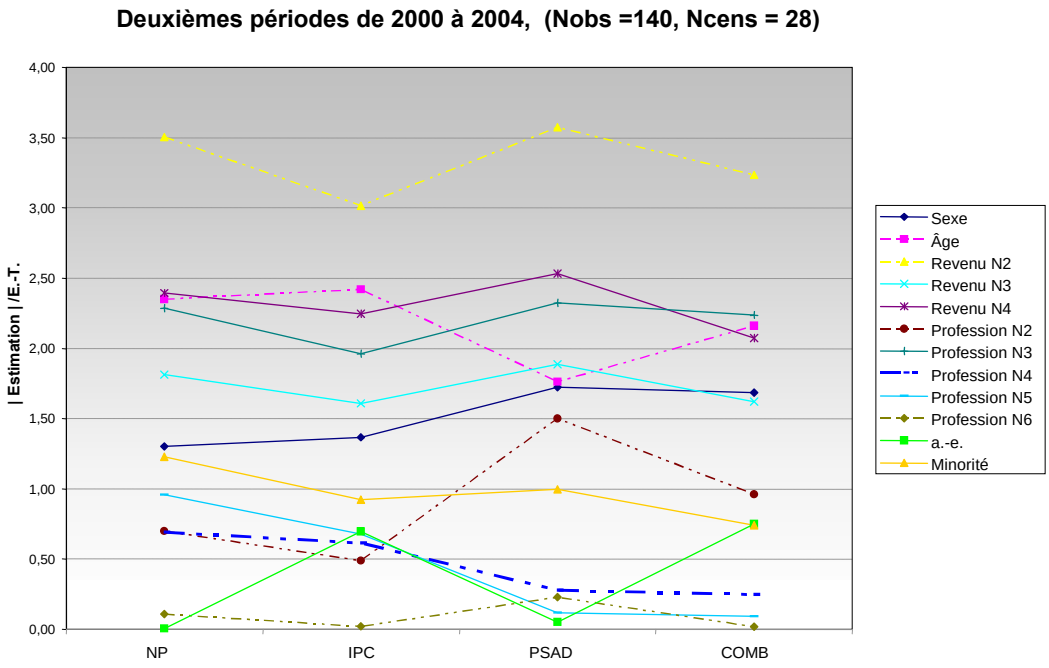
Nous illustrons brièvement les aspects dont nous avons discuté plus haut en considérant la durée d'une période de chômage pour le panel de l'EDTR de 1999. L'information sur la situation d'activité des années précédentes est recueillie annuellement; autrement dit, les données pour les années 1999 à 2004 sont recueillies, respectivement, au début de chaque année suivante. Une période de chômage est définie comme une période durant laquelle une personne ne travaille pas à cause d'une mise à pied permanente et est à la recherche de travail.

Notre exemple porte sur les durées des périodes de chômage qui ont débuté durant la période de 2000 à 2004, pour les personnes qui étaient âgées de 16 à 63 ans en 1999 et qui résidaient en Ontario cette année-là. Les périodes pour lesquelles la date de début manquait ont été éliminées. Une importante remarque au sujet de l'exemple est que nous supposons que les dates de début des périodes manquent entièrement au hasard. Dans de futurs travaux, nous prévoyons étudier les méthodes mentionnées à la section 4.2, ce qui nécessitera seulement l'hypothèse que les dates de début de période manquent au hasard. Nous avons modélisé les durées en utilisant des modèles à risques proportionnels de Cox et la fonction d'estimation $U^W(\theta) = 0$ dans (3). Nous considérons jusqu'à trois périodes par individu. Des 931 périodes ainsi définies, 236 présentaient une certaine forme de valeurs manquantes (date de début ou valeurs des covariables manquantes). Parmi les 695 périodes restantes utilisées pour la modélisation, 441 étaient des premières périodes, 168 étaient des deuxièmes et 86 étaient des troisièmes périodes. Environ 17 % étaient censurées. Les covariables catégoriques examinées ici sont le sexe (féminin/masculin), le revenu (faible, moyen-inférieur, moyen, élevé), la profession (services, fabrication, secteur primaire, construction, services administratifs, services professionnels), l'assurance-emploi (non/oui) et l'appartenance à un groupe minoritaire (oui/non). Dans chaque cas, le premier niveau a été utilisé comme catégorie de référence. L'âge a été traité comme une variable continue. L'année de début d'une période et, pour les deuxième et troisième périodes, la durée moyenne des périodes de chômage antérieures ont également été incluses.

La modélisation de la perte de vue au suivi pour chaque année portait sur tous les individus. Les valeurs manquantes pour les covariables catégoriques ont été représentées comme une catégorie distincte. Nous avons évalué 12 variables de l'EDTR et certaines de leurs interactions en vue de leur inclusion dans les modèles de décrochage (4) pour chaque année. En outre, nous avons inclus une variable indicatrice précisant si la personne était sans emploi à l'interview de l'année précédente. Les variables significativement associées au décrochage à tous les âges étaient l'âge, le niveau de scolarité, l'état matrimonial et la situation d'emploi. Les effets avaient le sens attendu; par exemple, les personnes mariées ou vivant en union de fait étaient moins susceptibles d'être perdues de vue au suivi, et le niveau de scolarité d'une personne était inversement relié à la probabilité de perte de vue au suivi.

Nous présentons ici les résultats pour quatre options de pondération pour l'estimation : a) non-pondération (« NP »), b) poids estimés d'après la modélisation du décrochage (« IPC »), c) poids de sondage longitudinaux de l'année de début de la période (« PSAD ») et d) poids donnés par l'inverse des IPC et des probabilités de sélection combinées (« COMB »). La figure qui suit donne les valeurs de |estimation|/e.-t.(estimation) pour chaque option de pondération dans le cas des deuxièmes périodes de chômage. La covariable donnant la durée de la première période de chômage n'est pas incluse. Elle est corrélée aux autres covariables et son effet n'était pas significatif. Les erreurs-types ont été calculées en tenant compte de la mise en grappes, mais non du fait que les poids IPC sont estimés. Nous nous attendons à ce que les variances estimées par une méthode qui tient compte du caractère aléatoire des poids IPC soient un peu plus faibles (Robin et coll., 1995), ce qui aurait tendance à rendre les valeurs p des covariables un peu plus faibles.

La figure montre que les estimations standardisées des coefficients de régression et le caractère significatif des variables sont à peu près les mêmes pour les quatre options de pondération. Ce résultat pourrait être dû au fait que la plupart des périodes de chômage ont une durée de moins d'un an ou deux, et que l'effet de la perte de vue au suivi n'est donc pas très important. En outre, l'estimation correcte de la variance tenant compte de la mise en grappes est utilisée même pour les estimations non pondérées. Pour les variables les plus significatives, la direction des effets est plausible. Par exemple, un niveau de revenu plus élevé l'année précédente est associé à des périodes de chômage plus courtes et un âge plus avancé est associé à des périodes plus longues.



6. Conclusion

Le rôle des données d'enquête longitudinale dans l'analyse explicative détaillée des processus tels que l'emploi semble être limité, à moins que puissent être résolus les problèmes que posent les données manquantes et l'exactitude des mesures. Nous n'avons pas discuté des problèmes d'erreurs de mesure ici, mais nous soulignons que pour certaines variables, ils peuvent être importants. Les dates attribuées aux événements passés au moment des interviews annuelles sont particulièrement sujettes à erreur quand un enregistrement concret de l'événement n'est pas disponible. Les effets dits de lisière (Cotton et Gilles 1998, Callegaro 2008) dans lesquels les dates enregistrées pour les événements sont biaisées vers les dates de la collecte des données sont bien décrits, mais les méthodes permettant de traiter ce problème de manière satisfaisante semblent limitées. De toute façon, il existe de nombreuses raisons d'entreprendre l'étude et l'élaboration de méthodes pour résoudre ces problèmes.

En guise de mot de la fin, soulignons que la contrainte voulant que les données manquent au hasard (MAR pour *missing at random*) pour pouvoir appliquer les méthodes discutées ici sera presque certainement violée dans toute enquête dans laquelle les membres du panel sont vus ou contactés à des intervalles de temps très grands. Dans ce cas, le rôle des données auxiliaires provenant de sources administratives, ou peut-être du dépistage de certaines personnes qui sont perdues de vue au suivi, est très important. Les méthodes en vue d'intégrer ce genre de données dans l'analyse des durées méritent qu'on leur accorde de l'attention. En l'absence de ce genre de données, une analyse de sensibilité peut être exécutée pour évaluer l'effet de la perte de vue au suivi non-MAR ou d'autres types de données manquantes (p. ex., Scharfstein et Robins, 2002).

Remerciements

Nous remercions Georgia Roberts, Christian Boudreau et Mary Thompson de leurs commentaires et suggestions, et Georgia Roberts de son aide durant le stage du deuxième auteur à Statistique Canada à l'automne de 2008. Nous sommes reconnaissants à Georgia Roberts et à Pat Newcombe–Welch du Southwestern Ontario Research Data Centre de l'Université de Waterloo de leur aide concernant les données de l'EDTR. Nous tenons aussi à remercier MITACS et le Programme national sur les structures de données complexes (CRDNG) de leur soutien.

Bibliographie

Andersen, P.K., Borgan, O., Gill, R.D. et Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.

Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data, *Biometrika, 79*, pp. 139-147.

Boudreau, C. et Lawless, J.F. (2006). Survival analysis based on the proportional hazards model and survey data, *The Canadian Journal of Statistics, 34*, pp. 203-216.

Callegaro, M. (2008). Seam effects in longitudinal surveys, *Journal of Official Statistics, 24*, pp. 387-409.

Cook, R.J. et Lawless, J.F. (2007). *The Statistical Analysis of Recurrent Events*, Springer, New York.

Cotton, C. et Giles, P. (1998). The seam effect in the Survey of Labour and Income Dynamics, *SLID working paper No.75F0002,* Statistics Canada, Ottawa, ON.

Kalbfleisch, J.D. et Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition, Wiley, New York.

Kenward, M.C. et Carpenter, J. (2007). Multiple imputation: current perspective, *Statistical Methods in Medical Research, 16*, pp.199-218.

Kovacevic, M. S. et Roberts, G.R. (2007). Modélisation des durées de périodes multiples à partir de données d'enquête longitudinale, *Techniques d'enquête, 33,* pp. 15-25.

Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd edition, Hoboken, NJ.

Lawless, J.F. et Boudreau, C. (2003). Modélisation et analyse des données sur la durée provenant d'enquêtes longitudinales*, Recueil du Symposium 2002 de Statistique Canada : Modélisation des données d'enquête pour la recherche sociale et économique*.

Lawless, J.F. et Mariaca Hajducek, D. (2010). Analysis of durations in longitudinal studies with intermittent ascertainment of information, Unpublished manuscript.

Lawless, J.F., Kalbfleisch, J.D. et Wild, C.J. (1999). Semiparametric methods for response-selective and missing data problems in regression, *Journal of the Royal Statistical Society B, 61*, pp. 413-438.

Lin, D.Y. (1994). Cox regression analysis of multivariate failure time data: the marginal approach, *Statistics in Medicine, 13,* pp. 2233-2247.

Lin, D.Y. (2000), On fitting Cox's proportional hazards models to survey data, *Biometrika, 87*, pp. 37-47.

Little L.J.A. et Rubin D.B. (2002). *Statistical Analysis of Missing Data*, 2nd edition, Wiley, New York.

Miller, M.E., Ten Have, T.R., Reboussin, B.A., Lohman, K.K. et Rejeski, W.J. (2001). A marginal model for analyzing discrete outcomes from longitudinal surveys with outcomes subject to multiple cause nonresponse, *Journal of the American Statistical Association*, *96,* pp.844-857.

Robins, J.M, Rotnitzky, A. et Zhao, L.P (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association, 90*, pp. 106-121.

Rubin, D.B. (1996). Multiple imputation after 18+ years, *Journal of the American Statistical Association, 91*, pp. 473-489.

Scharfstein, D.O. et Robins, J.M. (2002). Estimation of the failure time distribution in the presence of informative censoring, *Biometrika, 89,* pp. 617-634.

L'analyse d'enquêtes longitudinales comportant des réponses manquantes

Changbao Wu and Ivan Carrillo Garcia[1]

Résumé

Au cours des dernières années, les enquêtes longitudinales sont devenues un important outil de collecte de données dans les études démographiques axées principalement sur l'évolution des variations de population au niveau individuel. La méthode de l'équation d'estimation généralisée (EEG) est l'outil d'inférence statistique le plus utilisé dans les études longitudinales. Toutefois, la vaste majorité des études publiées sur la méthode EEG abordent cette méthode à d'autres fins que des enquêtes, sans tenir compte des enjeux liés aux plans d'échantillonnage complexes.

Nous proposons des méthodes pour analyser des enquêtes longitudinales lorsque la variable réponse comporte des valeurs manquantes. Nos méthodes reposent sur le cadre de l'EEG et sont fortement axées sur le recours à la méthode EEG lorsque les données manquantes sont traitées au moyen de l'imputation. Nous expliquons d'abord pourquoi, puis nous montrons comment on peut intégrer les poids de sondage à la méthode dite pseudo-EEG dans un cadre de randomisation conjointe où les réponses manquantes soit sont traitées par repondération ou par imputation. La cohérence des estimateurs EEG des coefficients de régression produits par cette méthode est établie dans certaines conditions de régularité. Nous calculons des estimateurs par linéarisation de la variance en supposant que la fraction d'échantillonnage de la population finie est minime ou négligeable; cette hypothèse est souvent retenue dans les vastes enquêtes auprès de la population. Nous examinons le rendement des estimateurs proposés à l'égard d'un échantillon fini au moyen d'une étude par simulation. Les résultats montrent que les estimateurs EEG proposés et les estimateurs par linéarisation de la variance s'avèrent efficaces selon plusieurs plans d'échantillonnage, tant pour des réponses continues que pour des réponses binaires.

[1] Changbao Wu, University of Waterloo, Canada (cbwu@math.uwaterloo.ca); Ivan Carrillo Garcia, Statistique Canada

Une analyse de l'intention de cesser de fumer à l'aide de l'enquête ITC4 en tenant compte des réponses et des covariables manquantes

Baojiang Chen et Mary Thompson[1]

Résumé

Les études longitudinales comprennent souvent des données incomplètes sur la réponse et sur les covariables. L'effet des données manquantes dépend habituellement de leur fréquence et de la force de l'association entre les indicateurs de données manquantes et les variables de réponse. Lorsque la réponse ainsi que les données sur les covariables sont incomplètes, il est important de tenir compte de l'association entre les indicateurs de données manquantes pour ces deux processus au moyen de modèles conjoints. Des équations d'estimations généralisées pondérées par l'inverse des probabilités sont formulées pour traiter les données qui manquent au hasard. Des études empiriques montrent que les estimateurs convergents produits par les méthodes proposées présentent un biais empirique très faible pour des échantillons de taille moyenne, et qu'ils sont plus efficaces que d'autres méthodes ne tenant pas compte de l'association entre les processus de création des données manquantes. La méthode proposée est appliquée aux données de l'International Tobacco Control (ITC) Four Country Survey afin d'en démontrer l'utilité.

Mots clés : association; réponse manquante; covariable manquante.

1. Introduction

La recherche sur le tabagisme a montré que de nombreux fumeurs ne veulent tout simplement pas fumer. Environ 80 % déclarent qu'ils veulent arrêter. De 40 % à 50 % essaient d'arrêter chaque année, mais de 3 % à 5 % seulement réussissent à le faire pendant au moins 12 mois (Centers for Disease Control and Prevention 2002; Santé Canada 2002; Hyland et coll. 2004). Il est souhaitable d'étudier les facteurs qui influent sur l'intention d'arrêter de fumer, qui est souvent un précurseur des tentatives de renoncement au tabac.

La présente étude s'appuie sur des données provenant des six premières vagues de l'International Tobacco Control (ITC) Four Country Survey, une étude de cohorte menée auprès d'échantillons représentatifs des fumeurs adultes au Canada, aux États-Unis, au Royaume-Uni et en Australie. Dans cette enquête, les répondants de chaque pays sont interviewés annuellement en utilisant les mêmes protocoles d'enquête et les mêmes mesures. Les données fournies par un répondant peuvent être incomplètes, parce que des réponses et des données sur les covariables manquent ou parce que les sujets de l'étude omettent de répondre aux questions. Des problèmes d'inférence se posent si le mécanisme donnant lieu aux données manquantes est relié aux valeurs des réponses ou des covariables. Par exemple, la fatigue du répondant ou le degré d'accoutumance au tabac pourrait être associé à la non-réponse à des questions particulières dans un questionnaire long et approfondi. Le cas échéant, les analyses basées seulement sur les personnes qui fournissent des données complètes peuvent donner lieu à des inférences invalides. Sous un mécanisme de données manquant entièrement au hasard (MCAR pour *Missing Completely at Random)* (Little et Rubin 1987), les analyses fondées sur des équations d'estimations généralisées (EEG) (Liang et Zeger 1986) produisent des estimations cohérentes des paramètres de régression. Quand les données manquent au hasard (MAR pour *Missing at Random*) ou ne manquent pas au hasard (MNAR pour *Missing Not At Random*) (Little et Rubin 1987), les analyses fondées sur les EEG produisent des estimations non cohérentes des paramètres. Robins, Rotnitzky et Zhao (1995) ont élaboré une classe d'estimateurs fondés sur des équations d'estimations généralisées pondérées par l'inverse des probabilités (EEGPIP) dans un contexte de régression quand les données sont de type MAR.

Le but de la présente analyse illustrative est de déterminer si l'intention d'arrêter de fumer dépend de certaines covariables dans quatre pays, à savoir le Canada, les États-Unis, le Royaume-Uni et l'Australie, quand des valeurs manquent pour les réponses et pour les covariables. L'approche adoptée ici s'appuie sur des équations d'estimations pondérées par l'inverse des probabilités dans lesquelles il est tenu compte de l'association entre la réponse manquante et les données manquantes sur les covariables. Pour des précisions, consulter Chen, Yi et Cook (2009).

[1]Baojiang Chen, Department of Biostatistics, University of Washington, Seattle, Washington, US 98195 (bjchen@u.washington.edu); Mary Thompson, Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1 (methomps@uwaterloo.ca)

La présentation de l'article est la suivante. À la section 2, nous examinons la méthode pour traiter la réponse manquante et les covariables manquantes proposée par Chen, Yi and Cook (2009). À la section 3, nous appliquons la méthode proposée à l'analyse des données de l'International Tobacco Control Four Country Survey. Enfin, à la section 4, nous présentons nos conclusions.

2. La méthode proposée

À la présente section, nous présentons brièvement la méthode élaborée par Chen, Yi et Cook (2009) pour résoudre le problème des réponses manquantes et des covariables manquantes dans l'analyse de données longitudinales.

2.1 Formulation du modèle

Soit Y_{ij} la réponse du sujet i à la vague j pour $i = 1, \ldots, I$, et $Y_i = [Y]_{i1}, \ldots, Y_{ij})'$. Soit $\mu_{ij} = E(Y_{ij} | X_i, Z_i)$, où $X_i = [X]_{i1}, \ldots, X_{ij})'$ est la covariable pour laquelle des données manquent, et $Z_i = [Z]_{i1}, \ldots, Z_{ij})'$ est le vecteur des covariables qui est toujours observé. Le modèle de réponse est souvent donné par

$$[g(\mu_{ij}) = X_{ij}\ \beta_x + Z'_{ij}\ \beta_z$$

pour $j = 1, \ldots, J, i = 1, \ldots, n$, où g est une fonction de lien dérivable monotone, et $\beta = (\beta_x, \beta'_z)'$ est un vecteur de dimension $p \times 1$ de coefficients de régression d'intérêt. Soit $\mu_i = (\mu_{i1}, \ldots, \mu_{ij})'$.

Pour les processus de formation des données manquantes, nous désignons par R^y_{ij} l'indicateur d'absence de la réponse Y_{ij}, qui est égal à 1 si Y_{ij} est observée et à 0 autrement, et nous définissons de la même façon R^x_{ij} comme l'indicateur d'absence de la covariable X_{ij}. Soit $\lambda^y_{ij} = P\left(R^y_{ij} = 1 \middle| H^y_{ij}, H^x_{ij}, X_i, Y_i\right)$ et $\lambda^x_{ij} = P\left(R^x_{ij} = 1 \middle| H^y_{ij}, H^x_{ij}, X_i, Y_i\right)$, où $H^y_{ij} = \left\{R^y_{i1}, \ldots, R^y_{i,j-1}\right\}$ et $H^x_{ij} = \left\{R^x_{i1}, \ldots, R^x_{i,j-1}\right\}$ désignent, respectivement, les historiques des indicateurs d'absence de la réponse et de la covariable pour le sujet i, à la vague j. Habituellement, nous employons des modèles de régression logistique marginaux pour λ^y_{ij} et λ^x_{ij} à chaque période d'évaluation. En particulier, nous spécifions

$$\text{logit}\left(\lambda^y_{ij}\right) = u'_{ij}\alpha_y \text{ et } \text{logit}\left(\lambda^x_{ij}\right) = v'_{ij}\alpha_x, \qquad (1)$$

où u_{ij} et v_{ij} contiennent les fonctions de $[H]^y_{ij}, H^x_{ij}, Y_i, X_i, Z_i\}, \quad j = 2, 3, \ldots, J$, et α_y et α_x sont les paramètres de régression. Soit $\alpha_{xy} = (\alpha'_y, \alpha'_x)'$.

À chaque point dans le temps j, les situations d'observation de la réponse et de la covariable peuvent être associées à l'intérieur des sujets, en raison de facteurs communs affectant les deux processus d'observation. Pour modéliser cette association, nous traitons les processus d'observation de la réponse et de la covariable symétriquement et définissons le rapport de cotes conditionnelles

$$\psi_{ij} = \frac{P(R^y_{ij}=1, R^x_{ij}=1 | H^y_{ij}, H^x_{ij}, X_i, Y_i, Z_i) P(R^y_{ij}=0, R^x_{ij}=0 | H^y_{ij}, H^x_{ij}, X_i, Y_i, Z_i)}{P(R^y_{ij}=1, R^x_{ij}=0 | H^y_{ij}, H^x_{ij}, X_i, Y_i, Z_i) P(R^y_{ij}=0, R^x_{ij}=1 | H^y_{ij}, H^x_{ij}, X_i, Y_i, Z_i)},$$

qui est habituellement modélisé sous la forme

$$\log(\psi_{ij}) = \phi_j, \qquad j = 2, \ldots, J.$$

Dans le présent article, nous considérons un mécanisme de données manquant au hasard et supposons que

$$P(R^y_{ij} = r^y_{ij}, R^x_{ij} = r^x_{ij} | H^y_{ij}, H^x_{ij}, X_i, Y_i, Z_i) = P(R^y_{ij} = r^y_{ij}, R^x_{ij} = r^x_{ij} | H^y_{ij}, H^x_{ij}, X_i^{(o)}, Y_i^{(o)}, Z_i) \qquad (2)$$

pour chaque points dans le temps j, où $X_i^{(o)}$ désigne les parties observées de X_i, et $Y_i^{(o)}$ désigne les parties observées de Y_i.

2.2 Estimation et inférence

Pour l'estimation du paramètre β, nous pouvons employer les équations d'estimations généralisées pondérées par l'inverse des probabilités augmentées (EEGPIPA). Définissons $D_i = \frac{\partial \mu_i}{\partial \beta^T}$, $\Delta_i(\alpha) = [w_{ijk}]_{J \times J}$, où $w_{ijk} = \frac{I\left(R^y_{ij} = 1, R^y_{ik} = 1, R^x_{ik} = 1\right)}{\pi^{xy}_{ijk}}$, et

$\pi_{ijk}^{xyz} = P(R_{ij}^{y} = 1, R_{ik}^{y} = 1, R_{ik}^{y} = 1 | X_i, Y_i, Z_i)$. Soit $M_i = \kappa V_i^{-\frac{1}{2}}[C_i^{-1} \bullet \Delta_i(\alpha)]V_i^{-\frac{1}{2}}$, où $A \bullet B = [a_{ij}b_{ij}]$ désigne le produit de Hadamard des matrices $[A = [a]_{ij}]$ et $[B = [b]_{ij}]$ de dimensions $J \times J$, κ est un paramètre de dispersion, $V_i = diag\{var(Y_{ij}|X_i, Z_i), j = 1, \ldots, J\}$, et C_i est une matrice de corrélation de travail. L'EEGPIPA est donnée par

$$\sum_{i=1}^{n} U_i(\beta, \alpha) = 0 \quad , \qquad (3)$$

où $U_i(\beta, \alpha) = D_i M_i (Y_i - \mu_i) - \eta A_i(\alpha)$, $A_i(\alpha)$ est un vecteur de données observé dans lequel ne figure pas β, η est le coefficient de régression de $A_i(\alpha)$ dans la régression en population de $D_i M_i (Y_i - \mu_i)$ sur $[[A]_i, S_i]^r$, et S_i est la fonction de score de α pour le sujet i .

En pratique, les paramètres α du modèle de données manquantes sont inconnus, et on doit remplacer α dans (2) par une estimation convergente. La procédure d'estimation étant relativement classique, nous ne présentons pas les détails ici. En désignant l'estimateur résultant par $\hat{\beta}$, et sous certaines conditions de régularité, $n^{\frac{1}{2}}(\hat{\beta} - \beta)$ suit asymptotiquement une loi normale de moyenne 0 et de matrice de covariance $\Gamma^{-1}\Sigma[[\Gamma]^{-1}]^r$, où $\Gamma = E\left[\partial \frac{U_i}{\partial \beta^r}\right]$, et Σ est la matrice de covariance des résidus de la régression de $D_i M_i (Y_i - \mu_i)$ sur $[[A]_i, S_i]^r$.

3. Enquête ITC – Quatre pays

L'Enquête ITC – Quatre pays a débuté en 2002 et a été réalisée par vagues approximativement annuelles. La taille de l'échantillon a été maintenue à 2000 environ dans chaque pays à chaque vague, grâce au recrutement de nouveaux répondants pour remplacer ceux qui avaient cessé de participer à l'enquête. Les répondants à l'Enquête ITC – Quatre pays sont des fumeurs âgés de 18 ans et plus au moment du recrutement; ceux qui arrêtent de fumer sont gardés dans l'échantillon.

La réponse d'intérêt est l'intention d'arrêter de fumer des fumeurs au moment de l'enquête dont la situation d'usage du tabac est 1, 2 ou 3. (Les personnes dont la situation d'usage du tabac est 4, 5, 6 ou 7 sont des personnes qui ont arrêté de fumer au moment de l'enquête et pour lesquelles les données ne sont pas recueillies pour la variable d'intention d'arrêter.) L'intention d'arrêter est exprimée sous forme d'une variable binaire (1=Oui, O=Non). Les covariables d'intérêt comprennent le sexe, l'âge (18 à 24 ans, 25 à 39 ans, 40 à 54 ans, 55 ans et plus), le revenu (faible 15 000 £/30 000 $, moyen – entre £ 15 001$/30 001 $ et 30 000 £/59 999 $, élevé >30 001 £/60 000 $), le nombre de cigarettes fumées par jour, l'heure de la première cigarette de la journée en minutes et le niveau de scolarité (faible, moyen, élevé). Cependant, des données manquent pour les variables d'intention d'arrêter et de niveau de revenu. Le tableau 3-1 donne les proportions de données manquantes pour ces variables à chaque vague pour chaque pays.

Tableau 3-1
Proportion de données manquantes pour la réponse et pour les covariables pour les situations d'usage du tabac 1, 2 et 3

Pays	Canada		États-Unis		Royaume-Uni		Australie	
	Intention %	Revenu%	Intention%	Revenu%	Intention%	Revenu%	Intention%	Revenu %
Vague 1	13,78	19,47	16,15	20,64	12,76	20,17	11,28	15,76
Vague 2	5,76	11,94	9,86	13,95	3,87	10,39	3,98	8,25
Vague 3	1,32	6,94	1,60	5,93	1,31	8,92	1,46	6,10
Vague 4	1,52	7,26	2,60	6,63	1,84	9,18	1,05	5,65
Vague 5	1,96	7,25	1,90	5,54	1,89	8,82	1,28	6,56
Vague 6	1,58	7,27	1,84	5,69	1,47	10,05	1,28	6,09
Total	4,92	10,64	6,30	10,40	4,44	11,81	3,74	8,40

La proportion de données manquantes pour la variable de revenu diminue quand les vagues augmentent. Cela peut être attribué en partie au fait qu'à chaque vague, on a demandé aux personnes qui avaient répondu à la vague précédente si leur catégorie de revenu avait changé, tandis qu'on a de nouveau posé la question sur le revenu à celles qui n'avaient pas répondu. La proportion de données manquantes pour la variable d'intention d'arrêter diminue aussi quand les vagues augmentent, et la raison pour laquelle il en est ainsi n'est pas claire.

3.1 Modèles statistiques

Ici, nous construisons certains modèles pour appliquer la méthode proposée aux données de l'enquête ITC –Quatre pays. Soit Y_{ij} la réponse du sujet i à la vague $j = 1, \ldots, 6$, et $\mu_{ij} = P(Y_{ij} = 1 | X_{ij})$. Le modèle de réponse est donné par

$$\text{logit}[E(\mu_{ij})] = X'_{ij}\beta ,$$

où X_{ij} est le vecteur des covariables, incluant le sexe, l'âge, la racine carrée du nombre de cigarettes par jour (rcncpj), le logarithme de l'heure de la première cigarette en minutes (logheureprem), le niveau de scolarité et le niveau de revenu.

Pour les associations entre les indicateurs de données manquantes, nous construisons les modèles $(\lambda^x_{ij}) = u'_{ij}\alpha_x$ et logit $(\lambda^y_{ij}) = v'_{ij}\alpha_y$, où u_{ij} et v_{ij} sont les covariables qui influent sur la façon dont la réponse et les covariables manquent respectivement, lesquelles pourraient comprendre une fonction d'indicateurs de données manquantes antérieurs, le sexe, l'âge, la racine carrée du nombre de cigarettes par jour (rcncpj), le logarithme de l'heure de la première cigarette (logheureprem), le niveau de scolarité et le niveau de revenu. Pour les associations des indicateurs de données manquantes, nous construisons les modèles $\log(\psi_{ij}) = \phi_j, \quad j = 2, \ldots, 6$.

3.2 Résultats

Les tableaux 3.2-1 à 3.2-5 donnent les résultats des modèles de réponse et des modèles de données manquantes. Ici, nous commentons en détails l'analyse des résultats pour le Canada. Les résultats pour les trois autres pays sont fort semblables. Les tableaux 3.2-1 et 3.2-2 contiennent les données pour les modèles de réponse et les modèles de données manquantes au Canada.

Pour l'ensemble de données de chaque pays, nous comparons trois méthodes, à savoir la méthode proposée qui tient compte du mécanisme de formation des données manquantes, l'analyse des cas complets qui tient compte du mécanisme de formation des données manquantes uniquement par modélisation de l'indicateur d'observation tant de la réponse que de la covariable, et l'analyse des cas complets en ignorant le mécanisme de formation des données manquantes. Nous soulignons que l'analyse des cas complets produit des estimations biaisées quand les données manquent au hasard, tandis que la méthode proposée produit des estimations cohérentes et efficaces quand le modèle de données manquantes est spécifié correctement (Chen, Yi et Cook, 2009).

L'examen du tableau 3.2-2 montre que l'hypothèse que les données manquent au hasard (MAR) pourraient être plus appropriée que l'hypothèse que les données manquent entièrement au hasard (MCAR), tant pour la réponse que pour la covariable, puisque le logarithme de l'heure de la première cigarette, l'âge et le niveau de scolarité ont un effet significatif dans le modèle de réponse manquante, et que le sexe, le logarithme de l'heure de la première cigarette, la racine carrée du nombre de cigarettes par jour, l'âge et le niveau de scolarité ont un effet significatif dans le modèle de covariable manquante. En particulier, plus l'heure de la première cigarette de la journée est tardive (c.-à-d. moins le fumeur est « accoutumé et dépendant »), plus la probabilité de répondre est grande. Les indicateurs de données manquantes pour la réponse et pour les covariables présentent des corrélations sérielles, puisque r^x_{j-1} est modérément significatif dans le modèle de réponse manquante, et que r^x_{j-1} et r^y_{j-1} sont significatifs dans le modèle de covariable manquante. En outre, il existe certaines associations entre le mécanisme de réponse manquante et de covariable manquante, puisque ϕ_1, ϕ_2, ϕ_3, ϕ_4 et ϕ_5 sont significatifs. Il semble ne pas y avoir d'association significative à la vague 6, puisque ϕ_6 n'est pas significatif. Notons que l'estimation empirique de ϕ_6 est 0,0402, valeur qui est très faible.

Le tableau 3.2-1 donne les estimations pour les modèles de réponse pour des hypothèses de corrélation de travail non structurées dans le modèle EEG en utilisant la méthode proposée, ainsi que l'analyse des cas complets avec et sans pondération classique par l'inverse des probabilités pour le Canada. Les deux analyses des cas de données complètes donnent des résultats proches, qui diffèrent de ceux obtenus au moyen de la méthode proposée. Ici, nous faisons le commentaire que, comme dans le cas des nombreuses méthodes de réduction du biais, les erreurs-types sont plus grandes pour la méthode proposée, puisque la variation liée aux processus de formation des données manquantes intervient également dans l'estimation et dans les inférences. Pour toutes les méthodes, nous voyons que le sexe n'a pas d'effet significatif, ce qui indique que la différence d'intention d'arrêter entre les hommes et les femmes n'est pas significative. Le logarithme de l'heure de la première cigarette n'est pas significatif dans la méthode proposée, ce qui indique que l'heure de la première cigarette n'a pas d'effet significatif sur l'intention d'arrêter de fumer, tandis que son effet est significatif dans l'analyse des cas de données complètes. Cette différence peut être causée par un biais inhérent à l'analyse des cas de données complètes. L'intention d'arrêter de fumer dépend peut-être plus fortement de l'heure de la première cigarette chez les personnes pour lesquelles on observe à la fois la réponse et la variable de revenu. L'effet significatif de la racine carrée du nombre de cigarettes par jour dans la méthode

proposée indique que plus le nombre de cigarettes par jour est élevé, moins l'intention d'arrêter est probable; par contre, ce paramètre n'a pas d'effet significatif dans l'analyse des cas de données complètes. Cette situation peut également être causée par le biais inhérent aux résultats de l'analyse des cas complets. L'intention d'arrêter de fumer dépend peut-être plus fortement du nombre de cigarettes fumées par jour chez les personnes pour lesquelles soit la réponse, soit la variable de revenu manque. La tranche d'âge de 25 à 39 ans et la tranche d'âge de 40 à 54 ans n'ont pas d'effet significatif dans la méthode proposée, ce qui indique qu'il n'y a pas d'écart significatif entre l'intention d'arrêter de fumer chez les personnes du groupe de 25 à 54 ans, comparativement au groupe des 18 à 24 ans, tandis que l'effet de la tranche d'âge de 40 à 54 ans est significatif dans l'analyse des cas de données complètes, ce qui indique que l'intention d'arrêter de fumer dépend plus fortement de l'appartenance au groupe des 40 à 54 ans chez les personnes qui fournissent des données complètes. Les résultats reflètent aussi les erreurs-types plus grandes dans l'analyse proposée. L'effet de la tranche d'âge de 55 ans et plus est significatif dans toutes les méthodes, ce qui indique que les personnes de 55 ans et plus sont moins susceptibles d'avoir l'intention d'arrêter de fumer que les personnes de 18 à 24 ans. L'effet significatif du niveau de scolarité moyen (scolarité M) dans toutes les méthodes indique que les personnes dont le niveau de scolarité est moyen sont plus susceptibles d'avoir l'intention d'arrêter de fumer que les personnes dont le niveau de scolarité est faible; cependant, l'effet du niveau de scolarité élevé (scolarité E) n'est significatif dans aucune des méthodes, ce qui donne à penser qu'il n'y a pas de différence significative d'intention d'arrêter de fumer entre les personnes dont le niveau de scolarité est élevé et celles dont le niveau est faible. L'effet significatif du revenu élevé (revenu E) dans toutes les méthodes indique que les personnes dont le niveau de revenu est élevé sont plus susceptibles d'avoir l'intention d'arrêter de fumer que celles dont le niveau de revenu est faible, mais l'effet du revenu moyen (revenu M) n'est pas significatif, ce qui indique qu'il n'y a pas de différence significative d'intention d'arrêter de fumer entre les personnes dont le niveau de revenu est moyen et celles dont le revenu est faible

Tableau 3.2-1
Modèles de réponse pour les données de l'ITC4

	Canada			États-Unis			Royaume-Uni			Australie		
Paramètre	Est.	e.-t.	p	Est.	e.-t.	p	E.st.	e.-t.	p	Est.	e.-t.	P
Méthode proposée :												
Ord. à l'origine	2,091	0,242	<,001	1,520	0,198	<,001	1,219	0,256	<,001	1,579	0,329	<,001
Sexe M	-0,074	0,081	0,357	-0,108	0,072	0,130	-0,130	0,105	0,211	-0,214	0,100	0,032
Logheureprem.	0,026	0,027	0,322	0,118	0,024	<,001	0,031	0,029	0,262	0,053	0,030	0,074
Rcncpj	-0,183	0,033	<,001	-0,119	0,029	<,001	-0,123	0,047	0,008	-0,057	0,032	0,072
Âge (25 à 39)	0,217	0,184	0,238	-0,260	0,140	0,063	-0,064	0,153	0,671	-0,012	0,272	0,964
Âge (40 à 54)	-0,200	0,180	0,264	-0,461	0,137	0,001	-0,484	0,145	0,001	-0,669	0,275	0,015
Âge (≥ 55)	-0,555	0,185	0,002	-1,107	0,142	<,001	-1,079	0,191	<,001	-0,990	0,283	0,001
Scolarité M	0,196	0,089	0,027	0,178	0,080	0,025	0,468	0,147	0,001	0,143	0,127	0,257
Scolarité E	0,088	0,122	0,469	0,256	0,099	0,009	0,570	0,125	<,001	0,406	0,142	0,004
Revenu M	0,127	0,085	0,135	0,159	0,077	0,038	0,206	0,113	0,067	0,187	0,104	0,071
Revenu E	0,237	0,100	0,017	0,189	0,086	0,028	0,145	0,152	0,340	0,053	0,126	0,669
Cas complets – EGG pondérées :												
Ord. à l'origine	1,251	0,132	<,001	1,397	0,124	<,001	1,057	0,144	<,001	1,558	0,129	<,001
Sexe M	-0,092	0,069	0,179	-0,132	0,058	0,022	-0,212	0,060	<,001	-0,143	0,066	0,029
Logheureprem.	0,067	0,015	<,001	0,058	0,013	<,001	0,021	0,015	0,143	0,042	0,014	0,002
Rcncpj	-0,009	0,100	0,332	-0,051	0,009	<,001	-0,042	0,009	<,001	-0,019	0,008	0,024
Âge (25 à 39)	0,162	0,126	0,198	-0,248	0,120	0,038	-0,035	0,132	0,788	-0,200	0,121	0,097
Âge (40 à 54)	-0,265	0,121	0,027	-0,488	0,116	<,001	-0,609	0,130	<,001	-0,821	0,118	<,001
Âge (≥ 55)	-0,672	0,128	<,001	-1,011	0,119	<,001	-1,073	0,134	<,001	-1,194	0,129	<,001
Scolarité M	0,250	0,074	0,001	0,220	0,064	0,001	0,292	0,071	<,001	0,196	0,078	0,011
Scolarité E	0,088	0,102	0,389	0,225	0,088	0,010	0,311	0,094	0,001	0,329	0,100	0,001
Revenu M	0,074	0,074	0,313	0,113	0,064	0,077	0,155	0,067	0,019	0,200	0,074	0,006
Revenu E	0,231	0,087	0,007	0,115	0,077	0,132	0,296	0,076	<,001	0,153	0,080	0,056
Cas complets – EEG :												
Ord. à l'origine	1,243	0,130	<,001	1,391	0,124	<,001	1,037	0,145	<,001	1,545	0,130	<,001
Sexe M	-0,091	0,068	0,181	-0,134	0,058	0,021	-0,193	0,060	0,001	-0,143	0,066	0,029
Logheureprem.	0,065	0,015	<,001	0,060	0,013	<,001	0,023	0,015	0,107	0,042	0,014	0,002
Rcncpj	-0,007	0,100	0,426	-0,052	0,009	<,001	-0,042	0,009	<,001	-0,020	0,008	0,017
Âge (25 à 39)	0,159	0,126	0,207	-0,247	0,120	0,039	-0,026	0,133	0,840	-0,195	0,122	0,110
Âge (40 à 54)	-0,268	0,121	0,026	-0,485	0,116	<,001	-0,612	0,130	<,001	-0,814	0,120	<,001
Âge (≥ 55)	-0,688	0,128	<,001	-1,016	0,119	<,001	-1,077	0,135	<,001	-1,188	0,130	<,001
Scolarité M	0,259	0,074	0,000	0,223	0,064	0,001	0,290	0,071	<,001	0,191	0,078	0,013
Scolarité E	0,097	0,102	0,338	0,226	0,087	0,009	0,318	0,094	0,001	0,331	0,099	0,001
Revenu M	0,082	0,073	0,262	0,113	0,064	0,075	0,149	0,067	0,025	0,208	0,074	0,004
Revenu E	0,228	0,086	0,007	0,109	0,077	0,153	0,286	0,076	<,001	0,163	0,080	0,041

Tableau 3.2-2
Modèles de données manquantes sur les covariables pour les données de l'ITC4 : Canada

Paramètre	Estimation	e.-t.	p	Estimation	e.-t.	p
Ord. à l'origine	2,3144	0,2140	<,0001	0,7122	0,1413	<,0001
Sexe M	0,0657	0,1217	0,5893	0,1897	0,0890	0,0330
Logheureprem.	0,4291	0,0125	<,0001	0,3415	0,0114	<,0001
Rcncpj	0,0063	0,0176	0,7194	0,0357	0,0127	0,0049
Âge (25 à 39)	0,0086	0,2233	0,9691	0,3949	0,1498	0,0084
Âge (40 à 54)	-0,0279	0,2194	0,8989	0,5171	0,1478	0,0005
Âge ($\geq$ 55)	-0,4601	0,2301	0,0456	-0,1443	0,1521	0,3426
Scolarité M	0,2516	0,1343	0,0610	0,1581	0,0958	0,0989
Scolarité E	0,2992	0,1939	0,1228	0,5916	0,1523	0,0001
r_{j-1}^{y}	0,3850	0,3283	0,2410	7,9643	0,2607	<,0001
r_{j-1}^{y} : Revenu M	0,5499	0,1614	0,0007			
r_{j-1}^{y} : Revenu E	-0,3822	0,1591	0,0163			
r_{j-1}^{y}	0,5883	0,3188	0,0650	-4,6311	0,1757	<,0001
Association						
ϕ_1	4,9701	1,2835	0,0001			
ϕ_2	3,5034	1,0584	0,0009			
ϕ_3	1,7849	0,8936	0,0458			
ϕ_4	2,5133	0,9230	0,0065			
ϕ_5	1,3455	0,6810	0,0482			
ϕ_6	0,8465	0,5096	0,0967			

Tableau 3.2-3
Modèles de données manquantes sur les covariables pour les données de l'ITC4 : États-Unis

Paramètre	Estimation	e.-t.	p	Estimation	e.-t.	P
Ord. à l'origine	2,1944	0,1939	<,0001	0,9384	0,1422	<,0001
Sexe M	-0,1029	0,1133	0,3638	-0,0597	0,0885	0,4997
Logheureprem.	0,4367	0,0112	<,0001	0,3705	0,0106	<,0001
Rcncpj	0,0122	0,0166	0,4609	0,0483	0,0124	0,0001
Âge (25 à 39)	0,0275	0,1972	0,8893	0,4901	0,1517	0,0012
Âge (40 à 54)	0,3348	0,1955	0,0867	0,4787	0,1467	0,0011
Âge ($\geq$ 55)	-0,0567	0,2050	0,7822	0,1914	0,1527	0,2102
Scolarité M	0,0755	0,1213	0,5339	0,0889	0,0950	0,3491
Scolarité E	0,1923	0,1787	0,2820	0,1821	0,1362	0,1813
r_{j-1}^{y}	-1,3784	0,3351	<,0001	7,9744	0,2903	<,0001
r_{j-1}^{y} : Revenu M	-0,0751	0,1867	0,6874			
r_{j-1}^{y} : Revenu E	0,0902	0,2021	0,6556			
r_{j-1}^{y}	2,4721	0,3347	<,0001	-4,7668	0,2047	<,0001
Association						
ϕ_1	5,0828	1,4512	0,0005			
ϕ_2	4,2512	1,0868	0,0001			
ϕ_3	3,4018	1,6228	0,0361			
ϕ_4	2,9335	1,0354	0,0046			
ϕ_5	0,7749	0,5244	0,1395			
ϕ_6	-0,3517	0,5328	0,5092			

Tableau 3.2-4
Modèles de données manquantes sur les covariables pour les données de l'ITC4 : Royaume-Uni

Paramètre	Estimation	e.-t.	p	Estimation	e.-t.	p
Ord. à l'origine	2,2744	0,2658	<,0001	0,4124	0,1688	0,0146
Sexe M	0,0795	0,1298	0,5403	0,2819	0,0871	0,0012
Logheureprem.	0,4240	0,0125	<,0001	0,3146	0,0116	<,0001
Rcncpj	-0,0125	0,0243	0,6084	0,0175	0,0163	0,2833
Âge (25 à 39)	0,2152	0,2591	0,4063	0,7454	0,1685	<,0001
Âge (40 à 54)	0,2305	0,2599	0,3753	0,6626	0,1668	0,0001
Âge ($\geq$ 55)	0,0493	0,2632	0,8515	0,0619	0,1659	0,7092
Scolarité M	-0,1635	0,1541	0,2887	0,2356	0,1061	0,0264
Scolarité E	0,1443	0,2218	0,5151	0,3588	0,1515	0,0179
r_{j-1}^{y}	0,2980	0,3720	0,4232	9,0376	0,3205	<,0001
r_{j-1}^{y} : Revenu M	0,4139	0,1817	0,0227			
r_{j-1}^{y} : Revenu E	-0,1813	0,1805	0,3150			
r_{j-1}^{y}	0,8882	0,3605	0,0138	-5,5074	0,2544	<,0001
Association						
ϕ_1	5,0221	1,4043	0,0003			
ϕ_2	3,3933	1,3434	0,0115			
ϕ_3	2,5297	1,2118	0,0368			
ϕ_4	2,7053	1,1533	0,0190			
ϕ_5	0,4146	0,7030	0,5553			
ϕ_6	1,0570	0,8856	0,2327			

Tableau 3.2-5
Modèles de données manquantes sur les covariables pour les données de l'ITC4 : Australie

Paramètre	Estimation	e.-t.	p	Estimation	e.-t.	p
Ord. à l'origine	2,3486	0,1975	<,0001	0,7787	0,1357	<,0001
Sexe M	-0,0340	0,1296	0,7930	0,1463	0,0964	0,1291
Logheureprem.	0,3953	0,0123	<,0001	0,3106	0,0113	<,0001
Rcncpj	-0,0282	0,0208	0,1748	0,0316	0,0143	0,0270
Âge (25 à 39)	0,4498	0,1993	0,0240	0,7526	0,1431	<,0001
Âge (40 à 54)	0,4592	0,2040	0,0244	0,7949	0,1466	<,0001
Âge ($\geq$ 55)	0,3425	0,2298	0,1361	0,2426	0,1570	0,1223
Scolarité M	-0,1353	0,1596	0,3964	0,0415	0,1199	0,7296
Scolarité E	-0,2087	0,1932	0,2802	0,2882	0,1564	0,0653
r_{j-1}^{y}	0,4696	0,3765	0,2124	8,0150	0,2688	<,0001
r_{j-1}^{y} : Revenu M	-0,0010	0,1792	0,9954			
r_{j-1}^{y} : Revenu E	0,0286	0,1808	0,8745			
r_{j-1}^{y}	0,6997	0,3645	0,0549	-4,7754	0,1876	<,0001
Association						
ϕ_1	5,1733	1,5114	0,0006			
ϕ_2	3,2662	1,4836	0,0277			
ϕ_3	1,7442	1,1269	0,1217			
ϕ_4	1,4635	0,7352	0,0465			
ϕ_5	1,2803	0,4740	0,0069			
ϕ_6	1,3352	0,8523	0,1172			

4. Conclusion

Dans le présent article, nous appliquons la méthode des équations d'estimations généralisées pondérées par l'inverse des probabilités augmentées aux données de l'enquête ITC–Quatre pays en supposant que les données manquent au hasard. Cette approche comprend la modélisation du processus de formation des données manquantes et la pondération des équations d'estimations par l'inverse d'une probabilité qui est calculée en se basant sur les modèles établis pour les processus de formation des données manquantes. Les résultats des analyses des données manquantes donnent à penser que l'hypothèse des données manquantes au hasard pourrait être raisonnable, de sorte que cette méthode permet de corriger le biais que l'analyse des cas complets ne corrige pas. Par exemple, dans notre exemple, le facteur commun important lorsque la réponse ainsi que les données sur les covariables manquent pourrait être le degré d'accoutumance et de dépendance mesuré par (l'inverse du) logarithme de l'heure de la première cigarette; par exemple, plus un fumeur est accoutumé et dépendant, plus il pourrait être susceptible de pas répondre à la question d'intérêt ainsi qu'à la question sur la covariable parce qu'il devient moins patient vers la fin de l'interview. Nous pouvons réduire le biais dû à l'arrêt de participation à l'étude par une pondération appropriée. La méthode proposée ne rend la conclusion concernant la question de savoir si l'intention d'arrêter de fumer dépend du revenu plus précise que dans le cas des États-Unis. Cependant, pour tous les pays, l'estimation de la relation de dépendance entre l'intention d'arrêter de fumer et la racine carrée du nombre de cigarettes par jour révèle un effet de plus grande taille et pourrait être meilleure parce que nous ramenons effectivement dans l'analyse un plus grand nombre de décrocheurs fortement accoutumés.

Remerciements

La présente étude est financée en partie par une subvention du Conseil de recherches en sciences naturelles et en génie du Canada. Le projet de l'ITC est financé par des subventions du Roswell Park Transdisciplinary Tobacco Use Research Center du National Cancer Institute des États-Unis (P50 CA11236) et des Instituts de recherche en santé du Canada (57897).

Bibliographie

Centers for Disease Control and Prevention. (2002). Cigarette smoking among adults-- United States, 2002, Morbidity and Mortality Weekly Report, 51, pp. 642--645.

Chen, B., Yi, G. Y. wr Cook, R. J. (2009). Weighted Generalized Estimating Functions for Longitudinal Response and Covariate Data which are Missing at Random, *Journal the American Statistical Association*, in press.

Health Canada, Tobacco Control Programme. (2002). *Canadian Tobacco Use Monitoring Survey*. February-December 2001. Ottawa, Canada.

Hyland, A., Li, Q., Bauer, J. Giovino, G., Steger, C. et Cummings, K. M. (2004). Predictors of cessation in a cohort of current and former smokers followed over 13 years, *Nicotine & Tobacco Research*, 6, pp. S363--S369.

Liang, K. Y. et Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, 73, pp. 13--22.

Little, R. J. A. et Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley and Sons, Inc., 1st ed.

Robins, J. M., Rotnitzky, A. et Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association*, 89, 846--866.

Robins, J. M., Rotnitzky, A. et Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association*, 90, 106--121.

SESSION 13 – FACTEURS ET EFFETS DE LA NON-RÉPONSE

Les facteurs liés aux différents profils de non-réponse dans la English Longitudinal Study of Ageing (ELSA)

Hayley Cheshire et David Hussey [1]

Résumé

Il est d'importance capitale pour les responsables d'études longitudinales de comprendre les facteurs associés à l'attrition. Si les analystes peuvent déterminer les groupes les plus susceptibles de continuer ou de cesser de participer à l'étude, le processus de conception de l'enquête peut être adapté (p. ex., au moyen des pratiques sur le terrain) de manière à maximiser la probabilité de réponse.

Une revue de la littérature actuelle portant sur les personnes de 55 ans et plus (Bhamra et coll., 2008) a permis de déterminer certains facteurs liés à l'attrition – par exemple, le fait d'être plus âgé, d'avoir une déficience cognitive, d'avoir un plus faible statut socioéconomique et d'être moins instruit.

Nous proposons d'étudier les facteurs associés à différents profils de participation de la vague 1 à la vague 3 de l'English Longitudinal Study of Ageing (ELSA). L'ELSA est une étude portant sur les personnes de 50 ans et plus et leurs partenaires plus jeunes. Au total, 12 100 personnes étaient incluses initialement et ont fait l'objet d'un suivi par la suite.

Nous utilisons les groupes de comparaison suivants :
ont achevé les interviews dans toutes les vagues
ont cessé de participer à la vague 2 mais sont revenus à la vague 3
ont cessé de participer à la vague 2
ont cessé de participer à la vague 3

Nos analyses élargissent l'accent mis actuellement sur les facteurs démographiques de manière à inclure les variables d'enquête qui peuvent être utiles pour donner une indication du niveau d'engagement à l'endroit de l'étude à la vague 1. Les facteurs d'intérêt clés sont ceux liés au comportement de réponse, par exemple l'utilisation de points extrêmes ou médians sur les échelles de réponse, la non-réponse partielle, la volonté de consulter des documents durant l'entrevue et le consentement au couplage de données administratives gouvernementales.

[1] Hayley Cheshire (Hayley.Cheshire@natcen.ac.uk) et David Hussey (David.Hussey@natcen.ac.uk), National Centre for Social Research, Royaume-Uni

Étude empirique du biais de non-réponse dû à l'érosion de l'échantillon de la National Survey of College Graduates (NSCG)

Donsig Jang, John Finamore, David Hall, Steve Cohen, Flora Lan et Fan Zhang[1]

Résumé

La NSCG est une enquête biennale dont le but principal est de produire des estimations représentant la population cible de scientifiques et d'ingénieurs des États-Unis à une date de référence fixe. Le questionnaire détaillé du recensement décennal est utilisé comme base de sondage de cette enquête. Comme la base de sondage complète est disponible seulement une fois tous les dix ans, la NSCG utilise les données pour plusieurs rondes de son enquête ainsi que des échantillons supplémentaires périodiques de nouveaux diplômés dans les domaines des sciences et du génie.

Dans le cas d'une enquête longitudinale comme la NSCG, les non-répondants devraient être inclus dans l'échantillon des enquêtes de suivi afin de minimiser le biais. Toutefois, la plupart des non-répondants initiaux deviennent des refus persistants et, par conséquent, il est difficile d'obtenir leur collaboration à la ronde suivante. Pour cette raison, l'échantillon de la NSCG a fait l'objet d'un suivi dans trois rondes subséquentes seulement si les personnes continuaient de participer à l'enquête. Toutefois, il est prévu que l'inclusion seulement des répondants dans la prochaine ronde de la NSCG entraînerait un biais de sondage assez important même si l'on apportait les corrections habituelles de pondération pour la non-réponse. Pour comprendre le biais dû à la non-réponse des non-répondants longitudinaux, nous avons comparé les estimations produites pour l'échantillon fondé sur les répondants ayant rempli le questionnaire détaillé décennal des années 1990, données recueillies en 2003, à celles produites pour un nouvel échantillon tiré du Recensement de 2000 initialement visé par l'enquête en 2003. Dans cette communication, nous présentons les résultats de cet examen qui fournissent des éclaircissements empiriques sur les effets de l'attrition sur le biais d'échantillonnage.

[1] Donsig Jang, Mathematica Policy Research, É.-U. (DJang@Mathematica-Mpr.com); John Finamore (john.m.finamore@census.gov) and David Hall, U.S. Census Bureau, É.-U.; Steve Cohen, Flora Lan and Fan Zhang, National Science Foundation, É.-U.

Facteurs associés à la participation dans la cohorte GAZEL

Marie Zins, Jean François Chastang, Mireille Coeuret-Pellicer, Annette Leclerc, Sébastien Bonenfant, Alice Guéguen, Anna Ozguler et Marcel Goldberg[1]

Résumé

Contexte : La cohorte GAZEL a été mise en place en 1989 parmi les employés d'Électricité de France-Gaz de France âgés de 35 à 50 ans. L'inclusion s'est faite par questionnaire postal, et 20 625 sujets (15 011 hommes et 5 614 femmes) ont accepté de participer. Le suivi comporte un questionnaire postal annuel, ainsi qu'une invitation à se rendre dans un centre d'examens de santé (CES) pour une investigation médicale. La participation initiale était de 44.5 % ; chaque année environ 75 % des sujets renvoient le questionnaire ; 44.7 % se sont rendus dans un CES.

Objectifs : Étudier les facteurs sociodémographiques, professionnels, de comportement et de santé associés respectivement aux effets de sélection à l'inclusion, lors du suivi annuel et lors de l'examen de santé, et quantifier leur rôle.

Méthodes : À l'inclusion, les volontaires ont été comparés aux non participants selon des variables recueillies systématiquement dans les bases médico-administratives de la compagnie (absentéisme, mortalité, expositions professionnelles) par des modèles de régression logistique ; des modèles mixtes ont permis d'étudier la probabilité de répondre aux questionnaires annuels durant le suivi ; des modèles de régression logistique ont été utilisés pour étudier la probabilité de participation à l'examen de santé.

Résultats : Ce ne sont pas toujours les mêmes facteurs qui jouent un rôle aux différentes étapes de l'inclusion, du suivi et de la venue dans un CES ; l'ampleur des effets de sélection varie selon l'étape concernée. Cette étude permet de caractériser les biais potentiels qui peuvent être ainsi induits.

[1] Marie Zins (marie.zins@inserm.fr), Jean François Chastang, Mireille Coeuret-Pellicer, Annette Leclerc, Sébastien Bonenfant, Alice Guéguen (alice.guéguen@inserm.fr), Anna Ozguler et Marcel Goldberg, INSERM, France

Stratégies pour étudier le biais de non-réponse dans le programme Coset et la cohorte Constances

Gaëlle Santin[1], Laetitia Bénézet[1], Alice Guéguen[2], Rémi Sitta[2,3], Stéphanie Gauvin[1], Hélène Sarter[1], Nicolas Razafindratsima[4], Marie Zins[2,3], Béatrice Geoffroy-Perez[1], Marcel Goldberg[1,2]

Résumé

Le programme Coset et la cohorte Constances, dont certains objectifs sont de décrire et de suivre l'évolution de la santé de travailleurs en fonction de l'activité professionnelle sont des enquêtes longitudinales actuellement en phase pilote. Le taux de participation attendu à l'inclusion ne devrait pas excéder 30%. Pour corriger les biais de non-réponse, une stratégie basée sur l'exploitation d'informations auxiliaires issues des fichiers existants relatifs aux remboursements de soins, aux hospitalisations et au parcours professionnel des participants et des non-participants sera adoptée. Afin de discuter de cette stratégie, une enquête complémentaire sera réalisée auprès d'un échantillon de non-participants. Ces stratégies seront présentées en s'appuyant sur la phase pilote de la cohorte "Travailleurs agricoles" du programme Coset.

Mots Clés : enquête longitudinale, cohorte, non-réponse, biais de non-réponse.

1. Introduction

La non-réponse peut entraîner des problèmes récurrents dans les enquêtes (estimations biaisées, inflation de la variance) qui se poseront dans le cadre du programme Coset (Cohortes pour la surveillance épidémiologique en lien avec le travail) et de la cohorte Constances (Cohorte des consultants des centres d'examens de santé). L'un des objectifs de ces cohortes, actuellement en phase pilote, est de décrire et de suivre l'évolution de l'état de santé de la population en activité ou l'ayant été en fonction, entre autres, de l'activité professionnelle. Le recours à des cohortes en épidémiologie des risques professionnels est essentiel, car elles permettent de prendre en compte des expositions professionnelles "vie entière" ; en effet, il existe, pour certaines pathologies telles que le cancer, un temps de latence important (parfois plusieurs décennies) entre l'exposition au risque et la survenue de la maladie (Goldberg, 2008). Le taux de participation attendu dans le cadre du programme Coset et de la cohorte Constances ne devrait pas excéder 20 à 30%.

Dans cet article, le programme Coset sera présenté, en mettant plus particulièrement l'accent sur la phase pilote de la cohorte "Coset-Travailleurs agricoles" qui débutera au premier semestre 2010 (la cohorte Constances a été présentée par ailleurs dans le cadre de ce symposium (Gueguen, 2009)). Les deux méthodes qui seront mises en œuvre lors de cette phase pilote afin de minimiser les biais de non-réponse dans les estimations à l'inclusion sont :

- D'une part l'exploitation de fichiers administratifs contenant des informations auxiliaires, disponibles à la fois pour les participants et les non-participants. Relatives aux remboursements de soins, aux hospitalisations et au parcours professionnel, leur recueil n'est pas soumis au volontariat et ces informations permettront de corriger les pondérations issues du plan d'échantillonnage ;
- D'autre part, le recours à un échantillonnage double, en réalisant une enquête complémentaire auprès d'un sous-échantillon de non-participants.

Les estimations obtenues après correction de la non-réponse par repondération et issues de l'échantillonnage double seront confrontées et permettront de discuter sous plusieurs angles les biais liés à la non-réponse à l'inclusion en vue de l'extension

[1]Gaëlle Santin, Laetitia Bénézet, Stéphanie Gauvin, Hélène Sarter, Béatrice Geoffroy-Perez, Marcel Goldberg, Institut de veille sanitaire, 12 rue du Val d'Osne 94415 Saint-Maurice Cedex, France
[2]Alice Guéguen, Rémi Sitta, Marie Zins, Marcel Goldberg, Unité mixte Inserm-Cnamts 687/ IFR69, 16 avenue Paul Vaillant Couturier 94807 Villejuif Cedex, France
[3]Rémi Sitta, Marie Zins, Équipe RPP-C du Cetaf, 16 avenue Paul Vaillant Couturier 94807 Villejuif Cedex, France
[4]Nicolas Razafindratsima, Institut national d'études démographiques, 133 boulevard Davout 75980 Paris Cedex 20, France

nationale des cohortes "Coset-Travailleurs agricoles" et Constances. La phase pilote se déroulant actuellement, aucun résultat ne sera présenté.

2. Présentation générale du programme Coset

2.1 Objectifs du programme Coset

Le programme Coset a pour objectifs principaux de décrire à un instant « t » la morbidité/mortalité des actifs selon l'activité professionnelle (catégorie socioprofessionnelle et secteur d'activité) et son évolution, et de décrire l'évolution des liens entre la morbidité/mortalité des actifs et les expositions professionnelles.
Le programme Coset devra donc permettre de décrire la distribution de la morbidité ou de la mortalité et des fréquences d'expositions à certaines nuisances à l'échelle de la population active en France, c'est-à-dire à la fois de produire des estimations sans biais de quantités (taux d'incidence, prévalences …) concernant la population active en France à l'inclusion (représentativité longitudinale), ainsi que des estimations sans biais de quantités concernant la population active en France une année donnée (représentativité transversale).

2.2 Méthode générale : Étude des données de trois cohortes

Les données du programme Coset seront recueillies auprès d'actifs affiliés aux trois principaux régimes de protection sociale, qui couvrent 95% de la population au travail en France. Elles seront donc issues de trois cohortes:

- pour la population active salariée, affiliée au Régime général de sécurité sociale, soit environ 80 % des actifs en France, les données seront issues de la cohorte Constances (Gueguen, 2009), cohorte généraliste des affiliés à ce régime, en cours de mise en place par l'Institut national de la santé et de la recherche médicale (Inserm). Les objectifs et les domaines de recherche de cette cohorte sont nombreux, et elle permet entre autres de recueillir les informations nécessaires à la surveillance des risques professionnels. Ces dernières seront exploitées dans le cadre du programme Coset ; à terme, la cohorte Constances comprendra environ 200 000 personnes ;

- une cohorte de la population active des travailleurs agricoles affiliés à la Mutualité sociale agricole (MSA), actuellement en phase pilote, et une cohorte de la population active des travailleurs indépendants affiliés au Régime social des indépendants (RSI), actuellement en cours de réflexion. Ces deux cohortes seront mises en place par l'Institut de veille sanitaire en partenariat avec les régimes de protection sociale concernés. A terme, chacune d'entre elles devrait comprendre 35 000 personnes.

La méthodologie des trois cohortes sera, autant que possible, similaire.

2.3 Phase pilote de la cohorte « Coset-Travailleurs agricoles »

Pour la phase pilote de la cohorte "Coset-Travailleurs agricoles", les personnes sollicitées pour participer seront tirées au sort parmi les salariés et les non-salariés agricoles affiliés à une des cinq Caisses départementales pilotes, âgées de 18 à 65 ans et ayant été assurées au moins 90 jours en 2008, quel que soit leur type d'activité. La base de sondage utilisée sera la base d'assurance retraite des affiliés à la MSA. Dix mille personnes seront tirées au sort, soit 2 000 par Caisse départementale. Dans chaque Caisse, le tirage au sort, par sondage aléatoire simple, sera stratifié selon le sexe, l'âge et le statut d'emploi (salarié ou non-salarié) et sera proportionnel à la taille des strates.

Le recueil des données se fera par questionnaire postal auto-administré ; les informations recueillies concerneront l'état de santé (problèmes musculaires et articulaires, problèmes de santé psychique, problèmes cardio-vasculaires et respiratoires, cancer…) , l'activité professionnelle actuelle et passée (statut, type de contrat, temps de travail) et les expositions à certaines nuisances passées ou actuelles (contraintes organisationnelles et psychosociales, pénibilité, bruit, nuisances d'origine chimique, physique ou biologique) subies sur le lieu de travail. Une relance postale sera effectuée un mois après le premier envoi. Le suivi des participants sera annuel : un court questionnaire, concernant les nouveaux événements professionnels survenus durant l'année et l'état de santé, leur sera envoyé par courrier postal.

3. Traitement de la non-réponse totale lors de la phase pilote

Le taux de réponse attendu pour cette phase pilote est compris entre 20 et 30 %. Certaines estimations seront donc probablement concernées par des biais liés à la non-réponse. Dans cette partie sont présentées les stratégies de prise en compte de la non-réponse lors de la phase pilote de la cohorte "Coset-Travailleurs agricoles".

3.1. Correction de la non-réponse totale par repondération

Afin d'obtenir des estimations sans biais, il est prévu d'ajuster les poids d'échantillonnage pour la non-réponse, étant donné que les non-participants diffèreront probablement des participants et que ces différences pourront être liées aux variables d'intérêt (état de santé, parcours professionnel). L'ajustement pourra être réalisé, en formant des groupes de réponse homogène, en utilisant la méthode du score (Eltinge, 1997), ou par calage généralisé (Sautory, 2003).

- Sur quelles variables ajuster pour la non-réponse ?

L'étude des biais de sélection fait partie intégrante de l'épidémiologie, car cette discipline est continuellement confrontée à la non-réponse dans ses enquêtes (Goldberg, 2001b). Il a été montré que la participation à une enquête épidémiologique est liée à l'âge, à la catégorie sociale, à l'état de santé de la personne (Oleske, 2007) et aux comportements à risque pour la santé tels que la consommation d'alcool et de tabac (Goldberg, 2001a). Toutes ces variables étant également liées aux variables d'intérêt de Coset, il est primordial de les prendre en compte pour traiter la non-réponse.

- Quelles sont les sources de données disponibles ?

Lors de l'inclusion pilote, il est prévu d'accéder à de nombreuses informations auxiliaires, disponibles dans les bases informationnelles de la MSA et dans le fichier du Système national d'information inter-régime de l'assurance maladie (SNIIR-AM). Elles sont, d'une part, relatives à la santé des personnes tirées au sort (remboursement de médicaments et consultations, hospitalisations, accidents du travail ou maladies professionnelles) et d'autre part, relatives à leur activité professionnelle (statut d'emploi, temps de travail). Des informations sociodémographiques telles que l'âge et la catégorie sociale seront par ailleurs disponibles via ces mêmes fichiers.

Ces nombreuses informations auxiliaires seront présentes à la fois pour les participants et les non-participants. A l'inclusion, elles seront par ailleurs disponibles avec un recul d'au moins deux ans. Lors du suivi de la cohorte, ces informations seront recueillies annuellement. Compte tenu de leur nature et de leur diversité, elles devraient permettre de réaliser un ajustement pour la non-réponse de bonne qualité pour minimiser les biais dans les estimations et se ramener autant que possible à un processus de non-réponse ignorable.

Pour obtenir des estimations sans biais, le recours à des informations auxiliaires peut être cependant discuté, d'autant plus que le taux de réponse attendu est faible. Afin de pouvoir discuter cette méthode, une enquête complémentaire auprès d'un échantillon de non-participants sera réalisée lors de la phase d'inclusion pilote.

3.2. Enquête complémentaire auprès d'un échantillon de non-participants

3.2.1 Principe général

Hansen et Hurwitz (Hansen, 1946) ont été les premiers à réaliser une enquête complémentaire auprès de non-participants pour obtenir des estimations sans biais de moyennes et totaux, en se replaçant dans le contexte de l'échantillonnage double (ou sondage en deux phases).

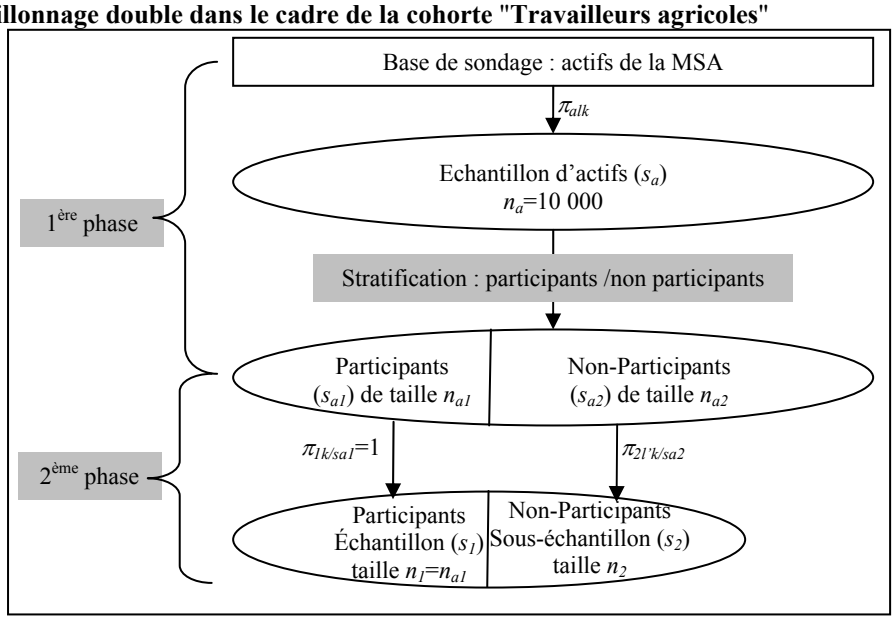
Appliqué à la cohorte "Coset-Travailleurs agricoles", le principe est le suivant (Särndal, 1992) (*cf*. figure 3.2.1-1) :
Un échantillon (s_a) de 10 000 actifs a été tiré au sort dans la base d'assurance retraite de la MSA, avec, pour l'individu k, une probabilité d'inclusion π_{alk} (l correspondant à la strate : Caisse départementale x sexe x classe d'âges x statut). A l'issue du recueil de données par autoquestionnaire postal, l'échantillon (s_a) est constitué d'un échantillon de participants (s_{a1}) et d'un échantillon de non-participants (s_{a2}) ; il peut donc être partitionné en une strate "participants" et une strate "non-participants". On sélectionne un échantillon de 2ème phase avec les probabilités d'inclusion suivantes :

- si l'individu k appartient à (s_{a1}), il aura une probabilité d'inclusion $\pi_{1k/sa1}$ égale à 1. L'échantillon obtenu (s_1) correspond exactement à (s_{a1}) et sa taille correspond au nombre de participants à l'enquête initiale.

- si l'individu k appartient à (s_{a2}), il aura une probabilité d'inclusion $\pi_{2l'k/sa2}$ (l' correspondant à la strate d'appartenance constituée par le département). On obtient alors un échantillon (s_2) de taille n_2. L'enquête complémentaire sera alors réalisée auprès de l'échantillon (s_2).

La condition suffisante pour obtenir des estimations sans biais est d'avoir un taux de réponse proche de 100% pour l'échantillon (s_2). C'est pourquoi le protocole opérationnel de l'enquête complémentaire auprès des non-participants prévoit des modes de collecte maximisant autant que possible le taux de réponse (questionnaire court, enquête par téléphone ou en face-à-face).

Figure 3.2.1-1
Principe de l'échantillonnage double dans le cadre de la cohorte "Travailleurs agricoles"



3.2.2 Échantillonnage pour l'enquête complémentaire

La population cible pour le tirage au sort de l'enquête complémentaire correspond à tous les individus non-participants à l'enquête postale initiale de la cohorte "Coset-Travailleurs agricoles". Les sujets seront sélectionnés par tirage au sort stratifié selon la Caisse départementale d'appartenance du non-participant. La taille de l'échantillon a été déterminée selon un principe de parcimonie entre les coûts de collecte et les effectifs, qui doivent être suffisamment grands pour obtenir des estimations fiables.

Un échantillon de 100 non-participants par Caisse départementale sera constitué par tirage au sort ; les personnes échantillonnées seront aléatoirement réparties en deux groupes, bénéficiant de modes de collecte différents : 70 seront enquêtées prioritairement par téléphone, 30 seront enquêtées prioritairement en face-à-face. Au total, 500 personnes seront donc échantillonnées.

3.2.3 Recueil des données

3.2.3.1 Données recueillies

Le questionnaire de l'enquête complémentaire est issu du questionnaire initial ; cependant afin de maximiser le taux de réponse, il a été considérablement réduit et la durée du recueil n'excèdera pas une dizaine de minutes. Seules ont été conservées les principales variables d'intérêt (état de santé général et emploi) et les variables *a priori* associées à la non-réponse telles que le sexe, l'âge, le statut marital, les comportements à risque pour la santé (alcool, tabac), l'état de santé général, le niveau d'étude, le statut vis-à-vis de l'emploi, le nombre d'épisodes professionnels au cours de la carrière, la profession et les principales expositions professionnelles.
Les données relatives au processus de collecte (paradonnées), telles que le nombre de contacts, les dates et heures des contacts, la durée d'interview, utiles pour comprendre la non-réponse dans l'enquête complémentaire seront également recueillies (Beaumont, 2005).

3.2.3.2 Modalités de recueil des données

Les modalités de recueil des données ont été choisies en fonction de leur capacité à obtenir un taux de réponse maximal. C'est pourquoi, le choix s'est porté au départ sur une enquête par téléphone plutôt qu'une enquête postale. Or, les numéros de téléphone n'étant pas disponibles dans les bases de la MSA, il est probable qu'un pourcentage non négligeable de numéros de téléphone ne sera pas retrouvé, car il n'existe pas à ce jour, en France, d'annuaire téléphonique exhaustif. Il a donc été décidé d'enquêter en face-à-face les personnes non-jointes par téléphone. Cependant, pour pouvoir étudier les biais potentiels liés aux différents modes de collecte des données (par téléphone et en face-à-face), il a été finalement décidé qu'un groupe soit enquêté en première intention en face-à-face.
Une saisie directe des réponses par les enquêteurs sera réalisée par l'utilisation des méthodes CATI (Computer Assisted Telephone Interview) et CAPI (Computer Assisted Personal Interview).
Afin de maximiser les chances de contacter les personnes et étant donné qu'il s'agit d'une population d'actifs, les appels ou les visites seront principalement réalisés aux créneaux horaires où les personnes ont le plus de chance de se trouver à leur domicile (en soirée ou le samedi). Jusqu'à 15 tentatives d'appels seront réalisées pour les enquêtes par téléphone et jusqu'à 3 visites pour les enquêtes en face-à-face à des jours et horaires différents.

3.3 Analyses des données

Au final, pour un paramètre d'intérêt Y , plusieurs estimations pourront être calculées et comparées :

- $\hat{Y}^{s_{a1}}$ obtenue à partir de l'échantillon (s_{a1}) ;
- $\hat{Y}^{s_{a1}Aj}$ obtenue à partir de l'échantillon (s_{a1}) après correction de la non-réponse par repondération ;
- $\hat{Y}^{s_2}$ obtenue à partir de l'échantillon (s_2) ;
- $\hat{Y}^{ED}$ obtenue à partir de l'échantillonnage double ($s_1 \cup s_2$).

Afin d'avoir une idée de l'ampleur du biais lié à la non-participation, et de l'effet de la correction, les estimations $\hat{Y}^{s_{a1}}$ et $\hat{Y}^{s_2}$, $\hat{Y}^{s_{a1}}$ et $\hat{Y}^{ED}$, $\hat{Y}^{s_{a1}}$ et $\hat{Y}^{s_{a1}Aj}$, $\hat{Y}^{s_{a1}Aj}$ et $\hat{Y}^{ED}$ pourront être comparées. On pourra dans un premier temps considérer que deux estimations ne sont pas trop différentes si leurs intervalles de confiance respectifs se chevauchent ; d'autres critères de comparaison plus élaborés pourront également être choisis ultérieurement.

Si le taux de réponse à l'enquête complémentaire est proche de 100% et s'il n'existe aucune autre source de biais que le biais de non-participation (par exemple de biais lié aux erreurs de mesure), l'estimation $\hat{Y}^{ED}$ sera approximativement sans biais. La comparaison de ces estimations permettra alors de mieux comprendre les biais liés à la non-participation et de discuter de la qualité des informations auxiliaires contenues dans les bases de données informationnelles pour corriger le biais de non-participation.

Il est cependant peu réaliste d'obtenir un taux de réponse à l'enquête complémentaire proche de 100% ; compte tenu des modalités de recueil des données mises en œuvre pour cette enquête, il semble néanmoins raisonnable d'attendre un taux de réponse autour de 70%. L'estimation $\hat{Y}^{ED}$ pourra alors être corrigée pour la non-participation à partir des informations auxiliaires contenues dans les bases de données informationnelles pour obtenir une estimation $\hat{Y}^{EDAj}$ et la comparer à $\hat{Y}^{s_{a1}Aj}$. Si ces deux estimations sont proches, on pourra considérer que le recours aux informations auxiliaires permet d'obtenir des estimations équivalentes à celles obtenues si l'on avait cherché à maximiser le taux de participation lors de l'enquête initiale. Si elles diffèrent, plusieurs hypothèses devront être discutées, comme l'insuffisance de l'information auxiliaire pour corriger la non-participation ou l'existence de biais de mesure dans l'enquête complémentaire (modes de recueil différents, réponses moins précises des personnes enquêtées lors de l'enquête complémentaire).

4. Discussion et perspectives

Lors de la phase pilote de la cohorte "Coset-Travailleurs agricoles", une comparaison des estimations obtenues après correction de la non-réponse par repondération et celles issues de l'échantillonnage double sera réalisée pour les variables communes à l'enquête initiale et à l'enquête complémentaire.

Cependant cette façon de procéder comporte certaines limites. Tout d'abord, cette comparaison ne pourra être faite pour toutes les variables d'intérêt puisque l'enquête complémentaire se limite à certaines questions. Le recours à des modes de collecte différents entre l'enquête initiale et l'enquête complémentaire pourrait en outre entrainer d'autres formes de biais. Par ailleurs, si l'information auxiliaire utilisée lors de la correction de la non-réponse par repondération n'est pas pertinente, cela risque

d'induire une augmentation des variances pour les estimations $\hat{Y}^{s_{a1}Aj}$, d'autant plus que le taux de participation attendu à l'enquête initiale est faible. De même, les variances des estimations $\hat{Y}^{ED}$ et $\hat{Y}^{s_2}$ risquent d'être élevées en raison des faibles effectifs de l'enquête complémentaire. Ceci peut être problématique car on envisage de se baser sur les intervalles de confiance pour comparer les estimations.

Cette approche comporte néanmoins certains atouts. Les informations auxiliaires sont nombreuses, liées aux variables d'intérêt (état de santé, parcours professionnel) et à la participation, exhaustives, indépendantes du volontariat et disponibles avec un recul d'au moins deux ans. L'enquête complémentaire pourra quant à elle être réalisée en maximisant au mieux le taux de réponse car elle sera réalisée sur un échantillon restreint, la base adresse est de très bonne qualité et le partenariat de la MSA, qui est très présente auprès de ses affiliés, est un gage d'une bonne collaboration de la part des personnes sollicitées.

Enfin, la confrontation de ces stratégies permettra d'avoir une approche globale pour comprendre et prendre en compte la non-participation dans la cohorte "Coset-Travailleurs agricoles".

Les résultats sont attendus au cours de l'année 2011, et permettront d'ajuster les modalités de mise en place de la cohorte "Coset-Travailleurs agricoles" lors de son extension nationale prévue pour 2012.

Remerciements

Les auteurs tiennent à remercier la Caisse centrale de la Mutualité sociale agricole pour son implication et sa collaboration active dans la mise en place de la cohorte « Coset-Travailleurs agricoles » et le recueil d'information auprès de ses affiliés.

Bibliographie

Beaumont, J.F. (2005). L'utilisation de renseignements sur le processus de collecte de données pour traiter la non-réponse totale au moyen de l'ajustement de poids, *Techniques d'enquête*, 31, p. 249-254.

Eltinge, J. L. et Yansaneh, I.S. (1997). Méthodes diagnostiques pour la construction de cellules de correction pour la nonréponse avec application à la nonréponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey, *Techniques d'enquête*, 23, p. 37-45.

Goldberg, M., Chastang, J.F., Leclerc, A., Zins, M., Bonenfant, S., Bugel, I., Kaniewski, N., Schmaus, A., Niedhammer, I., Piciotti, M., Chevalier, A., Godard, C. et Imbernon, E. (2001). Socioeconomic, demographic, occupational, and health factors associated with participation in a long-term epidemiologic survey: a prospective study of the French GAZEL cohort and its target population, *Am J Epidemiol,* 154, p. 373-84.

Goldberg, M. et Luce D. (2001). Les effets de sélection dans les cohortes épidémiologiques. Nature, causes et consequences *Rev Epidemiol Santé Publ*, 49, p. 477-492.

Goldberg, M., et Imbernon, E. (2008). Quels dispositifs épidémiologiques d'observation de la santé en relation avec le travail?, *RFAS*, 2-3, p.21-44.

Gueguen, A., Sitta, R., Bénézet, L., Santin, G., Lanoe, JL., Goldberg, M. et Zins M. (2009). L'apport des bases administratives et médico-administratives dans la cohorte Constances, *Recueil : Symposium 2009, Les enquêtes longitudinales: de la conception à l'analyse,* Statistique Canada.

Hansen, M. et Hurwitz, W. (1946). The problem of nonresponse in sample surveys, *JASA,* 41, p.517-29.

Oleske, D.M., Kwasny, M.M., Lavender, S.A. et Andersson, G.B. (2007). Participation in occupational health longitudinal studies : predictors of missed visits and dropouts, *Ann Epidemiol*, 17, p. 9-18.

Särndal, C., Swensson, B. et Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Sautory, O. (2003). Calmar 2 : Une nouvelle version du programme Calmar de redressement d'échantillon par calage, *Recueil : Symposium 2003, Défis reliés à la réalisation d'enquêtes pour la prochaine décennie,* Statistique Canada.

SESSION 14 – ENJEUX MÉTHODOLOGIQUES DIVERS

Répartition de l'échantillon de la 2010 Decade of the National Survey of College Graduates

John Finamore, David Hall, Donsig Jang, Stephen Cohen, Flora Lan et Fan Zhang[1]

Résumé

La National Survey of College Graduates (NSCG) est une enquête longitudinale biennale dont l'échantillon actuel a été établi à l'aide du questionnaire détaillé du Recensement de 2000. Étant donné que le questionnaire détaillé est en voie d'être remplacé par l'American Community Survey (ACS), nous prévoyons utiliser l'ACS comme base de sondage de la NSCG pour la décennie 2010. À l'étape de la planification, nous avons examiné les options en matière de conception de la NSCG pour la décennie 2010 et avons décidé d'adopter un plan de sondage avec renouvellement de panel. Au cours de la transition vers ce plan de sondage par panel rotatif, une partie de l'échantillon de la NSCG de 2010 sera sélectionnée à partir de la base de sondage de l'ACS et une autre partie provenant de l'échantillon de la NSCG pour la décennie 2000 sera reportée. Dans les cycles d'enquête ultérieurs, les cas de la décennie 2000 seront supprimés de l'échantillon par renouvellement et remplacés par les cas échantillonnés plus récemment pour l'ACS.

L'enquête NSCG pour la décennie 2010 sera conçue de manière à utiliser la base de sondage de l'ACS pour produire des estimations statistiquement fiables pour les principaux domaines d'analyse de la NSCG. Selon le financement consenti actuellement, environ 130 000 cas seront sélectionnés à partir de la base de sondage fondée sur l'ACS au cours de deux cycles de l'enquête NSCG pour la NSCG de 2010 et de 2012. Le présent article décrit nos recherches en vue de déterminer les principaux domaines d'analyse de la NSCG, d'établir les seuils de fiabilité pour ces domaines et d'élaborer un algorithme pour déterminer la répartition de l'échantillon sous les seuils de fiabilité. L'algorithme de répartition de l'échantillon portera initialement sur le plan de sondage de la NSCG de 2012 (le premier plan fondé intégralement sur l'ACS), mais permettra également d'examiner la répartition de l'échantillon de la NSCG en 2010, en 2014 et au delà.

[1] John Finamore (john.m.finamore@census.gov) and David Hall, U.S. Census Bureau, É.-U.; Donsig Jang, Mathematica Policy Research, É.-U. (DJang@Mathematica-Mpr.com); Stephen Cohen, Flora Lan and Fan Zhang, National Science Foundation, É.-U.

Life Pathways Project : conception et enjeux méthodologiques

Trivina Kang, Melvin Chan, Tan Teck Kiang, et David Hogan[1]

Résumé

Le présent article décrit le Life Pathways Project, une étude longitudinale d'environ 30 000 étudiants de plusieurs cohortes à Singapour. Malgré la fréquence de telles études à grande échelle au niveau international, elles sont relativement nouvelles à Singapour, particulièrement dans le domaine de l'enseignement. Cette étude présente la large gamme des résultats mesurés ainsi que le plan de recherche, et fournit des détails sur les modalités de collecte des données. Elle expose également l'important défi lié à l'attrition avec lequel les chercheurs se sont trouvés aux prises. Le problème s'est posé tout particulièrement en ce qui a trait à la cohorte d'étudiants au niveau postsecondaire dont une forte proportion a cessé de participer à l'enquête malgré les incitatifs offerts et les suivis effectués. Cela soulève des questions quant aux autres stratégies qu'il y a lieu d'utiliser pour faire en sorte que les jeunes continuent de participer activement à de telles enquêtes longitudinales.

Mots-clés : jeunes, scolarité, études longitudinales.

1. Introduction

En 2002, le Centre for Research in Pedagogy and Practice (CRPP) a été établi au National Institute of Education (NIE), seul établissement de formation des enseignants de Singapour. Grâce au soutien du ministère de l'Éducation, ce centre est devenu le plus grand centre de recherche pédagogique en Asie-Pacifique. Les recherches du CRPP portent principalement sur l'enseignement et l'apprentissage, et visent plus particulièrement à permettre de comprendre ce qui se passe dans les salles de classe de Singapour et d'élaborer des façons nouvelles et novatrices d'améliorer cette expérience.

Au fil des ans, les chercheurs au CRPP et au NIE ont exécuté de nombreux projets. Le présent article décrit l'un de ces projets, le Life Pathways Project (LPP), qui a été conçu pour recueillir des données sur une vaste gamme de résultats des étudiants. Ce projet se démarque d'autres projets non seulement par sa nature longitudinale, mais parce qu'il vise à mesurer les résultats scolaires et non scolaires d'un grand échantillon représentatif d'étudiants à Singapour. Ces résultats comprennent le rendement scolaire des étudiants aux évaluations des connaissances en anglais et en mathématiques ainsi que les données autodéclarées sur les compétences, dispositions et connaissances économiques, sociales, communautaires et psychologiques des étudiants.

2. Objet

Même si de telles études longitudinales ont été menées dans d'autres pays, il s'agit ici de la première étude à grande échelle de ce genre effectuée à Singapour. Les étudiants de Singapour se sont classés aux premiers rangs aux concours internationaux tels que la Troisième Enquête Internationale sur l'Enseignement des Mathématiques et des Sciences (TEIEMS), et leurs bons résultats scolaires sont connus. Toutefois, comme l'indique la récente série d'initiatives du ministère de l'Éducation, comme Enseigner moins Apprendre plus, Apprentissage social et émotionnel, Innovation et esprit d'entreprise, le ministère de l'Éducation de Singapour est de plus en plus résolu à former des étudiants qui non seulement visent l'excellence scolaire, mais sont en bonne position pour participer de façon utile dans divers domaines sociaux institutionnels et y sont préparés. Pour sa part, le LPP peut permettre de mieux comprendre les résultats de la scolarité puisqu'il porte sur l'acquisition d'une gamme de capacités dont les jeunes peuvent avoir besoin pour participer avec succès à la vie institutionnelle et sociale de Singapour et en retirer des avantages.

[1]Trivina Kang, Nanyang Technological University, 1 Nanyang Walk, Singapore, 637616 (trivina.kang@nie.edu.sg); Melvin Chan, Nanyang Technological University, 1 Nanyang Walk, Singapore, 637616; Tan Teck Kiang, Nanyang Technological University, 1 Nanyang Walk, Singapore, 637616; David Hogan, Nanyang Technological University, 1 Nanyang Walk, Singapore, 637616

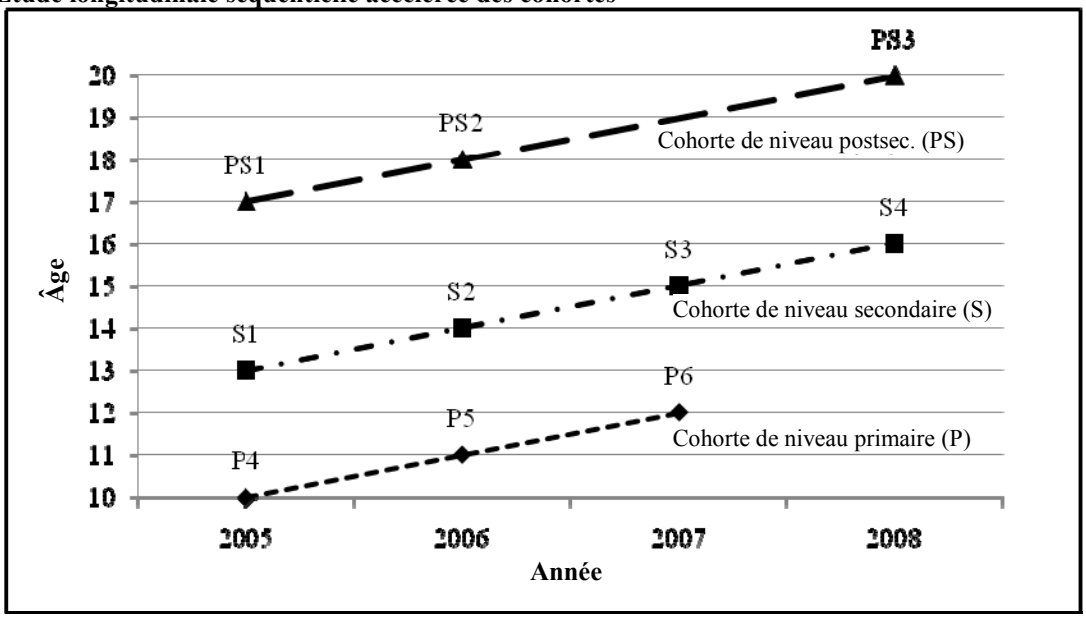
La gamme des compétences et capacités mesurées dans l'étude Life Pathways comprend les suivantes :

- Participation à la vie économique (« capital humain »)
  - Mesures conventionnelles du capital humain : rendement scolaire; niveau de scolarité (années de scolarité)
  - Capacités et compétences liées à « l'économie du savoir » (p. ex. initiation à la prise de risques; initiation à l'apprentissage permanent; initiation au travail individuel et d'équipe; adaptabilité; sens de l'action/auto-efficacité; compétences en résolution de problèmes interpersonnels; gestion du temps; littératie transversale; habiletés d'adaptation; gestion de l'information; compétences professionnelles interdisciplinaires, etc.)
  - Participation au marché du travail et réalisations (revenu, avancement, etc.)
- Aptitudes sociales et participation (« capital social »)
  - Réseaux sociaux; amitiés
  - Vie privée (famille, amis, ami(e)s de cœur)
  - Confiance et attachement/appartenance
  - Compétences en leadership
  - Appartenance à un groupe communautaire
- Participation à la vie communautaire (« capital communautaire »)
  - Croyances communautaires (civisme, justice, identité nationale, etc.)
  - Action communautaire (participation à des groupes politiques, intentions de vote, intérêt pour la lecture et les discussions sur l'actualité)
  - Dispositions communautaires (intérêt/engagement, confiance, respect d'autrui)
  - Identité communautaire/identités nationales (attachement, appartenance)
- Subjectivité (« auto-gestion »)
  - Croyance en soi (auto-efficacité/action subjective, image de soi, confiance en soi, source de détermination)
  - Valeurs
  - Buts et aspirations
  - Formations de l'identité (sociale, personnelle)
  - Réflexivité/connaissance de soi (y compris métaconnaissance)
- Bien-être subjectif
  - Aspirations et satisfaction existentielles
  - Problèmes de santé mentale (anxiété, dépression, anxiété face aux examens)
  - Incidence positive et négative
  - Satisfaction à l'égard de la vie liée à un domaine précis (école, famille, amitié, etc.)
  - Satisfaction à l'égard de la vie et bonheur
- Objectifs et choix de vie, cheminement
  - Objectifs et aspirations de vie
  - Engagement scolaire et réalisations
  - Travail, carrière et participation au marché du travail
  - Participation à la vie sociale (réseaux, amitiés, vie privée; société civile)
  - Fondation d'une famille
  - Participation à la vie communautaire et politique
  - Autoformation (biographie)

3. Plan de recherche

Nous étudions trois cohortes d'étudiants, de niveau primaire, secondaire et postsecondaire, et notre plan de recherche nous permet de disposer de données sur environ dix ans de scolarité, puisqu'il y a chevauchement de la cohorte de la 4^{e} primaire (P4, 10 ans) et de celle de la 1re secondaire (S1, 13 ans), puis avec la cohorte de la 1re postsecondaire (PS1, 17 ans).

Figure 1
Étude longitudinale séquentielle accélérée des cohortes



Les cohortes P4 et S1 comprennent les étudiants de toutes les classes de quelque 77 écoles. Cela permet de procéder à une analyse transversale multiniveaux des variables de niveau de l'école (niveau 3) portant sur les étudiants (niveau 2) et le temps (niveau 1).

Au total, nous avons élaboré quatre instruments d'enquête pour les deux cohortes plus jeunes, soit P4 et S1, et trois instruments pour la cohorte PS1. Les instruments d'enquête diffèrent quant à leurs mesures des résultats clés, mais tous comprennent des modules communs de prédicteurs. Les instruments portent entre autres sur les mesures de résultats suivantes : formulaire 1 : Capital économique et social; formulaire 2 : Bien-être subjectif; formulaire 3 : Planification de la vie et identités; formulaire 4 : Civisme et capital communautaire. Comme les étudiants au niveau postsecondaire sont plus âgés, et donc en mesure de répondre à un plus grand nombre de questions d'enquête, les questions dans le formulaire 3 de l'enquête au niveau secondaire ont été ajoutées aux formulaires 2 et 4 de l'enquête au niveau postsecondaire. Ainsi, seulement trois formulaires ont été administrés à cette cohorte.

Au moment de la première vague de l'enquête (auprès des 10, 13 et 17 ans), les formulaires ont été distribués aléatoirement aux étudiants. Dans les années subséquentes, ces étudiants ont été dirigés vers le même formulaire à leur entrée dans le site Web de l'enquête à l'aide de leur numéro d'identification individuel. On a communiqué aux répondants une adresse Internet unique (URL) et un mot de passe leur permettant d'avoir accès aux formulaires de l'enquête annuelle et de les remplir.

En outre, les étudiants faisant partie de la cohorte du secondaire ont participé à une évaluation papier crayon. En 2005, l'organisme Educational Assessment Australia (EEA) à l'University of New South Wales (UNSW) a été chargé de concevoir et de produire des instruments d'évaluation annuelle des connaissances en mathématiques et en anglais pour ce projet. Les instruments d'évaluation ont été élaborés par l'EAA dans le cadre de ceux fournis par l'organisme International Competitions and Assessments for Schools (anglais et mathématiques). Ces instruments sont appelés ci-après « évaluations UNSW ». L'instrument d'évaluation des connaissances en anglais comprenait 64 questions portant sur les domaines suivants : lecture (littératie); lecture (textes factuels); langue (rédaction de texte); langue (syntaxe); et orthographe. L'instrument d'évaluation des connaissances en mathématiques comprenait 40 questions portant, entre autres, sur les concepts suivants : les nombres; les mesures; le hasard et les données; l'espace et la géométrie; l'algèbre. Ces évaluations UNSW ont permis aux responsables du projet d'obtenir des données objectives sur le rendement scolaire de l'ensemble de la cohorte du secondaire qui ont été étalonnées au fil du temps.

Au niveau de la 1re secondaire, on a affecté au hasard l'évaluation UNSW des connaissances en anglais à la moitié des étudiants de la cohorte du secondaire dans chaque école sélectionnée et l'évaluation des connaissances en mathématiques, à l'autre moitié. Les étudiants ont continué de subir l'évaluation de la matière qui leur était affectée au cours des quatre années.

4. Le problème de l'attrition

Comme nous l'avons mentionné ci-dessus, les échantillons pour l'étude étaient grands et, une fois identifiée, l'entière cohorte des étudiants du niveau (P4, S1 ou PS1) a participé à l'étude. Comme les étudiants devaient être dirigés vers le même instrument d'enquête chaque année, nous avons utilisé leur numéro d'identification individuel aux fins d'administration des formulaires pertinents. Pour les cohortes P4 et S1, nous avons pu obtenir la collaboration et le soutien des écoles (37 écoles primaires et 38 écoles secondaires) pour prévoir des tranches de temps où les étudiants accéderaient à notre site Web et rempliraient nos formulaires d'enquête. Malgré une certaine attrition, tenant principalement à ce que des étudiants quittaient l'école, comme le montrent les tableaux 4-1 et 4-2, les taux de réponse ont été suffisants pour les cohortes des niveaux primaire et secondaire. Les taux de réponse aux enquêtes en ligne ont été 94,4 % au niveau primaire et 71,5 % au niveau secondaire. Toutefois, il est difficile d'empêcher une forte attrition de la cohorte de niveau postsecondaire. Cela est attribuable dans une large mesure au fait que les établissements postsecondaires étaient moins disposés à nous accorder l'accès à leurs étudiants. Malgré plusieurs appels de suivi et divers incitatifs, tels que des billets de cinéma et des tirages au sort offrant des prix comme des ordinateurs portatifs Apple, il a été difficile d'obtenir la participation des étudiants.

Tableau 4-1
Nombre de répondants participant à l'étude Life Pathways

	Primaire	Secondaire	Postsecondaire
Vague 1	9 620	11 627	8 836
Vague 2	10 849	11 904	5 819
Vague 3	10 978	11 743	2 318
Vague 4	--	11 198	--
Total	9 084	8 303	2 077

Tableau 4-2
Nombre de répondants participant aux évaluations UNSW des connaissances en anglais et en mathématiques

	UNSW Mathématiques	UNSW Anglais
Vague 1	6 052	6 063
Vague 2	6 099	6 109
Vague 3	5 501	5 536
Vague 4	5 815	5 977
Total	4 265	4 246

Étant donné ce taux d'attrition élevé, nous avons dû limiter notre utilisation des données longitudinales de la cohorte de niveau postsecondaire aux fins de modélisation. Nous nous sommes également rendu compte que, malgré la technologie et les enquêtes en ligne qui ont l'avantage de garantir que tous les champs dans le formulaire d'enquête sont remplis, il faut trouver une solution à un problème de longue date, celui de faire en sorte que les répondants participent et, dans le cas d'études longitudinales, qu'ils continuent de participer. Dans un contexte comme celui de Singapour, où il y a probablement moins d'engagement et d'intérêt pour la participation à une enquête, il faut veiller en élaborant le plan d'enquête à prendre des mesures pour garantir que les taux d'attrition demeurent faibles. À l'avenir, les chercheurs devraient également se pencher sur la possibilité que les enquêtes en ligne ne conviennent peut-être pas à tous les groupes de population, même si elles sont rentables et permettent de compiler les données facilement. En outre, un examen plus poussé des types d'incitatifs qui seraient attrayants pour les répondants, particulièrement les membres de la génération Y, serait fort utile.

La conception et l'analyse de données longitudinales à Statistique Nouvelle-Zélande

Deborah Brunning[1]

Résumé

Avant le tournant du siècle, Statistics New Zealand avait peu d'expérience de la conception et de l'analyse de données longitudinales. Au cours des dix dernières années, toutefois, les décideurs en Nouvelle-Zélande, comme leurs homologues dans de nombreux autres pays, ont déterminé qu'ils avaient besoin de plus de renseignements pour pouvoir étudier les profils et les dynamiques mieux que ne le permettent les données transversales ponctuelles recueillies périodiquement. Pour répondre à ce besoin, au cours des dix dernières années, Statistics New Zealand a conçu et exécuté sept vagues d'une enquête à huit vagues visant à mesurer la dynamique du revenu et de l'emploi (connue sous le nom de SoFIE); a réalisé, de concert avec le ministère du Travail, deux vagues d'une enquête longitudinale à trois vagues visant à mesurer l'expérience d'établissement des immigrants en Nouvelle-Zélande (LISNZ); a élaboré une base de données longitudinales sur les entreprises en combinant des données provenant de plusieurs sources; et a élaboré un ensemble de données qui lui permet d'examiner les dynamiques et les profils longitudinaux des employeurs et des employés, à partir des données administratives du système d'imposition (LEED).

La présente communication porte sur notre participation à ces travaux. Nous exposons nos réalisations ainsi que les importants défis qu'il a fallu relever dans des domaines comme la conception et la mise en œuvre de méthodes de collecte, l'utilisation de méthodes assistées par ordinateur, ainsi que la confidentialité et l'accès aux données. Nous examinerons comment ces expériences influencent notre réflexion concernant les futures collectes de données longitudinales à Statistics New Zealand.

Mots-clés : longitudinal, entreprise, données administratives, social, interview assistée par ordinateur, accès aux données.

1. Introduction

Au cours de la dernière décennie, Statistics New Zealand a procédé à quatre collectes de données longitudinales. La présente communication fournit un bref aperçu de ces quatre collectes, ainsi qu'un examen de certaines des expériences liées à ces travaux et de la façon dont ils influencent la réflexion pour les futures collectes de données longitudinales à Statistics New Zealand.

1.1 Aperçu des collectes longitudinales de données sociales de Statistics New Zealand

Enquête sur la famille, le revenu et l'emploi (SoFIE)
L'Enquête sur la famille, le revenu et l'emploi (SoFIE) est une enquête nationale de la Nouvelle-Zélande qui est conçue pour étudier les changements qui touchent le revenu des personnes, des familles et des ménages, ainsi que les facteurs qui les influencent chez les personnes, comme la participation au marché du travail et la composition de la famille. La SoFIE est constituée d'un panel fixe et vise à réinterviewer le même groupe de personnes sur une période de huit ans (ou vagues). À la vague 1, environ 15 000 ménages ont été sélectionnés de façon aléatoire pour participer à la SoFIE, et des données ont été recueillies auprès de 22 000 personnes admissibles dans 11 500 ménages.

On utilise à la fois un questionnaire destiné aux ménages et un questionnaire destiné aux personnes pour recueillir des données auprès des répondants de la SoFIE à chacune des vagues. Le questionnaire des ménages est rempli par un adulte de chaque ménage et sert à recueillir des données sur les caractéristiques du ménage (type de ménage, type de famille) et le niveau de vie (type de logement, appareils électroménagers). Le questionnaire destiné aux personnes est rempli par chaque répondant adulte de la SoFIE et sert à recueillir des données sur les caractéristiques démographiques, les antécédents et les activités actuelles sur le marché du travail, la famille, la scolarité et le revenu.

Enquête longitudinale sur l'immigration : Nouvelle-Zélande (LisNZ)
L'Enquête longitudinale sur l'immigration : Nouvelle-Zélande (LisNZ) est une enquête longitudinale auprès des immigrants qui a été mise au point par le ministère du Travail, en partenariat avec Statistics New Zealand (qui a mené l'enquête). Le

[1]Deborah Brunning, Statistics New Zealand, P.O. Box 2922, Wellington, New Zealand, deborah.brunning@statsgovt.nz, www.stats.govt.nz

principal objectif de la LisNZ est d'éclairer les politiques en matière d'immigration et d'aider à l'élaboration des services d'établissement.

Dans le cadre de la LisNZ, on a interviewé le même groupe d'immigrants pendant trois vagues; 6 mois (vague 1), 18 mois (vague 2) et 36 mois (vague 3), après qu'ils soient devenus résidents permanents de la Nouvelle-Zélande. L'échantillon de l'enquête a été sélectionné à partir des immigrants de 16 ans et plus qui ont obtenu le statut de résident permanent de la Nouvelle-Zélande du 1er novembre 2004 au 31 octobre 2005. Les immigrants ont été échantillonnés au moment où ils ont obtenu leur statut de résident. Dans le cas des immigrants dont le statut de résident a été approuvé à l'étranger, l'interview de la première vague a pu avoir lieu jusqu'à 18 mois après la date d'approbation, ces immigrants ayant 12 mois pour venir en Nouvelle-Zélande et s'y établir. Les interviews de la vague 1 ont commencé en mai 2005 et celles de la vague 3 ont pris fin en novembre 2009.

La LisNZ a été conçue pour produire des estimations pour plusieurs sous-populations définies selon la catégorie d'immigrant et la région d'origine. On estime que 5 000 interviews complètes à la vague 3 sont nécessaires pour produire des estimations comportant les niveaux d'exactitude requis. Par conséquent, un échantillon initial important de plus de 12 000 personnes a été sélectionné au moyen d'un plan de sondage aléatoire systématique stratifié.

Pour des raisons pratiques et opérationnelles, la population d'enquête de la LisNZ s'est limitée aux immigrants qui, à la première vague, vivaient à North Island, South Island ou Waiheke Island (excluant ainsi la plupart des petites îles au large de la Nouvelle-Zélande) et qui pouvaient aussi comprendre au moins une des sept langues désignées pour l'enquête (anglais, mandarin, cantonais, samoan, coréen, hindi et pendjabi).

1.2 Aperçu des collectes longitudinales de données auprès des entreprises de Statistics New Zealand

Données couplées employeur-employé (LEED)
Les Données couplées employeur-employé (LEED) intègrent les données existantes sur les employeurs et les employés, afin de donner un aperçu du fonctionnement du marché du travail en Nouvelle-Zélande et de ses liens avec le rendement des entreprises. Les LEED utilisent les données administratives existantes de deux sources : données fiscales d'Inland Revenue et données sur les entreprises du registre des entreprises de Statistics New Zealand (BF). L'ensemble de données LEED résulte du couplage d'une série de données longitudinales sur les employeurs du BF et d'une série de données longitudinales sur la paie des horaires mensuels des employeurs (EMS). Les données EMS portent sur toutes les personnes qui touchent un revenu au sujet duquel des retenues à la source sont effectuées. À l'intérieur de ces données fiscales, les employés individuels sont liés à un employeur à un moment donné, ce qui produit des paires uniques d'employeur-employé. L'ensemble de données LEED fournit un lien important entre les données sur les entreprises et les données sur les personnes.

La première diffusion des statistiques LEED a eu lieu en février 2006, ce qui a permis de disposer d'une série rétroactive de données de juin 1999 à décembre 2004. Au niveau trimestriel, les statistiques LEED ont continué d'être produites pour chaque trimestre depuis la première diffusion.

Base de données longitudinales sur les entreprises (LBD)
La Base de données longitudinales sur les entreprises (LBD) est un prototype de base de données longitudinales comprenant des renseignements sur les entreprises qui ont été créées à l'origine par suite de deux années de travaux du projet de faisabilité IBULDD (Improved Business Understanding via Longitudinal Database Development).

La LBD intègre des données administratives et des données d'enquête qui existent déjà à Statistics New Zealand et dans d'autres organismes gouvernementaux. Le cœur de la LBD est constitué de la base longitudinale des entreprises (LBF). La LBF est un registre longitudinal des entreprises qui comprend des renseignements longitudinaux (comme l'industrie, le type de propriété et le secteur) au sujet d'une vaste population d'entreprises. Parmi les autres composantes clés figurent les données fiscales d'Inland Revenue, qui comprennent des renseignements sur les ventes et les achats des entreprises; des variables de rendement financier et de situation; ainsi que des chiffres sur les salaires, les traitements et le nombre d'employés. La LBD comprend aussi un certain nombre d'enquêtes sur échantillon de Statistics New Zealand qui mesurent les pratiques et le rendement des entreprises.

La LBD est un registre exhaustif d'entreprises qui est actif depuis 2000. La dynamique de la création, de la suppression, de l'expansion et des contractions d'entreprises est saisie dans les données longitudinales.

2. Expériences

2.1 Expériences sociales

La SoFIE est la première expérience d'enquête sociale longitudinale jamais menée par Statistics New Zealand, ce qui fait qu'il n'existe pas d'expérience ou d'expertise longitudinale antérieure dont Statistics New Zealand pourrait tirer parti. L'approche initiale adoptée a été influencée dans une large mesure par les expériences transversales précédentes. Toutefois, cela n'était pas suffisant, étant donné que nous avons appris depuis que les collectes longitudinales exigent une approche différente en ce qui a trait au processus d'enquête (de la conception à l'analyse) de celle des enquêtes transversales. Dans la plupart des cas, l'élaboration et le fonctionnement permanent de la LisNZ ont connu moins de turbulences que la SoFIE, ce qui fait que l'expertise et les leçons apprises des premières expériences de la SoFIE ont été appliquées à la LisNZ, au besoin.

Le système de traitement de la SoFIE englobe toutes les transformations requises (p. ex., calcul des variables, imputation, ajustements pour la non-réponse et étalonnage) des données, qui vont de la post-vérification (après la collecte) à la production d'un ensemble de données convenant à l'analyse et de microdonnées pour la diffusion. Différents problèmes concernant les aspects du système de traitement se sont posés depuis la conception et l'élaboration originales, et la fréquence et l'ampleur des problèmes ont continué d'augmenter au fur et à mesure des vagues. Au cours des cinq dernières années, une somme considérable de ressources a servi à examiner les problèmes, à les documenter et à en résoudre certains. Toutefois, en dépit de ces efforts, le système est devenu tellement instable globalement qu'il était impossible à utiliser, ce qui a mené à la diffusion des données de la vague 5 à la fin de 2008. Les diffusions prévues des données de la SoFIE ont dû être reportées, afin de permettre au système de traitement de se stabiliser. Le projet de stabilisation du système de la SoFIE est actuellement en cours et devrait prendre fin au milieu de 2010. Avec un système de traitement fonctionnant convenablement, Statistics New Zealand entend recommencer à respecter les calendriers de diffusion de la SoFIE, la diffusion des données des vagues 1 à 7 étant prévue à la fin de 2010. Les problèmes de traitement des données de la SoFIE n'ont pas eu de répercussions sur la collecte proprement dite des données, qui s'est poursuivie sur le terrain comme prévu.

On a recours aux interviews sur place assistées par ordinateur (IPAO), tant pour la SoFIE que pour la LisNZ. La SoFIE était la première enquête de Statistics New Zealand pour laquelle on a commencé l'élaboration du questionnaire et la collecte sur le terrain au moyen d'une forme d'interviews assistées par ordinateur (IAO). À l'heure actuelle, toutes les enquêtes sociales auprès des ménages de Statistics New Zealand (tant transversales que longitudinales) utilisent l'IPAO ou les interviews téléphoniques assistées par ordinateur (ITAO). L'utilisation prévue de l'IAO, à partir du début des deux collectes longitudinales sociales, a constitué un avantage important du point de vue de la collecte des données requises. Le type d'information recueillie auprès des répondants et les détails de celle-ci, de même que le cheminement complexe requis entre les questions, n'auraient simplement pas été possibles dans des interviews sur place avec crayon et papier. Parmi les autres avantages de l'IAO figurent les suivants :

- utilisation possible de données dépendantes, les données recueillies dans les vagues précédentes étant intégrées dans l'interview de la vague courante;
- possibilité de procéder à des vérifications des enregistrements unitaires et de l'uniformité pendant l'interview, afin de répondre aux interrogations et de donner suite aux incohérences en présence du répondant;
- collecte et récupération facilitées des données sur les répondants pour la gestion de la réponse.

L'élaboration, la mise à l'essai et la programmation du questionnaire représentent un processus à forte intensité de ressources. La SoFIE a été élaborée en vue d'intégrer des modules du questionnaire des personnes (ou sections), dont les questions ne seront posées qu'à des vagues particulières. Le module sur la « santé » comprend une section de questions liées à la santé (état de santé autoévalué, incidence des problèmes de santé chroniques, événements majeurs liés à la santé, facteurs de risque et utilisation des services de santé), qui sont posées uniquement aux interviews des vagues trois, cinq et sept. Les questions sur le type et la valeur des actifs et des passifs, quant à elles, sont posées dans un module différent des interviews des vagues deux, quatre, six et huit.

Même si des modules ont été intégrés à la SoFIE, les changements apportés au questionnaire entre les vagues (après le développement initial) ont été limités.

2.2 Expériences des entreprises

L'utilisation des données du système fiscal de la Nouvelle-Zélande représente une composante clé des LEED et de la LBD. L'utilisation des données fiscales par Statistics New Zealand est autorisée en vertu de la *Tax Administration Act*, à condition de respecter certaines conditions. Par exemple, l'utilisation des données ne doit pas avoir de répercussions sur l'intégrité du système fiscal. Parmi les défis initiaux liés aux deux collectes figurait la démonstration que toutes les exigences en matière

législative, de sécurité et de confidentialité liées à l'utilisation des données fiscales et à l'intégration des différentes sources de données pouvaient être respectées.

Les données fiscales sont recueillies par Inland Revenue, aux fins de l'administration du système fiscal de la Nouvelle-Zélande. L'utilisation de ces données par Statistics New Zealand pour les LEED et la LBD vise la production de statistiques, c'est-à-dire un objectif différent de celui de la collecte. Par conséquent, des processus d'épuration, de transformation et d'intégration sont nécessaires pour produire les microdonnées requises. Parmi les défis liés à l'application de ces processus figurent les suivants :

- différence entre les unités de collecte et les unités statistiques;
- modifications des données administratives, par suite des changements dans les structures des entreprises, comme les fusions ou les divisions;
- numéros d'enregistrement fiscaux manquants ou mal codés pour les employés ou les entreprises;
- méthodologie utilisée pour coupler les différentes sources de données.

La LBD intègre en outre d'autres données administratives (non fiscales) et des données d'enquête. Un élément clé des travaux d'élaboration de la LBD avait trait à l'établissement de couplages, d'unités et d'échéances communs pour toutes les données de source. Par exemple, l'une des enquêtes auprès des entreprises visées est l'Enquête annuelle auprès des entreprises (AES), dont les unités de collecte sont fondées sur les unités de type d'activité (KAU). L'unité commune utilisée dans la LBD correspond à l'entreprise. Les unités de collecte de l'AES doivent être agrégées au niveau de l'entreprise avant de pouvoir être couplées à la LBD.

L'accès aux microdonnées LEED a constitué la toute première expérience de Statistics New Zealand en ce qui a trait à l'accès par des utilisateurs externes aux microdonnées sur les entreprises (longitudinales et transversales). L'accès aux microdonnées dans le cas des données longitudinales sur les entreprises a représenté un défi plus grand que pour les données sociales longitudinales, en raison de la constitution proprement dite des données sur les entreprises. Par exemple, la Nouvelle-Zélande compte très peu de grandes entreprises et, dans certaines industries, comme les télécommunications ou le transport, ces grandes entreprises présentent un risque élevé d'être identifiées grâce aux microdonnées. Des mesures additionnelles de sécurité et de confidentialité sont utilisées à la fois pour les LEED et la LBD, afin de tenir compte de ce risque. Par exemple, des restrictions sont imposées à l'emplacement physique des microdonnées, aux personnes qui y ont accès et à la façon dont les produits des microdonnées sont diffusés à l'extérieur.

2.3 Accès aux données et confidentialité

Parmi les défis clés liés à l'accès aux données auxquels font face les organismes officiels de statistiques figure le maintien de la sécurité (protection de la confidentialité des répondants) et de l'utilité des microdonnées fournies aux utilisateurs externes. Statistics New Zealand offre aux utilisateurs externes plusieurs façons différentes d'accéder aux microdonnées, selon leurs besoins et les exigences des données. Les techniques de confidentialité appliquées aux microdonnées pour en assurer la protection dépendent du niveau de restriction imposé à l'accès aux microdonnées. En général, plus l'accès est public, plus le contrôle de la divulgation appliquée est grand.

Il est possible de produire un fichier de microdonnées, à partir d'une enquête transversale, qui comporte des niveaux appropriés de sécurité et d'utilité en vue de sa diffusion publique. Le risque d'identification des répondants à partir des microdonnées longitudinales est beaucoup plus grand que dans le cas des données transversales, en raison simplement du caractère longitudinal. Par exemple, les données concernant les transitions au fil du temps peuvent facilement permettre d'identifier les répondants. L'obtention d'un niveau approprié de sécurité dans les microdonnées longitudinales destinées à être diffusées dans le public rendrait probablement le niveau d'utilité beaucoup trop faible pour répondre à l'un ou l'autre des besoins des utilisateurs.

L'approche de Statistics New Zealand en ce qui a trait à l'accès aux microdonnées longitudinales consiste à les rendre disponibles dans l'environnement des laboratoires de données seulement. Les laboratoires de données sont installés dans nos bureaux, dans trois villes différentes et représentent des environnements contrôlés destinés à des utilisateurs et des projets de recherche approuvés au préalable. Tous les produits font l'objet d'une vérification au chapitre de la confidentialité avant d'être diffusés à l'extérieur des laboratoires de données. Le délai nécessaire pour l'approbation d'un projet, l'établissement initial dans le laboratoire de données, ainsi que la vérification des produits et leur diffusion, est souvent perçu par les utilisateurs comme le plus grand obstacle à l'accès. Des réflexions et un examen de nos processus, objectifs et techniques de vérification de la sécurité sont en cours, afin de déterminer les aspects pour lesquels l'actualité peut être améliorée, tout en continuant d'assurer la sécurité des microdonnées.

L'environnement des laboratoires de données et l'accès aux microdonnées en général constituaient aussi une nouvelle expérience pour Statistics New Zealand par rapport à la décennie précédente. Une part importante des raisons motivant l'accès aux microdonnées (ainsi que les procédures et les lignes directrices relatives à cet accès) ont été le résultat de l'intervention de nombreux utilisateurs externes des données longitudinales de Statistics New Zealand. Il s'agit notamment de la diffusion demandée des chiffres non pondérés par les chercheurs de la SoFIE dans les laboratoires de données. Les discussions qui en ont résulté et l'examen de cette demande a entraîné une modification de la pratique de confidentialité de Statistics New Zealand. En mars 2008, une nouvelle politique a été adoptée, qui permet pour la première fois la diffusion et la publication, selon certaines conditions, de chiffres non pondérés des enquêtes sociales.

2.4 Ressources

Les ressources sont limitées et les organisations doivent faire face au défi permanent de décider où et comment ces ressources sont affectées ou dépensées.

Lorsque l'on réfléchit rétroactivement aux premières étapes de la SoFIE et que l'on tient compte de la somme réelle de ressources utilisées jusqu'à maintenant, on constate que la somme des ressources requises initialement et de façon permanente a été sous-estimée. La collecte de données longitudinales nécessite un investissement de ressources important à long terme. Toutefois, ce n'est pas seulement l'importance des ressources disponibles qui compte, mais aussi la façon dont elles sont affectées et utilisées dans les différents secteurs ou aux diverses étapes de la collecte. Le remaniement actuel du système de traitement de la SoFIE a fait ressortir l'importance d'un investissement initial de ressources approprié. La correction ou le remaniement des systèmes à une étape ultérieure entraînera l'utilisation de plus de ressources au total.

Les personnes sont l'une des ressources les plus importantes. L'établissement et le maintien d'une expertise et de connaissances longitudinales chez les employés représentent un défi permanent, particulièrement du fait de la nature à long terme et de la complexité de la collecte de données longitudinales par rapport à la collecte des données transversales. Au cours de la dernière décennie, Statistics New Zealand a investi considérablement dans la mise en valeur des capacités des employés pour la conception, la collecte sur le terrain et le traitement des collectes longitudinales. Par le passé, on n'a pas mis autant d'accent sur l'analyse ou l'utilisation proprement dite de nos propres données longitudinales. À l'avenir, il faudra mettre constamment l'accent sur une meilleure utilisation de nos données, ainsi que sur une aide améliorée aux tiers qui utilisent nos données.

3. Conclusion

Pour Statistics New Zealand, les aspects longitudinaux de la SoFIE, de la LisNZ, des LEED et de la LBD ont fourni aux utilisateurs de données la capacité d'examiner et d'explorer les changements qui touchent les entreprises ou les personnes en Nouvelle-Zélande au fil du temps. Cela n'aurait simplement pas été possible avec des collectes de données transversales. L'expérience acquise dans le cadre de ces collectes a fourni à Statistics New Zealand la possibilité de développer des connaissances et une expertise longitudinales sur lesquelles on pourra compter à l'avenir.

SESSION 15 – REMANIEMENT D'ENQUÊTES LONGITUDINALES DE GRANDE ENVERGURE

La continuité et l'innovation dans la conception de Understanding Society: The UK Household Longitudinal Study

Dr Heather Laurie[1]

Résumé

Dans le présent article, nous traitons du processus décisionnel ayant permis d'intégrer la British Household Panel Survey (BHPS), une enquête par panel à long terme auprès des ménages au Royaume-Uni, à une nouvelle enquête élargie, *Understanding Society*: the UK Household Longitudinal Study. Nous mettons la nouvelle enquête en contexte et nous en énonçons les caractéristiques clés et les possibilités d'analyse, le plan de sondage et les décisions prises pour y intégrer l'échantillon de la BHPS, et pour assurer la continuité et l'innovation dans le contenu du questionnaire.

1. Introduction

Les concepteurs d'enquêtes longitudinales font face à des problèmes d'adaptation au changement, surtout lorsque le plan de sondage a été plus ou moins constant dans le contexte d'une étude menée à long terme. Dans le présent article, nous étudions la BHPS (British Household Panel Study) et sa transition à une enquête nouvelle et élargie, *Understanding Society*: the UK Household Longitudinal Study. La BHPS, une enquête par panel menée auprès d'environ 8 000 ménages du Royaume-Uni, s'est déclinée sur 18 cycles annuels de collecte des données. Elle constitue la principale source de données de panel de ménages au Royaume-Uni depuis 1991 et elle sert largement aux besoins de la recherche universitaire et stratégique. À l'issue d'une vaste consultation des utilisateurs, on a passé un contrat avec l'équipe de recherche chargée de la BHPS, qui relève de l'Institute for Social and Economic Research (ISER), University of Essex, en vue de concevoir une enquête élargie par panel auprès de 40 000 ménages, appelée *Understanding Society*: the UK Household Longitudinal Study. Le cycle 1 de l'enquête a débuté en janvier 2009 et le cycle 2, en janvier 2010. Le plan de sondage de l'enquête permettra d'intégrer les participants de l'échantillon de la BHPS l'enquête à partir du cycle 2. Dans le présent article, nous décrivons la conception d'*Understanding Society* et nous expliquons comment l'expérience tirée de la BHPS a éclairé le processus décisionnel ayant permis de concevoir l'enquête et de la mettre en œuvre. Nous abordons les facteurs qui ont motivé la décision d'y intégrer la BHPS, soit la nécessité de concilier les diverses exigences de continuité et d'innovation, l'échéancier des interviews au cours de l'année, le contenu du questionnaire et la fidélisation du panel, malgré le recours à une autre organisation pour les travaux sur le terrain.

1.1 Contexte d'*Understanding Society*

Understanding Society est une nouvelle enquête par panel auprès des ménages financée par l'Economic and Social Research Council du Royaume-Uni, de concert avec les ministères publics. Elle s'inscrit dans le droit fil d'autres enquêtes longitudinales qui ont été menées à bien au Royaume-Uni et qui forment un portefeuille riche et diversifié où figurent non seulement la BHPS mais aussi les enquêtes de la British Birth Cohort (National Child Development Survey, British Cohort Study 1970, Millennium Cohort Study), les enquêtes sur le vieillissement, comme l'English Longitudinal Study of Ageing (ELSA), les enquêtes sur les jeunes, comme la Longitudinal Study of Young People in England (LSYPE) et les enquêtes longitudinales du recensement. La BHPS représente un ensemble de données largement utilisé au Royaume-Uni, mais du fait de la maturité du panel, on a jugé nécessaire d'examiner l'avenir des sources de données longitudinales pour les prochaines décennies au Royaume-Uni. Pour amener la BHPS dans une ère nouvelle, il était important d'assurer la continuité avec le plan de sondage et les données qui existaient déjà. Sans compter que de nouvelles demandes allaient changer inévitablement le modèle de la BHPS et nécessiter un examen soigné. L'un des éléments clés de la nouvelle enquête était la taille de l'échantillon, puisqu'il était admis qu'en l'élargissant par rapport à celui de la BHPS, on permettrait d'étendre le champ des analyses et d'affiner

[1] Dr Heather Laurie, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK. laurh@essex.ac.uk

l'examen des sous-groupes de la population. L'échantillon total d'*Understanding Society* a été fixé à 40 000 ménages, y compris l'échantillon des ménages établi pour la BHPS.

1.2 Principales caractéristiques et possibilités d'analyse

Le plan de sondage d'*Understanding Society* se détaille en huit particularités qui reflètent sa raison d'être et qui peuvent servir à produire des innovations majeures en recherche scientifique. Ces particularités permettent de pousser l'analyse au-delà du champ de la BHPS. En voici le détail.

i) La **large taille de l'échantillon,** établie à 40 000 ménages, permet d'explorer des questions qui échappent aux enquêtes longitudinales de moindre portée. Les analyses des petits sous-groupes, comme les parents adolescents, les personnes handicapées ou les jeunes en chômage, suscitent des préoccupations clés en matière de politiques, mais lorsqu'ils sont désagrégés selon d'autres caractéristiques, ils risquent de ne plus résister à une analyse robuste. On peut aussi les analyser pour observer aux niveaux régional et infrarégional les effets de la variation géographique des politiques, devient aussi possible. La large taille de l'échantillon permet d'analyser finement les événements au fil du temps, en s'intéressant par exemple aux cohortes d'une année d'âge. Des événements qui sont relativement rares dans la population, comme le divorce, seront plus nombreux pendant un cycle donné, ce qui permet de renforcer l'analyse sans avoir à attendre que les données d'un certain nombre de cycles s'accumulent pour disposer d'un nombre suffisant d'événements pour l'analyse.

ii) L'**importance accordée aux ménages dans le plan de sondage** signifie que des données seront recueillies sur tous les membres des ménages échantillonnés et sur leurs interactions à l'intérieur du ménage. Elle comporte des avantages non négligeables pour les domaines clés de la recherche, comme la consommation et le revenu, le partage des ressources dans le ménage étant important en l'occurrence, ou encore l'évolution démographique, lorsque le ménage proprement dit est souvent l'objet de l'enquête. Par rapport aux cohortes de naissance distinctes, elle permet d'améliorer la qualité et la continuité des données sur le milieu familial et le ménage où se produit le développement de la petite enfance. La capacité d'observer plusieurs générations et toutes les fratries permet d'examiner les processus de transmission à long terme et d'isoler les effets des caractéristiques communes liées aux antécédents familiaux. Par rapport aux enquêtes existantes, *Understanding Society* élargit les possibilités d'explorer les liens à l'extérieur du ménage.

iii) Un **échantillon comportant toute la fourchette d'âge** vient compléter les études par cohorte axées sur l'âge au Royaume-Uni, comme celles reposant sur des échantillons de personnes âgées ou de jeunes, et fournit un aperçu unique des comportements et des transitions au milieu de la vie. Ainsi, même pour les questions de pensions et de soins de longue durée, qui sont habituellement associées à la vieillesse, l'établissement des politiques est fortement influencé par les comportements tôt dans la vie des gens, lorsque ceux-ci prennent des décisions au sujet de l'épargne, des pensions et des investissements. Par ailleurs, la large taille de l'échantillon signifie que toutes les cohortes peuvent être analysées au même moment.

iv) L'inclusion d'un **panel de l'innovation pour la recherche et les essais méthodologiques** sert d'outil pour une vaste gamme de travaux méthodologiques visant non seulement à éclairer l'élaboration de l'enquête, mais aussi à contribuer aux progrès méthodologiques de l'enquête en général.

v) Le **plan de sondage comportant plusieurs sujets** vise à répondre à un éventail de besoins disciplinaires et interdisciplinaires en matière de recherche. *Understanding Society* explore plusieurs sujets et s'intéresse à la diversité des domaines liés à la vie. L'enquête répond à la fois aux besoins des disciplines quantitatives « traditionnelles » des sciences sociales, comme l'économie, la sociologie et la politique sociale, et à ceux d'autres disciplines (p. ex., les sciences biomédicales) et elle offre un jeu élargi d'approches méthodologiques. Elle facilitera la recherche interdisciplinaire véritable : dans le domaine des sciences sociales (p. ex., géographie et économie); dans celui des sciences biomédicales (p. ex., psychologie et génétique); et entre les deux.

vi) L'intégration d'un volet **ethnicité** vise à reconnaître l'importance de la recherche sur les différences ethniques pour permettre de saisir la constitution de la société britannique et les questions de diversité et de communauté. Elle fait aussi ressortir la possibilité que cette recherche transforme notre compréhension des processus individuels et sociaux. Du fait qu'il couvre un grand nombre de membres de groupes minoritaires au fil du temps et de domaines de recherche liés à l'ethnicité, le volet ethnicité permettra de réaliser des progrès essentiels dans la connaissance de ces sujets et des processus concernés.

vii) L'obtention de données sur un éventail de biomarqueurs et d'indicateurs de la santé ouvre des perspectives de progrès intéressants, à la limite des sciences sociales et de la recherche biomédicale. On pourra ainsi évaluer les facteurs d'exposition et d'antécédents de l'état de santé; comprendre les mécanismes de la maladie (p. ex., l'interaction entre les gènes et l'environnement et les liens entre les gènes et les fonctions), les effets sur les ménages et les effets socioéconomiques; et analyser les résultats, à partir d'évaluations directes ou du couplage des données.

viii) Il est prévu un couplage exhaustif des données et des dossiers administratifs ainsi que des données géocodées. À cette fin, on devra obtenir des participants qu'ils autorisent que leurs données soient couplées à celles d'autres domaines

comme la santé, l'éducation, les pensions et les prestations publiques. Les résultats fourniront de nouveaux thèmes d'analyse lorsque l'abondance des données d'une enquête sociale contextuelle peut se conjuguer aux données administratives.

2. Plan de sondage

2.1 Plan d'échantillonnage par panel des ménages

Le plan d'*Understanding Society* s'aligne sur celui de la BHPS et d'autres panels nationaux de ménages. Il comporte un échantillon longitudinal de personnes représentant l'ensemble de la population du Royaume-Uni, qui ont été interviewées à l'intérieur d'un ménage. L'échantillon est constitué d'adresses résidentielles privées tirées du fichier des codes postaux d'adresses (PAF) au Royaume-Uni, et tous les membres des ménages privés se trouvant à ces adresses au cycle 1 sont désignés comme membres de l'échantillon originel. L'échantillon est un échantillon avec probabilités égales comptant au total 2 640 unités primaires d'échantillonnage (l'UPE est un secteur de code postal au Royaume-Uni) et 18 adresses par UPE. En Irlande du Nord, l'échantillon est un échantillon aléatoire simple. Chaque échantillon mensuel comprend 110 UPE, et chaque mois produit des données représentatives pour le Royaume-Uni, avec comme objectif principal de permettre aux analystes de produire des estimations trimestrielles pour le Royaume-Uni. La taille de l'échantillon présuppose un taux de réponse de 60 % au minimum, une fois exclues les adresses non admissibles déterminées pendant les travaux sur le terrain. À chaque cycle, tous les membres de l'échantillon de 10 ans et plus peuvent être interviewés, et les personnes sont suivies à mesure qu'elles déménagent et qu'elles forment de nouveaux ménages. D'autres personnes qui forment des ménages avec des membres de l'échantillon après le cycle 1 deviennent admissibles aux interviews, à condition qu'elles résident avec un membre de l'échantillon originel. Comme pour la BHPS, les règles qui suivent signifient que l'échantillon demeurera représentatif de la population du Royaume-Uni en 2009, au gré de son évolution, sous réserve de la pondération et en excluant les nouveaux immigrants. L'inclusion d'un nouvel échantillon d'immigrants est prévue pour l'avenir (voir http://www.iser.essex.ac.uk/survey/bhps pour des détails sur la conception et le contenu de la BHPS et http://www.understandingsociety.org.uk pour des détails sur *Understanding Society*).

Outre les spécifications de base du plan d'échantillonnage par panel des ménages, il existe certaines différences clés entre les plans de sondage d'*Understanding Society* et de la BHPS. La première a trait à l'inclusion d'un panel de l'innovation, un échantillon longitudinal de 1 500 ménages auprès de qui la collecte commence une année avant celle de l'échantillon principal et qui sert pour les essais et l'expérimentation méthodologiques. Comme dans le cas de l'échantillon principal, il s'agit d'un échantillonnage avec probabilités égales de la Grande-Bretagne (dans ce cas, au sud du canal calédonien), à l'exclusion de l'Irlande du Nord, comportant 120 UPE et 23 adresses par UPE. Le panel de l'innovation sert à des expériences aléatoires des procédures sur le terrain, comme la fourniture d'incitatifs et de matériel aux répondants, des expériences de questionnaires à deux formes concernant l'utilisation de cartes, d'autres libellés, le recours aux interviews dépendantes dans les mesures du changement, ainsi que des essais de sensibilité pour certains éléments (voir Laurie, H., Burton, J. et Uhrig, S.C. Noah (sous la dir.) (2008) pour les résultats préliminaires du cycle 1 du panel de l'innovation).

En deuxième lieu, le plan de sondage comprend un suréchantillonnage des minorités ethniques dans les régions de densité moyenne à élevée de minorités ethniques qui, selon les données du Recensement du Royaume-Uni de 2001, devraient englober de 80 % à 90 % des populations de minorités ethniques d'intérêt au Royaume-Uni. Le suréchantillonnage est conçu pour tenir 1 000 interviews additionnelles dans chacun des cinq groupes clés : Indiens, Pakistanais, Bangladais, Caribéens et Noirs d'Afrique, les ménages admissibles étant identifiés au moyen d'une enquête de dépistage. Outre tous les répondants d'origine mixte identifiés dans l'interview de dépistage, y compris les Chinois, d'autres répondants d'Asie et du Moyen-Orient peuvent être inclus. Le suréchantillonnage permet aussi d'intégrer un échantillon de comparaison, tiré de l'échantillon de la population générale, qui est constitué d'environ 500 ménages. Les tailles d'échantillons estimées pour les diverses composantes d'*Understanding Society* figurent dans le tableau 2.1-1 ci-après.

Tableau 2.1-1

Tailles d'échantillon estimées (ménages), cycles 1 et 2

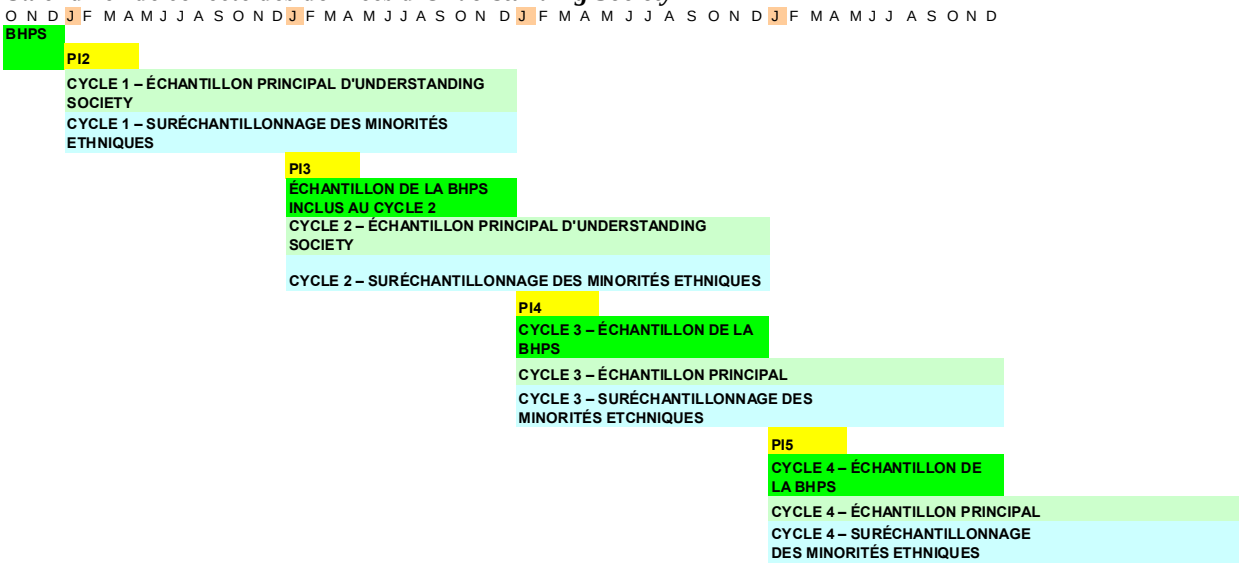
	Nouvel échantillon*	Échantillon de la BHPS**	Suréchantillonnage des ME	Total
Cycle 1	29 850	8 100	4 220	42 170
Cycle 2	25 370	6 900	3 590	35 860

* Comprend le panel de l'innovation de 1 500 ménages et l'échantillon de comparaison de 500 ménages pour le suréchantillonnage.
** À noter que le cycle 18 de la BHPS est considéré dans ce cas comme faisant partie du cycle 1 d'*Understanding Society*.

2.2 Calendrier de collecte des données

La BHPS est une enquête annuelle et, pendant 18 ans, les travaux sur le terrain ont commencé le 1er septembre et se sont échelonnés principalement jusqu'en décembre de la même année civile. Cela a été suivi par une queue relativement longue de trois autres mois pour traiter des cas qui étaient difficiles à contacter ou absents pendant la période de travaux sur le terrain, dépister les personnes qui avaient déménagé et convertir les refus. Le plan de sondage d'*Understanding Society* comporte aussi une interview annuelle à des intervalles de 12 mois, mais compte tenu de la taille importante de l'échantillon et des travaux sur le terrain nécessaires pour mener les interviews, chaque cycle comporte un effort continu de 24 mois. Les répondants conservent leur mois d'échantillonnage du cycle 1 tout au long de l'étude, même s'ils changent d'adresse entre les cycles. Cela fait en sorte que des cycles se chevauchent, l'année 2 du cycle 1 chevauchant l'année 1 du cycle 2. Des échantillons mensuels pour *Understanding Society* ont commencé en janvier 2009, et l'échantillon de la BHPS est intégré à partir du cycle 2 de l'étude en 2010. Le tableau 2.2-1 fait état du calendrier de collecte des données pour chaque cycle.

Tableau 2.2-1
Calendrier de collecte des données d'*Understanding Society*



Parmi les autres décisions difficiles à prendre au sujet de l'échantillon de la BHPS figurait le maintien du calendrier de collecte des données du cycle 1 au cycle 18, ou encore l'intégration complète de l'échantillon de la BHPS dans le calendrier des travaux sur le terrain d'*Understanding Society* et sa répartition entre les échantillons mensuels. Des arguments ont été fournis pour et contre le maintien de la collecte des données de la BHPS dans la même période de travaux sur le terrain que pour les cycles précédents. Les répondants s'attendaient à être interviewés entre septembre et décembre chaque année et étaient généralement prêts à ce que l'intervieweur les appelle à ce moment-là. Dans le cas des analystes, le maintien de la même période de collecte de données devait permettre d'assurer l'uniformité du point de vue des effets saisonniers et de la continuité des analyses effectuées à partir d'un certain nombre de cycles de données. Les arguments en faveur de l'intégration de l'échantillon de la

BHPS dans le calendrier de collecte d'*Understanding Society* avaient principalement à voir avec le désir d'intégrer pleinement la BHPS dans la nouvelle étude pour l'avenir. Par ailleurs, on s'attendait à ce que le maintien de la collecte des données de l'échantillon de la BHPS à la même période que pour les cycles précédents produise une distribution inégale et difficile à gérer des travaux sur le terrain.

Finalement, on a décidé que l'intégration de la BHPS dans le plan d'échantillonnage mensuel améliorerait les possibilités d'utilisation de cette dernière de concert avec le nouvel échantillon, plutôt qu'elle soit considérée comme une sous-étude distincte à l'intérieur d'*Understanding Society*. Du point de vue de la stratégie à plus long terme de l'étude globalement, l'intégration de la BHPS dans le plan d'échantillonnage mensuel semblait l'approche la plus cohérente. En dépit de cette décision, des préoccupations ont été soulevées concernant l'intervalle plus long que la normale entre les interviews pour l'échantillon de la BHPS. Les interviews du cycle 18 se sont tenues à la fin de 2008 et au début de 2009, ce qui fait que l'échantillon de la BHPS n'a pas pu être intégré au cycle 1 de la nouvelle étude à partir de janvier 2009. Le cycle 2 d'*Understanding Society* a commencé en janvier 2010, et si l'échantillon de la BHPS était réparti également entre les deux périodes d'année civile pour le cycle 2, l'intervalle entre les interviews pourrait aller de 36 à 38 mois. Cela a été jugé inacceptable, ce qui fait qu'on a décidé de répartir de façon aléatoire l'échantillon de la BHPS sur la première période de 12 mois des travaux sur le terrain du cycle 2, avec un intervalle maximum de deux ans entre les interviews, 50 % de l'échantillon étant interviewé de 18 à 20 mois après l'interview précédente du cycle 18. Cela signifie que l'échantillon de la BHPS sera inclus dans la première année de collecte des données de chaque cycle à l'avenir.

2.3 Mode de collecte des données

Depuis 1999, la BHPS a toujours été menée au moyen d'interviews sur place assistées par ordinateur (IPAO) dans les foyers des répondants. Dans le cadre d'*Understanding Society*, on a l'intention de passer à des modes mixtes de collecte des données pour certains cycles, en partie en raison des coûts des travaux sur le terrain, mais aussi pour fournir aux répondants un plus grand choix de modes de réponse. Les cycles 1 à 3 d'*Understanding Society* sont menés sur place, au moyen de l'IPAO, l'établissement du panel dès les premiers cycles, la maximisation des taux de réponse et la réduction de l'attrition dans ces cycles étant reconnus comme essentiels pour assurer la qualité et la santé à long terme de l'étude. Il est probable que pour le cycle 4 ou le cycle 5 de l'étude, on aura recours à une stratégie mixte de collecte de données, et le panel de l'innovation est utilisé pour mettre à l'essai différentes stratégies de mode séquentiel, afin de déterminer laquelle est la plus efficace du point de vue des taux de réponse globaux et du déroulement des travaux sur le terrain. Le cycle 2 du panel de l'innovation (mené d'avril à juin 2009) a permis de mettre à l'essai des modes séquentiels mixtes utilisant à la fois les interviews téléphoniques et les interviews sur place, et nous prévoyons inclure des interviews par Internet à la dernière étape des travaux de développement. L'analyse de ces données, y compris des paradonnées exhaustives sur les modèles d'appel, est en cours. Le panel de l'innovation représente par conséquent une ressource, non seulement pour éclairer le développement de l'étude principale, mais aussi pour mener des recherches méthodologiques innovatrices présentant un intérêt plus large pour les praticiens d'enquête ou les méthodologistes.

2.4 Changements touchant les organismes chargés des travaux sur le terrain

Pour l'échantillon de la BHPS, le premier contact avec *Understanding Society* prendra la forme d'une interview sur place au cycle 2 (2010). Pendant 18 ans, la BHPS a été menée par GfKNOP, un organisme d'enquête établi à Londres, dont l'équipe de l'enquête de l'ISER a retenu les services pour mener les interviews et pour fournir les données. Mais par suite de la nouvelle étude, le contrat de collecte des données pour les cycles 1 et 2 a été attribué à un nouveau fournisseur, le National Centre for Social Research (NatCen). Ce changement a suscité certains problèmes d'intégration pour l'échantillon de la BHPS, dont nombre des membres ont été interviewés par le même intervieweur pendant plusieurs ou tous les cycles de la BHPS. Nous estimons que ce changement de responsabilité des travaux sur le terrain est susceptible d'affecter les taux de réponse de l'échantillon de la BHPS. Même si les participants ont été loyaux à la BHPS au fil des ans, ce changement pourrait en inciter certains à décider de se retirer de l'étude, ce qui n'est évidemment pas souhaitable dans la mesure du possible. Outre les modifications concernant les intervieweurs et la responsabilité des travaux sur le terrain, les répondants de la BHPS ont dû se familiariser avec un nouveau nom, logo et format.

On a élaboré des procédures spéciales, en vue de présenter aux répondants de la BHPS la nouvelle enquête et les nouveaux intervieweurs. Cela comprenait une série d'envois par la poste entre l'interview du cycle 18 de la BHPS et le premier contact du cycle 2, pour les informer des changements apportés à l'étude et pour les inciter à les aborder favorablement. On a souligné leur importance comme membres fondateurs de l'étude, on leur a communiqué le nouveau nom et logo et de la documentation, on leur a envoyé des documents de promotion additionnels, on leur a fourni des détails concernant NatCen et on les a informés de leur nouveau mois d'interview. Avant la prise de contact, les intervieweurs leur ont écrit en se servant des renseignements de l'échantillon de la BHPS et leur ont fourni leur nom et numéro de téléphone. Comme cela est courant dans la BHPS, on a prévu une ligne téléphonique réservée, une adresse de courriel et un site Internet, afin que les participants puissent

communiquer directement avec l'équipe de l'enquête s'ils avaient des préoccupations. Comme le cycle 2 vient de commencer, il nous manque des données sur l'avancement de la transition à la nouvelle enquête pour l'échantillon de la BHPS, mais nous suivrons l'évolution de près de manière à intervenir s'il le faut pendant les travaux sur le terrain.

3. Conception et contenu du questionnaire

3.1 Continuité et innovation dans le contenu

La structure du questionnaire d'*Understanding Society* est similaire à celle de la BHPS et comprend :

- le questionnaire de 15 minutes à l'intention des ménages (y compris une liste des membres du ménage);
- des interviews individuelles de tous les adultes de 16 ans et plus (32 minutes);
- un questionnaire à remplir soi-même pour tous les adultes interviewés (7 minutes);
- des interviews par personne interposée pour les adultes de 16 ans et plus incapables d'être interviewés (10 minutes);
- un questionnaire à remplir soi-même destiné aux enfants de 10 à 15 ans (dans la BHPS, le questionnaire destiné aux jeunes est distribué à ceux de 11 à 15 ans).

Tous les questionnaires du script de l'IPAO et les documents connexes pour les travaux sur le terrain qui sont destinés aux répondants sont traduits en neuf langues y compris le gallois, et des intervieweurs bilingues ou interprètes sont fournis au besoin.

Même si la structure des interviews utilise les mêmes instruments que le questionnaire de la BHPS, il existe une forte demande de la part des utilisateurs pour l'inclusion dans *Understanding Society* d'un contenu nouveau et élargi par rapport à la BHPS. L'échantillon de la BHPS recevra le nouveau questionnaire du cycle 2, et il est important de maintenir une certaine comparabilité avec la BHPS, afin de réduire les ruptures des séries chronologiques à long terme de mesures clés. Par conséquent, on utilise dans une large mesure les questions de la BHPS dans *Understanding Society*. L'un des principes qui a guidé les décisions concernant des questions particulières était qu'en l'absence d'une question de rechange permettant de produire des données de meilleure qualité pour une question existante de la BHPS, on utiliserait cette question sans modification. Toutefois, lorsqu'il était clairement justifié du point de vue de l'amélioration de la qualité des données de modifier ou d'améliorer une question de la BHPS, nous avons inclus des questions nouvelles, révisées ou de rechange.

Au nombre des domaines ayant obligé à créer ou à élargir le contenu par rapport à la BHPS figurent les suivants :

- Famille et réseaux sociaux à l'extérieur du ménage
- Attitudes et comportements liés aux questions environnementales
- Précisions sur les comportements illicites et risqués, surtout chez les jeunes
- Précisions sur l'attachement au quartier et l'engagement social
- Activités culturelles et emploi du temps pendant les loisirs
- Précisions sur les attitudes, le comportement et les styles parentaux
- Attributs psychologiques
- Mesures des capacités/fonctions cognitives
- Précisions sur les résultats en matière de santé et les comportements influant sur la santé
- Qualité du sommeil
- Bien-être
- Qualité des relations conjugales
- Risque et confiance
- Collecte de données auprès des parents sur les enfants de moins de 10 ans
- Précisions sur le passage à l'âge adulte
- Discrimination et harcèlement
- Identité ethnique et nationale

Outre l'importance de ces domaines, l'enquête doit recueillir des données clés sur les dimensions sociales, économiques et de l'emploi qui sous-tendent le contenu longitudinal. L'un des principaux problèmes auxquels on a fait face était la durée réduite de l'interview dans *Understanding Society*, celle-ci étant significativement plus courte que pour la BHPS (32,5 minutes contre 40 minutes). Cette contrainte temporelle a fait qu'une assez grande proportion de questions de la BHPS sont carrément omises de la nouvelle enquête ou vont y être posées moins souvent que chaque année.

3.2 Plan de sondage modulaire annuel et avec renouvellement de panel

La solution aux contraintes temporelles a consisté à laisser de côté le modèle de la BHPS, dans lequel environ 80 % des questionnaires concernaient des mesures annuelles répétées et 20 % un contenu thématique renouvelé, toutes les questions pouvant être posées à la majorité des personnes. Pour *Understanding Society*, on a adopté un modèle dans lequel on utilise davantage : i) des questions posées régulièrement, mais pas à chaque cycle; ii) des questions posées uniquement après des événements clés, comme la naissance d'un enfant ou un déménagement; iii) des questions posées uniquement à des âges particuliers; et iv) des sous-échantillons fondés sur des caractéristiques démographiques. Contrairement à la BHPS, environ 50 % du questionnaire individuel est consacré aux mesures annuelles répétées, et le reste, aux modules thématiques renouvelés, qui sont menés tous les deux ou trois ans selon le sujet et les taux de changement prévus. Les mesures annuelles répétées sont dans une large mesure des questions reprises de la BHPS, qui permettront la poursuite de l'analyse de séries chronologiques plus longues pour cette portion de l'échantillon, ainsi que la continuité avec le nouvel échantillon d'*Understanding Society*, au gré de l'accumulation des données longitudinales dans les prochains cycles.

Une des tâches clés des étapes initiales a consisté à déterminer les mesures essentielles pour le plan de sondage annuel par panel et les questions à poser chaque année. Les mesures annuelles répétées comprennent des caractéristiques démographiques de base; les changements entre les cycles – emploi, scolarité, formation, fécondité, partenaire, mobilité géographique, problèmes de santé, état de santé (p. ex. SF12), incapacité; activité sur le marché du travail et situation d'emploi, recherche d'emploi; caractéristiques de l'emploi actuel, conditions d'emploi de base, heures de travail rémunérées, deuxième emploi; soins des enfants, autres soins à l'intérieur et à l'extérieur du ménage; développement de l'enfant à des âges clés, revenu et gains; satisfaction à l'égard de la vie; affiliation politique; accès au transport et aux communications; scolarité; planification de la retraite/attentes à des âges clés; dépenses de consommation; caractéristiques du logement; dépenses de logement; installations du ménage, propriété d'une voiture. En outre, le questionnaire du cycle 1 comprenait des mesures contextuelles factuelles, comme le lieu de naissance, le niveau d'études et les détails sur les origines des parents, ainsi que des données sur les événements de la vie, y compris la cohabitation, le mariage et la fécondité et un historique de la migration pour le suréchantillon des minorités ethniques.

3.3 Collecte de biomesures

La collecte de mesures physiques directes et d'échantillons devrait commencer au cycle 2 d'*Understanding Society* et se poursuivre jusqu'à la fin du cycle 3. Même si le financement disponible ne permet pas la collecte auprès de tous les membres de l'échantillon, nous estimons qu'entre 25 000 et 30 000 cas feront l'objet de ces mesures, ce qui représente une forte proportion de l'échantillon total. Le but visé est de recueillir ces mesures pour tout l'échantillon de la BHPS, afin que les données de l'enquête longitudinale puissent être utilisées de concert avec les données sur les biomesures et produire des résultats longitudinaux précoces.

Deux types de biomesures sont recueillis. L'un a trait aux mesures anthropométriques directes et l'autre, à des échantillons devant être entreposés pour une analyse ultérieure. Les chercheurs dans le domaine de la santé intègrent de plus en plus des mesures sommaires tirées de ces biomarqueurs dans l'analyse du cheminement vers la maladie et dans l'exploration de l'émergence de disparités en matière de santé (p. ex., Chandola, Brunner et Marmot 2006; Seeman et coll., 2004; Sabbah et coll., 2008). La collecte des données pour cet élément de l'étude se fera par des infirmières-intervieweuses ou par des intervieweurs d'enquête formés. Les mesures à obtenir sont les suivantes :

- Sang entier (infirmière seulement)
- Fonction respiratoire (spirométrie) (infirmière seulement)
- Tension artérielle et pouls
- Taille et poids
- Tour de taille
- Impédance bioélectrique
- Force de préhension
- Salive (intervieweur seulement)
- Pastilles de sang séché (intervieweur seulement)

Au Royaume-Uni, des protocoles et des normes sont bien établis pour la collecte des mesures physiques par des infirmières-intervieweuses, et cela se fait dans plusieurs études à grande échelle. La collecte de certaines des mesures proposées par des intervieweurs formés, et plus particulièrement la tension artérielle et le pouls, la salive et les pastilles de sang séché, ne s'est jamais faite dans des enquêtes à grande échelle au Royaume-Uni. Les protocoles et la formation des intervieweurs pour cet élément sont innovateurs au Royaume-Uni. Il existe une expérience considérable, dans des enquêtes majeures auprès de la population aux États-Unis, de l'utilisation réussie de ces approches « peu invasives » par des

intervieweurs d'enquête bien formés (McDade et coll., 2007, Lindau et McDade, 2008; Weinstein et coll., 2008), et nous tirons parti de cette expérience pour élaborer nos procédures.

Outre les mesures physiques, le cycle 3 permettra de recueillir des mesures des capacités/du fonctionnement cognitifs pour l'ensemble de l'échantillon. Globalement, ces données devraient fournir l'occasion d'évaluer les facteurs d'exposition et d'antécédents de l'état de santé actuel des personnes, permettre de mieux comprendre les mécanismes de la maladie, comme les interactions entre les gènes et l'environnement, les effets du ménage et les effets socioéconomiques sur la santé, et rendre possible une analyse des résultats à partir d'évaluations directes et de couplage des données.

4. Conclusion

Understanding Society est une étude très ambitieuse, tant du point de vue de la portée que de la couverture, et elle est conçue pour que le Royaume-Uni demeure à la fine pointe de la recherche en sciences sociales. Parmi les caractéristiques majeures figure le caractère biosocial de l'enquête, qui fournit de nouvelles possibilités de recherche interdisciplinaire dans les domaines des sciences de la santé et des sciences sociales. En tant que source de données, *Understanding Society* offre par conséquent des données de types et de sources multiples pour un échantillon longitudinal de ménages et de personnes au fil du temps.

La BHPS a obtenu un succès énorme au cours des 18 dernières années, et elle est très appréciée comme source de données longitudinales au Royaume-Uni. Les données sont largement utilisées par les chercheurs universitaires au Royaume-Uni et dans le monde, ainsi que par les ministères publics et les décideurs, comme sources de données longitudinales de grande qualité faisant autorité pour le Royaume-Uni. La BHPS a été conçue selon les normes de recherche les plus élevées et a constamment évolué tout au long de sa durée de vie, la transition à *Understanding Society* constituant le progrès le plus récent en vue d'assurer un avenir prometteur pour la poursuite de la BHPS au cours des années qui viennent. *Understanding Society* est une étude ambitieuse dont la portée est plus large que celle de la BHPS du point de vue du plan de sondage, du contenu et des utilisations possibles de l'analyse. Il est par conséquent essentiel qu'elle se déroule bien, la transition de l'échantillon de la BHPS à la nouvelle étude étant un élément important du succès de celle-ci. Les décisions de conception auxquelles l'équipe d'enquête de l'ISER doit faire face ne sont pas faciles et, dans nombre de cas, le bilan ne se dégagera que lorsque paraîtront les analyses à partir de l'échantillon de la BHPS, de concert avec le nouvel échantillon *Understanding Society*, commencent à paraître. Le processus décisionnel a dû s'appuyer sur la mise à disposition des données et faire intervenir les conseils des utilisateurs, le respect des besoins d'analyse longitudinale et les objectifs de l'enquête, ainsi que l'expérience acquise dans la conception et la gestion d'enquêtes longitudinales. Nous souhaitons avoir trouvé le bon équilibre entre la continuité et l'innovation nécessaires pour *Understanding Society*, tout en demeurant fidèles aux principes directeurs de la BHPS, à savoir fournir des données longitudinales de grande qualité.

Bibliographie

Chandola, T., Brunner, E. et Marmot, M. (2006). Chronic stress at work and the metabolic syndrome: prospective study. *British Medical Journal* 332: 521-525.

Laurie, H., Burton, J. et Uhrig, S.C. Noah (eds) (2008). Understanding Society: Some preliminary results from the wave 1 Innovation Panel, *Understanding Society Working paper Series, 2008 – 03, Institute for Social and Economic Research, University of Essex* http://research.understandingsociety.org.uk/files/working-papers/2008/usocwp-2008-03.pdf

McDade, T.W., Williams, S. et Snodgrass, J.J. (2007). What a drop can do: dried blood spots as a minimally-invasive method for integrating biomarkers into population-based research. *Demography* 44(4): 899-925.

Lindau, S.T. et McDade, T.W. (2008). Minimally-invasive and innovative methods for biomeasure collection in population-based research. In M. Weinstein, J.W.Vaupel and K.W. Wachter (Eds) *Biosocial Surveys*. Washington D.C., National Academies Press.

Sabbah, W., Watt, R.G., Shelham, A. et Tsakos, A. (2008). Effects of allostatic load on the social gradient in ischaemic heart disease and periodontal disease: evidence from the Third National Health and Nutrition Examination Survey. *Journal of Epidemiology and Community Health* 62: 415-420.

Seeman, T.E., Crimmins, E., Huang, M.-H. et al. (2004). Cumulative biological risk and socio-economic differences in mortality: MacArthur Studies of Successful Aging. *Social Science and Medicine* 58: 1985-1997.

Weinstein, M., Vaupel, J.W. et Wachter, K.W. (Eds, 2008). *Biosocial Surveys*. Washington D.C., National Academies Press.

Pour plus de renseignements, voir :

http://www.understandingsociety.org.uk
http://www.iser.essex.ac.uk

La survie et la reprise de la Survey of Income and Program Participation

David S. Johnson [1]

Résumé

Depuis deux décennies, la Survey of Income and Program Participation (SIPP) constitue la principale source de données sur le bien-être économique des Américains. De nombreux établissements d'enseignement et organismes aux niveaux fédéral, des États et des municipalités, ainsi que des organismes privés de recherche et d'étude de politiques, utilisent la SIPP pour évaluer l'efficacité des programmes gouvernementaux. Récemment, le Census Bureau des États-Unis a entrepris de remanier la SIPP afin de fournir des renseignements essentiels et ponctuels à coût réduit en remaniant le plan d'enquête, en améliorant l'efficacité du traitement et en ciblant davantage le contenu. La SIPP vise avant tout à produire un échantillon représentatif à l'échelle nationale afin d'évaluer la dynamique annuelle et infra-annuelle du revenu, le recours occasionnel aux programmes de transferts gouvernementaux et leur incidence sur le contexte familial et social des particuliers et des ménages. Les principales activités de ce processus de remaniement sont les suivantes : 1) amélioration de l'instrument de collecte et du système de traitement; 2) intégration d'un calendrier rétrospectif à l'instrument d'enquête; 3) utilisation des données de dossiers administratifs pour compléter et évaluer les données d'enquête; 4) développement du contenu de l'enquête et ajout de suppléments remboursables par le biais d'interactions avec les parties intéressées.

La consultation et l'intervention des parties intéressées au sujet du contenu et de la conception des améliorations proposées a constitué une activité importante dans le processus de remaniement de la SIPP. De plus, le Bureau a commencé en septembre 2008 à mener sur le terrain la nouvelle collecte de données de la SIPP. Plus récemment, les activités suivantes se sont ajoutées au processus de remaniement :

- Évaluation d'une version papier provisoire d'un questionnaire à calendrier rétrospectif.
- Planification d'un essai à grande échelle d'un questionnaire à calendrier rétrospectif automatisé au début de 2010.
- Reconstitution d'un sous-comité consultatif de l'American Statistical Association sur la méthodologie de recherche statistique de la SIPP.
- Assemblées publiques portant sur les renseignements nécessaires pour obtenir des recommandations d'un panel de la National Academy of Sciences et du Committee on National Statistics chargé de conseiller le Bureau sur la planification de l'utilisation de dossiers administratifs dans le cadre de la SIPP remaniée et sur les recherches connexes.
- Acquisition de dossiers administratifs et recours à des services de consultation sur des données nationales et certaines données au niveau des États concernant les programmes gouvernementaux pour évaluer la qualité des données papier et des données d'essai automatisées.
- Planification d'une consultation approfondie sur la formation des représentants sur le terrain en ce qui concerne la méthode d'interview liée au calendrier rétrospectif.

[1] David S. Johnson, U.S. Census Bureau, É.-U. (david.s.johnson@census.gov)

Les résultats de l'Enquête pilote par panel auprès des ménages canadiens

Andrew Heisz[1]

Résumé

En janvier 2006, une conférence sur les enquêtes longitudinales organisée par Statistique Canada, le Conseil de recherches en sciences humaines et les Instituts de recherche en santé du Canada concluait que le Canada n'a pas d'enquête longitudinale permettant de recueillir des données sur des sujets multiples, comme la famille, le capital humain, le revenu, le travail et la santé et de suivre les répondants pendant une longue période. Après cette conférence, Statistique Canada a reçu des fonds de l'Initiative visant à combler les lacunes statistiques pour la recherche sur les politiques, à l'appui d'une nouvelle Enquête par panel auprès des ménages canadiens sur une base pilote (EPMC pilote). Des consultations relatives à la conception et au contenu se sont tenues avec des universitaires et des spécialistes de l'élaboration de politiques, en 2007 et 2008, et l'enquête pilote a été menée à l'automne 2008. Les objectifs de l'enquête pilote étaient les suivants : 1) mettre à l'essai un questionnaire et mesurer la qualité des données recueillies; 2) évaluer plusieurs caractéristiques du plan de sondage; et 3) évaluer la réaction à l'enquête des répondants et des intervieweurs sur le terrain. L'enquête pilote a obtenu un taux de réponse de 76 %, et la durée moyenne des interviews auprès des ménages a été de 68 minutes. Plusieurs innovations dans les caractéristiques du plan de sondage ont été évaluées et se sont révélées viables. La réaction des répondants et des intervieweurs à l'enquête a été généralement positive. La présente communication fait état de cela et d'autres résultats de l'EPMC pilote.

Mots clés : Enquête par panel auprès des ménages canadiens; enquêtes longitudinales, enquêtes pilotes.

1. Introduction

À l'automne 2008, Statistique Canada, de concert avec Ressources humaines et Développement social Canada (RHDSC) et la collectivité des chercheurs universitaires canadiens, a tenu une Enquête par panel auprès des ménages canadiens sur une base pilote (EMPC pilote). La présente communication décrit le contexte du projet, les étapes suivies pour l'élaboration de l'enquête pilote, les résultats méthodologiques et la réaction à l'enquête pilote des répondants et des intervieweurs sur le terrain. Une étude distincte permettra d'évaluer le questionnaire et la qualité des données recueillies dans l'enquête pilote.

2. Élaboration de l'enquête pilote

2.1 Contexte

En janvier 2006, Statistique Canada, le Conseil de recherches en sciences humaines et les Instituts de recherche en santé du Canada ont tenu une conférence intitulée « Enquêtes longitudinales sociales et de santé dans une perspective internationale[2] ». L'objectif de cette conférence était de se pencher sur les enquêtes longitudinales sociales et sur la santé tenues au Canada et ailleurs. À ce moment-là, le programme d'enquêtes longitudinales sociales et sur la santé de Statistique Canada avait près de 15 ans, et la conférence a fourni l'occasion d'évaluer les succès et les lacunes du portefeuille d'enquêtes longitudinales de Statistique Canada. On a examiné toutes les enquêtes longitudinales de Statistique Canada et on a effectué des comparaisons avec des enquêtes longitudinales d'autres pays.

La conférence a notamment permis de déceler des lacunes importantes dans les données au Canada : le Canada n'a pas d'« enquête par panel générale auprès des ménages ». Une enquête par panel générale auprès des ménages est une enquête longitudinale auprès des ménages portant sur plusieurs sujets, qui est menée auprès d'un échantillon représentatif de la population. Le Canada mène des enquêtes longitudinales qui portent sur des sujets particuliers, comme l'Enquête nationale sur la santé de la population et l'Enquête sur la dynamique du travail et du revenu (EDTR). Il mène aussi des enquêtes longitudinales qui mettent l'accent sur des sous-populations particulières, comme l'Enquête longitudinale nationale sur les enfants et les jeunes, l'Enquête auprès des jeunes en transition et l'Enquête longitudinale auprès des immigrants du Canada.

[1]Andrew Heisz, Division de la statistique du revenu, Statistique Canada, 5ième étage, Édifice Jean Talon, 170 promenade Tunney's Pasture, Ottawa, Canada, K1A 0T6, andrew.heisz@statcan.gc.ca

[2] La plupart des communications présentées à cette conférence sont disponibles à l'adresse suivante : www.ciqss.umontreal.ca/Longit/index.html.

Ces enquêtes sont utiles, étant donné qu'elles permettent une recherche approfondie sur une sous-population ou un sujet particulier. Toutefois, une enquête par panel générale auprès des ménages permettrait de pousser la recherche au-delà des domaines spécialisés traditionnels, et offrirait aux chercheurs la possibilité de voir comment les événements dans un domaine peuvent en affecter d'autres, parfois plus tard dans la vie. Elle permettrait en outre aux chercheurs d'examiner comment la vie évolue dans les contextes sociaux en général, le plus immédiat étant celui de la famille, mais aussi dans des contextes sociaux plus vastes, comme les amis, les voisins et les services publics. La conception de telles enquêtes nécessite d'interviewer tous les membres du ménage, ce qui permet une analyse de la dynamique de la famille et de ses interactions avec d'autres domaines, qui ne serait pas possible pour d'autres enquêtes.

Un certain nombre d'autres recommandations ont découlé de la conférence « Enquêtes longitudinales sociales et de santé dans une perspective internationale » et ont trait à la conception d'une nouvelle enquête par panel auprès des ménages.

• L'enquête devrait mettre l'accent sur quatre domaines principaux : 1) travail et revenu, 2) famille, 3) développement du capital humain, et 4) santé.
• Elle devrait permettre d'élargir le nombre de sujets pouvant être abordés pendant la durée de vie d'un panel, grâce à des modules de renouvellement.
• Elle devrait comporter une durée de panel indéfinie.
• Elle ne devrait pas être utilisée pour la production d'estimations transversales du revenu.
• Elle devrait faire participer des chercheurs universitaires et des spécialistes de l'élaboration de politiques et maintenir leur engagement comme partenaires pour la gouvernance permanente de l'enquête.
• Elle devrait être souple et pouvoir s'adapter aux besoins émergents en matière de recherche et de politiques.
• Elle devrait être comparable du point de vue de la conception et du contenu aux enquêtes par panel générales auprès des ménages au niveau international, comme l'enquête par panel socioéconomique de l'Allemagne (GSOEP), la British Household Panel Survey (BHPS) et l'enquête Household Income and Labour Dynamics in Australia (HILDA).
• Elle devrait être facile à utiliser et mettre l'accent sur l'importance de réduire le délai de traitement, de simplifier la pondération et la méthodologie et de réduire le délai d'apprentissage nécessaire pour comprendre les ensembles de données[3].

Par suite de la conférence de Montréal, l'Initiative visant à combler les lacunes statistiques pour la recherche sur les politiques a fourni trois années de financement pour l'EPMC pilote.

2.2 Gouvernance et élaboration du contenu

Le projet d'EPMC pilote a été élaboré au moyen d'un système de gouvernance tripartite constitué de représentants de Statistique Canada, de RHDSC et de la collectivité des universitaires. Un comité directeur comprenant deux directeurs généraux de Statistique Canada, un directeur général de RHDSC et deux chercheurs a dirigé le projet. L'enquête a été gérée par la Division de la statistique du revenu de Statistique Canada.

L'élaboration du contenu a été effectuée de février 2007 jusqu'à la fin de mars 2008. Elle a été dirigée par quatre groupes d'universitaires spécialistes, chacun étant responsable d'un des quatre grands domaines spécialisés mentionnés précédemment. Ces quatre groupes de spécialistes, constitués d'environ 20 chercheurs canadiens, ont fourni des conseils sur les besoins et les priorités en matière de contenu et ont discuté des utilisations de recherche possibles de la nouvelle enquête. Chacun des quatre groupes de spécialistes a préparé un rapport indiquant quelles données seraient recueillies pour chaque domaine, y compris les éléments qui pourraient être utilisés comme contenu de renouvellement, par opposition au contenu repris chaque année. Ces rapports ont été élaborés sur plusieurs mois, pendant lesquels un certain nombre de réunions se sont tenues pour échanger des idées préliminaires. Une ébauche de l'enquête pilote a alors été produite.

Les experts en recherche sur les politiques des ministères fédéraux ont été consultés et ont commenté l'ébauche de l'enquête. Des chercheurs de RHDSC, de la Société canadienne d'hypothèques et de logement et de la Banque du Canada ont été consultés sur les besoins de données, ce qui a donné lieu à de nombreux changements à l'ébauche de l'enquête. Un essai qualitatif du questionnaire, sous forme d'interviews d'essai individuelles, a été mené de janvier jusqu'à la fin de mars 2008.

L'élaboration de l'application et le traitement des données ont été effectués à la Division des enquêtes spéciales de Statistique Canada. Comme pour toutes les enquêtes sociales de Statistique Canada, la méthodologie relevait de la Division des méthodes d'enquêtes auprès des ménages, tandis que la collecte a été assurée par la Division de la planification et de la gestion de la collecte.

[3] Voir Picot, Berthelot et Webber (2006) pour plus de détails.

2.3 Conception de l'enquête

L'EPMC pilote a été menée entre le 15 octobre et le 31 décembre 2008. Aux fins des travaux sur le terrain, l'enquête pilote a été appelée l'enquête Vivre au Canada – version pilote. Le plan de sondage a suivi le modèle établi par les enquêtes HILDA, GSOEP et BHPS pour les enquêtes par panel générales auprès des ménages. En résumé :

Population cible : la population cible était constituée de tous les Canadiens vivant dans des ménages, sauf les populations vivant en établissement institutionnel et dans les réserves. Les ménages ont été échantillonnés (la base de sondage et l'approche d'échantillonnage sont abordées ci-après), et tous les membres des ménages sont devenus des membres permanents de l'échantillon, peu importe leur âge.

Répondants visés : l'enquête visait à interviewer tous les membres de 15 ans et plus des ménages échantillonnés.

Mode : la collecte a pris la forme d'interviews sur place assistées par ordinateur (IPAO), sans personne interposée, même si on a donné aux intervieweurs l'option de tenir des interviews par téléphone au moyen de l'application d'IPAO, si une rencontre en personne n'était pas possible.

Taille de l'échantillonnage et base de sondage : l'échantillon était constitué de 1 200 logements sélectionnés à partir de la base de renouvellement de l'Enquête sur la population active (EPA), et de 1 400 logements sélectionnés à partir de la base aréolaire de l'EPA. Des échantillons de tailles à peu près égales ont été tirés pour chacune des quatre provinces – Ontario, Québec, Saskatchewan et Nouveau-Brunswick. Pour réduire les coûts de la collecte, l'enquête pilote comportait une petite taille d'échantillon et a utilisé un plan de sondage comprenant un nombre de grappes important. L'enquête pilote ne devait par conséquent pas produire d'estimations de la population.

Règles de suivi : les règles de suivi pour l'enquête n'ont pas été établies de façon rigide, le but étant de tenir une enquête pilote d'une vague seulement. Toutefois, elles devaient être similaires à celles utilisées dans les enquêtes HILDA, BHPS ou GSOEP. Dans ces enquêtes, tous les membres permanents de l'échantillon sont suivis indéfiniment, les cohabitants des membres permanents de l'échantillon sont aussi interviewés, et les enfants des membres permanents de l'échantillon deviennent eux-mêmes des membres permanents de l'échantillon.

2.4 Contenu de l'enquête pilote

L'enquête pilote a consisté à élaborer le questionnaire de la vague 1, celui-ci visant à recueillir des données de base pour l'élaboration des vagues subséquentes de l'enquête. Par conséquent, l'enquête pilote comprenait certaines questions de nature rétrospective, par exemple, des questions conçues pour brosser un portrait de base du capital humain, du travail, de la santé et du bien-être.

L'enquête a été structurée en fonction d'un certain nombre de composantes.

- La composante entrée a permis de dresser une liste de tous les membres des ménages et de recueillir des données démographiques de base au sujet de chacun. Cette composante a été administrée à la personne identifiée comme connaissant le mieux les membres du ménage (« personne la mieux renseignée » ou PMR).
- Une composante ménage comportait un ensemble des questions concernant le ménage ou ses membres, auxquelles on pouvait raisonnablement s'attendre à ce que la PMR réponde au nom des membres du ménage.
- Une composante membre, qui a été administrée à chacun des membres du ménage de 15 ans et plus. Les membres des ménages de 0 à 14 ans n'ont pas été interviewés.

La composante ménage comportait des questions sur le logement, le recours à des services de garde, les dépenses mensuelles pour des articles clés et le revenu mensuel total du ménage. Elle comprenait aussi des questions sur la sécurité alimentaire, financière et du logement, ainsi que la privation matérielle.

La composante membre comprenait des questions rétrospectives sur les mariages et les unions de fait, les antécédents parentaux et les intentions de fécondité, les antécédents scolaires et les périodes sans emploi. Elle comprenait en outre des questions sur l'activité actuelle sur le marché du travail, les caractéristiques des emplois courants, ainsi que de nouvelles questions destinées aux travailleurs autonomes et des questions sur les compétences utilisées au travail et sur les attentes en matière d'emploi. D'autres questions portaient sur les activités principales des répondants qui ne travaillaient pas. Quatre modules ont servi à recueillir des données sur la transition du travail à la retraite : deux ont été utilisés dans l'enquête pilote, avec comme objectif d'utiliser les deux autres pour la deuxième vague de l'enquête. Ces questions ont été posées aux

répondants de 45 ans et plus. Enfin, la composante membre comprenait des questions sur les caractéristiques démographiques et la satisfaction à l'égard de la vie.

3. Résultats de l'enquête pilote

3.1 Objectifs de l'enquête pilote

Les objectifs de l'EPMC pilote étaient : 1) de mettre à l'essai un questionnaire et de mesurer la qualité des données recueillies; 2) d'évaluer plusieurs plans de sondage; et 3) d'évaluer les réactions à l'enquête des répondants et des intervieweurs sur le terrain. Comme il est noté précédemment, un document décrivant le questionnaire et la qualité des données doit paraître sous peu. Du point de vue des caractéristiques du plan de sondage, les objectifs particuliers étaient les suivants :

1. Établir un taux de réponse.
2. Recueillir des données pour choisir la meilleure base de sondage pour l'enquête.
3. Examiner la difficulté des interviews dans les ménages plus importants.
4. Examiner le recours aux interviews téléphoniques.
5. Obtenir des données sur la longueur du questionnaire et des questions.

Dans la présente section, nous évaluons les caractéristiques du plan de sondage et nous rendons compte des réactions à l'enquête.

3.2 Taux de réponse

Il est important d'obtenir un taux de réponse élevé pour n'importe quelle enquête, mais peut-être plus encore dans le cas d'une enquête longitudinale, un faible taux de réponse à la première vague pouvant jeter des doutes sur la représentativité de l'enquête pour l'ensemble de la vie du panel. Par ailleurs, les réponses à la vague 1 de l'enquête permettent l'élaboration de données auxiliaires, afin de procéder à des ajustements en fonction de l'attrition pour les vagues subséquentes, ce qui pourrait prolonger la durée de vie du panel.

Le Canada, comme d'autres pays dans le monde, a assisté à une diminution du nombre de personnes disposées à répondre à la première vague des enquêtes longitudinales. Par exemple, dans le cadre de l'EDTR, une enquête par panel par ITAO dans laquelle une nouvelle première vague a été lancée tous les trois ans depuis 1993, le taux de réponse à la vague 1 est passé de plus de 90 % pour les premiers panels, à 75 % environ à la fin des années 2000.

Le tableau 3.2-1 montre le taux de réponse à l'EPMC pilote. Pour les vagues suivantes de l'enquête, on aurait visé un taux de réponse individuel, mais pour la première vague, dans laquelle la composition de l'échantillon était inconnue parce que la base d'échantillonnage se situait au niveau du ménage, le taux de réponse cible devait être établi au niveau du ménage. Le taux de réponse cible pour l'EPMC pilote était de 80 % d'interviews partielles des ménages (une interview partielle est une interview auprès de certains membres admissibles du ménage, mais pas tous). Ce taux de réponse cible a été établi en fonction des taux de réponse récents d'autres enquêtes par IPAO de Statistique Canada. Au total, 2 356 adresses ont été émises, soit 2 122 ménages admissibles à l'interview. Le taux de réponse (complet et partiel) a été de 76 %, ce qui est légèrement supérieur à l'objectif. Globalement, 5 453 personnes ont été dénombrées, et 3 205 étaient admissibles à l'interview. Au total, 91 % des personnes admissibles à l'interview ont été interviewées.

Tableau 3.2-1
Taux de réponse à l'EPMC pilote

Résultats des ménages de la vague 1		
Adresses émises	2 356	
Hors du champ de l'enquête	237	
Multiples (ajouts à l'échantillon)	3	
Ménages admissibles	2 122	100%
Refus et absence de contact	501	24%
Couverture complète et partielle des ménages	1 621	76%
Couverture complète des ménages	1 405	66%
Résultats individuels de la vague 1		
Personnes dénombrées	5 453	
Enfants non admissibles (moins de 15 ans)	2 248	
Adultes dénombrés	3 205	100%
Refus, absence de contact et interviews partielles	298	9%
Interviews complètes	2 907	91%

L'une des façons d'évaluer le succès de ce taux de réponse serait d'examiner les taux de réponse à la vague 1 d'autres enquêtes par panel auprès des ménages qui ont obtenu du succès, afin de déterminer si l'EPMC pilote répondait aux normes établies par les enquêtes internationales. Les résultats publiés pour la vague 1 de la BHPS en 1991 (Lynn, 2006)) et la vague 1 d'HILDA en 2001 (Watson et Wooden, 2002) montrent que la BHPS de 1991 a obtenu un taux de réponse partielle et complète de 74 %, tandis que l'HILDA de 2001 a obtenu un taux de réponse partielle et complète de 66 %, ce qui montre que le taux de réponse de 76 % à l'EPMC pilote était suffisant pour que cette enquête obtienne le même niveau de succès que les enquêtes internationales. Parmi les domaines possibles d'amélioration de l'EPMC pilote figurait la part d'interviews complètes, qui se situait à 95 % dans la BHPS, mais à seulement 91 % dans l'EPMC pilote.

3.3 Bases de sondage

Comme il est noté précédemment, l'EPMC pilote a utilisé deux bases de sondage pour tirer son échantillon, la base de sondage de renouvellement de l'EPA et la base aréolaire de l'EPA. Cette approche visait à évaluer quelle base de sondage produirait un taux de réponse plus élevé. La base de sondage de renouvellement comprenait les logements utilisés précédemment dans l'EPA de septembre 2007. Dans la plupart des cas, les répondants de ces logements avaient passé jusqu'à six mois dans l'EPA, ce qui peut avoir eu des effets négatifs sur la réponse. Toutefois, l'avantage d'utiliser la base de sondage de renouvellement est qu'elle contient une somme importante de données auxiliaires sur les répondants, qui peuvent servir à améliorer l'ajustement pour la non-réponse à la vague 1. Par contre, les répondants des logements de la base aréolaire n'ont pas d'autre expérience des enquêtes de Statistique Canada (mis à part le recensement), mais les données auxiliaires disponibles dans cette base de sondage sont beaucoup plus limitées. Il convient de souligner qu'aucun dépistage des anciens répondants de l'EPA sélectionnés dans la base de sondage de renouvellement n'a été effectué dans l'enquête pilote; on a plutôt mené des interviews avec les occupants actuels des logements. Idéalement, nous aurions dû comparer les taux de réponse des deux bases de sondage après avoir tenté de dépister les membres de l'échantillon de renouvellement qui étaient déménagés au cours de l'année suivant l'EPA de septembre 2007.

Tableau 3.3-1
Résultats de la comparaison de la base aréolaire de l'EPA et de la base de renouvellement de l'EPA

	Base aréolaire		Base de renouvellement	
Ménages compris dans le champ de l'enquête	1 180		942	
Non-réponse totale	265	22%	236	25%
Absence de contact	18	2%	18	2%
Refus	174	15%	167	18%
Limitations mentales/physiques	9	1%	6	1%
Obstacle linguistique	9	1%	4	0%
Autre	55	5%	41	4%

Les résultats des deux bases de sondage figurent dans le tableau 3.3-1. On peut voir que le taux de refus de la base aréolaire (15 %) a été plus faible que celui de la base de renouvellement (12 %), mais pas dans une très grande mesure. En outre, l'absence de contact pour la base de renouvellement (2 %) aurait probablement été plus élevée si on avait procédé au dépistage.

Toutefois, le taux de réponse légèrement plus élevé de la base aréolaire se fait au détriment de la perte de données auxiliaires utiles disponibles dans la base de renouvellement. Par conséquent, il est difficile de déterminer avec précision laquelle de la base aréolaire ou de la base de renouvellement aurait été la plus appropriée pour l'enquête. Les résultats de cet essai laissent cependant supposer que l'effet du fardeau de réponse antérieur de l'EPA sur le taux de réponse serait minime si l'on choisissait d'utiliser la base de renouvellement.

3.4 Difficulté à mener des interviews complètes dans les ménages importants

L'une des considérations importantes du succès de l'enquête est le fardeau de réponse que le plan de sondage impose aux ménages importants, et la difficulté correspondante d'interviewer chaque répondant dans ces ménages. Le tableau 3.4-1 comprend des statistiques sur les taux d'achèvement pour les ménages de diverses tailles. Tout d'abord, les ménages importants étaient assez rares. Seulement 16,3 % des ménages ont nécessité trois interviews ou plus, tandis que 4,2 % en ont nécessité quatre ou plus. Toutefois, il a été plus difficile de mener des interviews complètes dans les ménages importants. Au total, on a mené des interviews complètes dans 99 % des ménages comptant un répondant, comparativement à 85 % pour les ménages à deux répondants, 66 % pour les ménages à trois répondants, 76 % pour les ménages à quatre répondants, et 56 % pour les ménages à cinq répondants

La plupart (52 %) des répondants vivaient dans des ménages comptant deux répondants ciblés. En moyenne, il a fallu 4,6 tentatives pour interviewer les deux répondants du ménage. Il est intéressant de constater que le nombre moyen de contacts requis pour mener une interview complète d'un ménage comptant deux répondants n'était pas plus élevé que pour un ménage à un répondant, le nombre moyen de contacts se situant à 4,5 pour un ménage à un répondant. Toutefois, davantage de contacts et, par conséquent, un plus grand effort des intervieweurs, ont été requis pour les ménages comptant trois répondants ou plus.

Tableau 3.4-1
Résultats de la comparaison de la base aréolaire de l'EPA et de la base de renouvellement de l'EPA

Taille du ménage	1	2	3	4	5
% des ménages	31,2	52,2	12,1	3,3	0,9
% d'interviews complètes	98,8	85,3	66,2	76,1	56,0
Tentatives en vue d'une interview complète	4,5	4,6	6,4	6,9	7,3

3.5 Interview téléphonique

Comme il est noté précédemment, l'EPMC pilote a pris la forme d'interviews par ITAO. On a demandé aux intervieweurs de faire de leur premier contact avec le ménage un contact personnel et, dans la mesure du possible, d'interviewer chaque répondant en personne. Toutefois, les intervieweurs avaient la liberté de mener des interviews par téléphone avec les répondants difficiles à joindre. On a procédé ainsi afin de réduire le fardeau de réponse pour les ménages importants, d'augmenter la proportion de ménages complets et de réduire les coûts de la collecte. Par ailleurs, les cas soumis aux intervieweurs principaux étaient plus susceptibles d'avoir fait l'objet d'interviews téléphoniques.

Le tableau 3.5-1 montre le nombre de tentatives, le nombre de contacts et le nombre de cas complets pour les deux méthodes de collecte. Une tentative est un effort en vue de contacter un ménage, un contact est une tentative qui a permis de joindre un membre du ménage, et un cas complet est un contact qui a permis de finaliser le cas, ce qui signifie que les intervieweurs n'ont pas eu à contacter le ménage à nouveau.

Tableau 3.5-1
Comparaison des tentatives d'interview sur place et par téléphone

Méthode de collecte	Nombre de tentatives	Nombre de contacts	Nombre de cas complets
Sur place	9 266	4 715	1 164
Par téléphone	5 069	3 015	442

Au total, 35 % des tentatives, 39 % des contacts et 27 % des cas complets ont découlé d'interviews téléphoniques, ce qui représente des taux relativement élevés pour une enquête menée au moyen d'interviews sur place. Ces taux sont plus faibles,

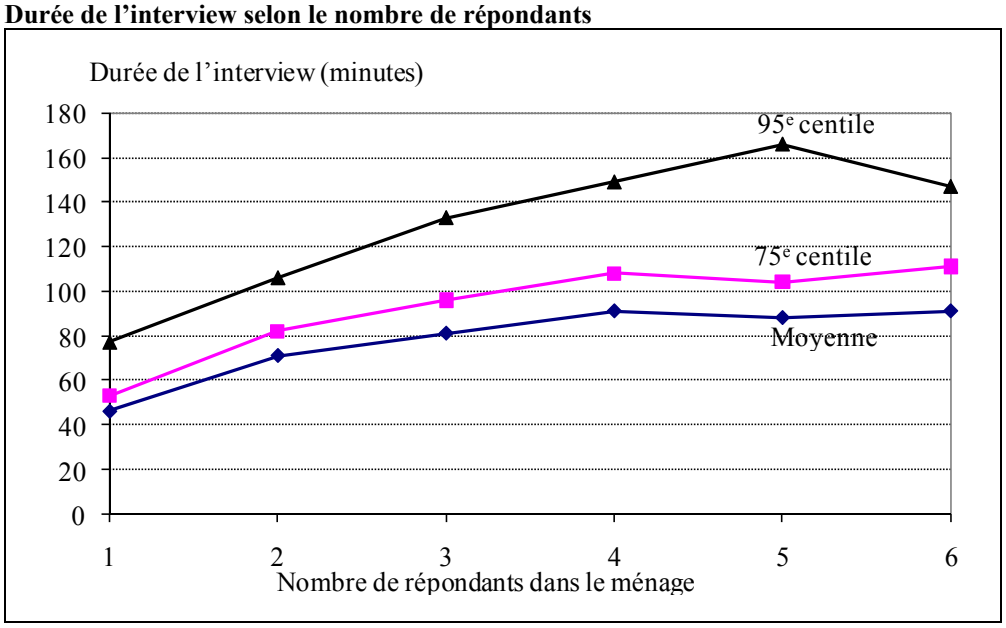
mais ils demeurent élevés si nous excluons du calcul les tentatives des intervieweurs principaux (qui sont habituellement effectuées par téléphone). Cela souligne l'importance de la collecte téléphonique pour obtenir le taux de réponse décrit précédemment. Il se peut que les interviews téléphoniques doivent être intégrées davantage dans le plan de sondage, grâce peut-être à l'élaboration d'une procédure plus formelle pour l'utilisation du téléphone, ou encore au recours aux instruments d'ITAO (outre l'IPAO) et aux centres d'appel d'ITAO.

3.6 Durée de l'interview et des questions

La durée prévue de l'interview a probablement été l'élément le plus examiné pendant la conception de l'EPMC pilote, étant donné les préoccupations selon lesquelles les interviews de tous les membres du ménage de 15 ans et plus feraient en sorte que l'on passerait trop de temps dans certains ménages. Même si on a étudié la possibilité de limiter le nombre d'interviews à quatre par ménage, il a été décidé de limiter plutôt la longueur des interviews, grâce à une épuration soigneuse du questionnaire. À cette fin, on a utilisé des données de piste de vérification concernant la durée d'interview par question d'autres enquêtes, et on a eu recours à des discussions avec des experts spécialisés et des experts de la collecte. Lorsqu'on ne disposait pas de données de piste de vérification, on a procédé à des évaluations ponctuelles, par exemple, en minutant les questions dans des situations d'interview simulée.

À la fin, la durée moyenne des interviews était de 12,6 minutes pour la composante ménage et de 24,8 minutes pour la composante membre. Ces durées d'interview étaient assez près de ce que l'on avait prévu avant la collecte. Si l'on tient compte uniquement des ménages dont tous les répondants ciblés ont été interviewés, la durée moyenne de l'interview a été de 68 minutes (y compris l'entrée et la sortie). La figure 3.6-1 montre les durées d'interview selon le nombre de répondants ciblés (pour les cas complets seulement) et montre que la durée moyenne d'interview a été de 46 minutes dans un ménage à un répondant, de 71 minutes dans un ménage à deux répondants, de 81 minutes dans un ménage à trois répondants, et de 91 minutes ou moins dans les ménages à quatre, cinq ou six répondants. Même si les interviews de 91 minutes étaient considérées comme longues dans le contexte de l'EPMC pilote, dont la durée des interviews est répartie entre les membres du ménage, on les a jugées acceptables. Toutefois, la figure 3.6-1 révèle aussi que, dans certains ménages, la durée des interviews a été beaucoup plus longue que la moyenne. Par exemple, 5 % des ménages à quatre répondants ont nécessité plus de 149 minutes d'interview, ce qui montre que les mesures visant à réduire la durée des interviews inhabituellement longues sont justifiées.

Figure 3.6-1
Durée de l'interview selon le nombre de répondants



3.7 Réactions à l'enquête

Au total, quatre gestionnaires de programme, dix intervieweurs principaux et 109 intervieweurs réguliers ont participé à la collecte des données de l'EPMC pilote, et la formation des intervieweurs sur le terrain a fourni de nombreuses occasions aux personnes chargées de l'élaboration de l'enquête d'obtenir des réactions informelles à l'égard de celle-ci. En outre, des conférences téléphoniques hebdomadaires se sont tenues avec les gestionnaires de programme pendant la période de travail sur

le terrain, afin de les informer des progrès réalisés, ainsi que de soulever des questions ou des préoccupations. Enfin, un questionnaire de compte rendu a été distribué aux intervieweurs après la collecte, afin d'obtenir leur rétroaction directe au sujet de l'enquête.

En général, on a déterminé que les intervieweurs étaient très réceptifs à l'enquête. Selon les résultats du questionnaire de compte rendu, les répondants étaient aussi réceptifs : 95 % des intervieweurs ont indiqué que les répondants avaient une attitude positive ou neutre à l'égard de l'enquête. Un nombre important d'intervieweurs (80 %) étaient d'avis que les répondants comprenaient clairement l'objectif de l'enquête, et 75 % des intervieweurs croyaient que la tenue d'interviews sur place ne causait pas de problème. Plusieurs intervieweurs ont déclaré que l'utilisation des interviews téléphoniques leur permettait de surmonter les problèmes de planification de la charge de travail.

Néanmoins, les intervieweurs ont fait état d'un certain nombre de préoccupations en ce qui a trait au plan de sondage. Parmi les préoccupations majeures, presque la moitié des intervieweurs croyaient que l'interview était trop longue, et 61 % étaient d'avis que les répondants multiples ou les visites multiples entraînaient des problèmes pendant la collecte. La planification des interviews et les visites répétées aux ménages importants sont deux aspects qui ont présenté des difficultés.

4. Conclusion

On continue d'évaluer les résultats de l'EPMC pilote. De façon plus particulière, l'attention de l'équipe d'enquête à Statistique Canada et des chercheurs se concentre maintenant sur l'évaluation du contenu de l'EPMC pilote, qui figurera dans un rapport distinct.

L'EPMC pilote a permis de démontrer la faisabilité d'une enquête par panel générale auprès des ménages au Canada, d'établir le taux de réponse probable pour la vague 1, de déterminer une longueur cible pour les interviews et de produire des données utiles sur l'utilisation de la base de sondage. Ces résultats et d'autres expériences tirés de l'EPMC pilote permettront d'éclairer les travaux à venir sur l'élaboration d'enquêtes longitudinales auprès des ménages à Statistique Canada.

Bibliographie

Picot, G., Berthelot J.-M., et Webber M. (2006). Possible Future Directions for Longitudinal Surveys at Statistics Canada, *Conférence sur les enquêtes longitudinales sociales et de santé dans une perspective internationale*, Montréal, Canada.

Watson, N. et Wooden, M. (2002). The Household, Income and Labour Dynamics in Australia (HILDA) Survey: Wave 1 Survey Methodology, *Hilda Project Technical Paper Series,* No.1/02, May 2002, University of Melbourne, Melbourne, Australie.

Lynn, P. (2006). Quality Profile: British Household Panel Survey, Version 2.0: Waves 1 to 13: 1991-2003, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Royaume-Uni.

SESSION 16 – MODÉLISATION DE VARIABLES LATENTES ET ESTIMATION BAYÉSIENNE

Évolution de la satisfaction de vivre dans la British Household Panel Survey : modélisation de la courbe de croissance latente

Maria de Fátima Salgueiro, Marcel D. T. Vieira et Peter W. F. Smith[1]

Résumé

Les trajectoires de satisfaction à l'égard de la vie sont modélisées en se servant de données tirées de la British Household Panel Survey (BHPS), qui est une enquête nationale représentative de la population réalisée annuellement. On demande aux répondants d'indiquer dans quelle mesure ils sont satisfaits de leur revenu, de leur logement, de leur emploi, de leur vie sociale, de leur nombre d'heures de loisirs et de leur utilisation de ces heures. Notre étude porte sur les personnes occupées qui ont été interviewées à chacune des vagues de l'enquête réalisées de 1991 à 2005 et qui ont donné des réponses complètes pour toutes les variables de satisfaction à l'égard de la vie lors de chacune des vagues étudiées. Nous recourons à la modélisation des courbes de croissance latente pour modéliser la variation des deux facteurs latents de satisfaction à l'égard de la vie pris en considération. Nous discutons des avantages qu'il y a à tenir compte du plan de sondage complexe.

Mots-clés : modèles de courbe de croissance latente, données longitudinales, bien-être subjectif.

1. Introduction

Ces dernières années, les spécialistes des sciences sociales ont manifesté un intérêt croissant pour le bien-être subjectif (BES). Plusieurs mesures ont été proposées, le choix de l'instrument ayant une influence sur l'évaluation du bien-être subjectif et de ses déterminants (Peasgood, 2007). La British Household Panel Survey (BHPS), qui est une enquête nationale représentative de la population réalisée annuellement depuis 1991, fournit plusieurs mesures du bien-être subjectif. Depuis la vague de 1996, en plus d'une question sur la satisfaction globale à l'égard de la vie, on demande aux répondants d'indiquer leur niveau de satisfaction à l'égard de huit dimensions du domaine (santé, revenu, logement, conjoint(e)/partenaire, emploi, vie sociale, nombre d'heures de loisirs et utilisation de ces heures. Les approches statistiques décrites dans la littérature spécialisée pour modéliser le bien-être subjectif comprennent souvent des modèles probit ordonnés et des modèles à effets fixes. On a également proposé des modèles à effets aléatoires et des modèles à équations structurelles avec décalage pour modéliser les données d'enquête longitudinale (p. ex., voir Berrington et coll., 2008).

Dans le présent article, nous étudions et modélisons les trajectoires de satisfaction à l'égard de la vie en nous servant des données de la BHPS. Notre étude porte sur les personnes occupées qui ont été interviewées à chacune des vagues 1 à 15 et qui ont donné des réponses complètes pour toutes les variables de satisfaction à l'égard de la vie à chacune de ces vagues. Nous recourons à des modèles de courbe de croissance latente (MCCL) pour modéliser la variation, aux niveaux intra-individu et inter-individus, des deux facteurs latents de satisfaction perçue à l'égard de la vie pris en considération. Les déterminants éventuels de la satisfaction à l'égard de la vie comprennent l'âge, le sexe, le fait d'avoir des enfants, le revenu familial, le niveau de scolarité et le nombre d'heures de travail. Nous discutons des avantages qu'il y a à tenir compte du plan de sondage complexe. La présentation de l'article est la suivante. À la section 2, nous décrivons le cadre conceptuel et les variables analysées, et nous présentons certaines statistiques descriptives pour caractériser l'échantillon. À la section 3, nous présentons le MCCL conditionnel proposé pour décrire et expliquer les trajectoires de croissance de la *satisfaction à l'égard des loisirs* et de la *satisfaction matérielle*, et nous résumons les options méthodologiques examinées pour la modélisation statistique. La section 4 décrit les principaux résultats et la section 5 contient une brève discussion.

[1]Maria de Fátima Salgueiro, ISCTE-IUL – Lisbon University Institute, Av. Forças Armadas, 1649-026 Lisboa, Portugal (fatima.salgueiro@iscte.pt); Marcel D.T. Vieira, Universidade Federal de Juíz de Fora, Brasil (marcel.vieira@ufjf.edu.br); Peter W.F. Smith, Southampton Statistical Sciences Research Institute, University of Southampton, United Kingdom (P.W.Smith@soton.ac.uk)

2. Les données

1.1 Cadre conceptuel et variables analysées

L'analyse porte sur les données de la British Household Panel Survey (BHPS). Nous considérons quatre vagues de l'enquête, couvrant la période de 2002 à 2005. L'échantillon comprend 2 255 personnes occupées, qui sont restées occupées tout au long de la période étudiée, âgée de 16 ans et plus, qui ont participé à l'enquête depuis la première vague (afin de pouvoir utiliser les poids longitudinaux) et qui ont répondu complètement à l'interview.

Six variables de satisfaction à l'égard de la vie son utilisées comme dimensions du bien-être subjectif. Les variables sont mesurées sur une échelle de type Likert de sept points et comprennent la satisfaction à l'égard de la vie sociale, du nombre d'heures de loisirs, de l'utilisation des heures de loisirs, du revenu du ménage, du logement, ainsi que de l'emploi. Les trois premières variables devraient être des indicateurs d'un facteur de *satisfaction à l'égard des loisirs*, tandis que les trois dernières devraient mesurer la *satisfaction matérielle*. Les deux facteurs latents devraient être corrélés positivement. Neuf déterminants possibles de la satisfaction à l'égard de la vie sont considérés comme variables explicatives, à savoir le sexe, le groupe d'âge, le nombre d'enfants dans le ménage, la qualification, la classe sociale, l'état matrimonial, l'état de santé perçu, le nombre d'heures de travail par semaine et le revenu du ménage.

En ce qui concerne le plan de sondage, l'échantillon initial (première vague) de la BHPS a été sélectionné selon un plan en grappes à plusieurs degrés stratifié. L'information sur les unités primaires d'échantillonnage (UPE), les strates et les poids longitudinaux est disponible dans les ensembles de données de la BHPS et est utilisée dans l'analyse. Les poids longitudinaux de la BHPS tiennent compte de l'érosion de l'échantillon dans l'intervalle entre chaque paire de vagues adjacentes, ainsi que du poids initial à la première vague (pour plus de renseignements, voir Taylor et coll., 2008).

1.2 Certaines statistiques descriptives

L'échantillon comprend 2 255 personnes, dont 49 % sont des femmes et 51 % sont des hommes. La répartition selon le groupe d'âge est la suivante : 4,7 % de personnes âgées de 16 à 21 ans, 13,9 % âgées de 22 à 29 ans, 23,7 % âgées de 30 à 39 ans, 32,1 % âgées de 40 à 49 ans, et les 25,6 % restants âgées 50 ans et plus. En ce qui concerne l'état matrimonial, 74,2 % de ces personnes vivent en compagnie de quelqu'un et 25,8 % vivent seules. Le nombre d'enfants dans le ménage varie de 0 (74,2 %) à 6, et est inférieur ou égal à 2 dans 94 % des ménages. Pour 77 % des répondants, le nombre d'heures de travail est de 30 heures par semaine ou plus. L'état de santé est perçu comme étant excellent par 28 % des répondants, comme très bon par 50,3 %, comme passable par 16,9 %, et comme mauvais ou très mauvais par 4,8 %.

Les statistiques descriptives pour les six variables de satisfaction en 2002 montrent que, pour la plupart de celles-ci, la fréquence relative des réponses variant de 5 (satisfait) à 7 (entièrement satisfait) était supérieure à 60 %. Seule fait exception la satisfaction à l'égard du nombre d'heures de loisirs, pour laquelle la valeur est de 50,6 %. Ces résultats donnent à penser que la majorité des répondants éprouvent des niveaux de satisfaction relativement élevés. Si l'on considère les valeurs moyennes de satisfaction, le score moyen est supérieur à 4,3 pour chacune des six variables durant chacune des quatre années prises en considération dans l'analyse. Les niveaux moyens de satisfaction sont demeurés relativement stables au cours de la période de référence de l'étude.

3. Le modèle de courbe de croissance latente proposé

Les modèles de courbe de croissance latente (MCCL) font partie de la famille générale des modèles à effets aléatoires – pour une introduction, voir Preacher et coll. (2008). Un avantage important de cette méthode est qu'elle est appliquée en recourant à la modélisation par équations structurelles (MES), dont elle partage de nombreux points forts et faiblesses (voir Duncan et coll., 2006). L'analyste des données peut ainsi tirer parti des caractéristiques standard de la MES, telles que les tests globaux et l'évaluation de l'ajustement du modèle aux données, les variables latentes à indicateurs multiples et la modélisation simultanée de plusieurs processus avec régressions sur les paramètres de croissance latente. Toutefois, de grandes tailles d'échantillon demeurent nécessaires.

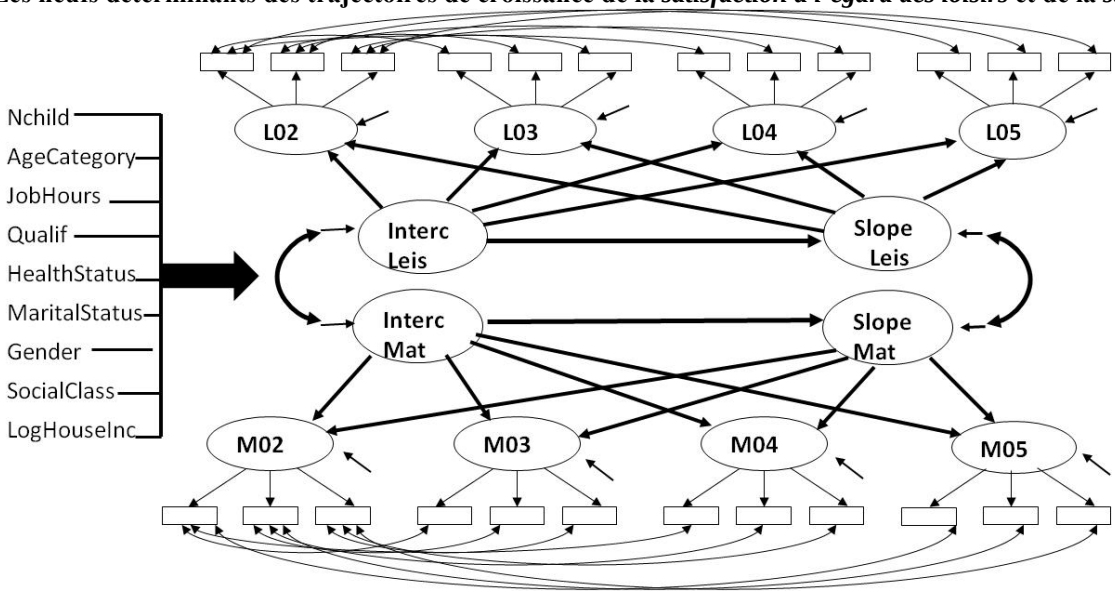
Les MCCL permettent d'étudier les différences inter-individus du changement au cours du temps, ainsi que les antécédents et les conséquences du changement. L'intégration de covariables variant ainsi que ne variant pas avec le temps est également possible. Dans les MCCL, le temps est introduit dans le modèle en imposant des contraintes de saturation factorielle dans un modèle à variables latentes. Les paramètres de croissance sont par conséquent spécifiés sous forme de variables latentes, et des contraintes de saturation factorielle sont imposées. La forme de la trajectoire de croissance (linéaire, quadratique) dépend du nombre de variables latentes spécifiées dans le modèle. Les observations répétées sont considérées comme des mesures

imparfaites d'une trajectoire latente sous-jacente. La forme de ce genre de trajectoire peut être estimée en utilisant la moyenne et la structure de covariance des mesures répétées observées. Pour une explication plus détaillée, consulter Boolen et Curran (2006).

3.1 Le MCCL conditionnel proposé pour décrire et expliquer les trajectoires de croissance de la *satisfaction à l'égard des loisirs* et de la *satisfaction matérielle*

La figure 3.1-1 montre le MCCL conditionnel à deux procédés proposé pour décrire et expliquer les trajectoires de croissance conjointes des facteurs latents de *satisfaction à l'égard des loisirs* et de *satisfaction matérielle*.

Figure 3.1-1
Les neufs déterminants des trajectoires de croissance de la *satisfaction à l'égard des loisirs* et de la *satisfaction matérielle*



Les facteurs latents de *satisfaction à l'égard des loisirs* et de *satisfaction matérielle* sont mesurés tous deux au moyen de trois variables manifestes à chaque point dans le temps. Nous émettons l'hypothèse d'une structure factorielle temporellement invariante et permettons la corrélation temporelle des termes d'erreur des mesures répétées (pour simplifier le graphique, une partie seulement des flèches correspondantes sont représentées à la figure 3.1-1). Nous supposons que les niveaux initiaux de satisfaction à l'égard des *loisirs* et de satisfaction *matérielle* influent sur les trajectoires de croissance correspondantes et, par conséquent, nous procédons à la régression des pentes sur les ordonnées à l'origine.

Afin d'expliquer les trajectoires de croissance conjointes, nous calculons la régression des deux ordonnées à l'origine aléatoires et des deux pentes aléatoires sur les neuf variables explicatives. Le nombre d'enfants dans le ménage et le logarithme du revenu du ménage sont modélisés comme des variables continues. Pour les sept variables explicatives catégoriques, nous créons des variables indicatrices que nous introduisons dans le modèle. Toutes les variables manifestes sont traitées comme des variables ordinales. Nous supposons que la croissance est linéaire, utilisons la convention de codage habituelle de 0, 1, 2 ou 3 pour les saturations, et prenons 2002 comme année de référence.

3.2 Principales options méthodologiques

Nous effectuons la modélisation statistique en trois étapes. En premier lieu, nous utilisons l'analyse factorielle confirmatoire (AFC) pour évaluer l'adéquation d'un modèle contenant deux facteurs corrélés (*Loisirs* et *Matérielle*) en vue de tenir compte des six dimensions de la satisfaction à l'égard de la vie à chaque point dans le temps. En deuxième lieu, nous considérons le MCCL conditionnel de deuxième ordre à deux processus présenté à la figure 3.1-1 pour expliquer conjointement les trajectoires de croissance des deux facteurs latents de satisfaction à l'égard de la vie, en tenant compte du plan de sondage complexe. En troisième lieu, nous estimons le même modèle sans y intégrer l'information sur le plan de sondage et nous comparons les résultats à ceux obtenus à la deuxième étape. Nous estimons tous les modèles en Mplus 5 (Muthén et Muthén, 1998-2007), en utilisant un estimateur par les moindres carrés robuste (MVMCP) disponible (Asparouhov, 2005). Nous ne prenons en considération que les cas valides. Nous utilisons l'option Complex pour intégrer les caractéristiques du plan de

sondage dans les procédures d'inférence et nous estimons les erreurs-types au moyen d'un estimateur de type sandwich (Muthén, 1998-2004).

4. Résultats

4.1 Principaux résultats en tenant compte du plan de sondage complexe

Le tableau 4.1-1 donne les estimations (et les erreurs-types correspondantes entre parenthèses) des principaux paramètres du MCCL ajusté.

Tableau 4.1-1
Estimations (et erreurs-types) des principaux paramètres dans le MCCL conditionnel

	LOISIRS	MATÉRIELLE
Moyenne de l'ordonnée à l'origine	0 (---)	0 (---)
Moyenne de la pente	-0,133 (0,235) n.s.	-0,178 (0,171) n.s.
Variance de l'ordonnée à l'origine	0,599 (0,021)	0,372 (0,025)
Variance de la pente	0,088 (0,016)	0,046 (0,009)
Régression de la pente sur l'ordonnée à l'origine	0,051 (0,039) n.s.	-0,023 (0,035) n.s.
Covariance entre les ordonnées à l'origine aléatoires	0,373 (0,017)	
Covariance entre les pentes aléatoires	0,054 (0,008)	

n.s. désigne une estimation non significative.

Nous obtenons des résultats comparables si nous ne tenons pas compte du plan de sondage complexe. Nous pouvons également conclure qu'il existe plusieurs prédicteurs non significatifs des trajectoires de croissance moyenne de la satisfaction à l'égard de la vie. En effet, aucune des neuf variables explicatives n'a un effet significatif sur la pente aléatoire du facteur *satisfaction à l'égard des loisirs* et la plupart des variables aléatoires n'ont pas d'effet significatif sur la trajectoire de croissance moyenne du facteur *satisfaction matérielle* entre 2002 et 2005. Il convient de noter que les neuf variables explicatives sont modélisées comme étant temporellement invariantes. Seules les valeurs de 2002 ont été prises en considération dans le modèle.

4.2 Effets de la non-prise en compte du plan de sondage complexe

Le tableau 4.2-1 donne les estimations (et les erreurs-types correspondantes) des coefficients de régression associés aux variables explicatives du modèle. Les résultats sont présentés quand on ne tient pas compte du plan de sondage et quand on en tient compte. Les valeurs en caractères gras représentent des estimations significatives.

Tableau 4.2-1
Prédicteurs des niveaux moyens de satisfaction à l'égard de la vie en 2002

		LOISIRS Ordonnée à l'origine aléatoire		MATÉRIELLE Ordonnée à l'origine aléatoire	
	Estimation (erreur-type)	**Sans tenir compte du plan de sondage**	**En tenant compte du plan de sondage**	**Sans tenir compte du plan de sondage**	**En tenant compte du plan de sondage**
Sexe	Féminin	-0,04 (0,047)	-0,05 (0,05)	**0,095** (0,043)	0.083 (0.049)
Groupe d'âge	22 à 29 ans	-0,087 (0,106)	-0,089 (0,112)	**-0,199** (0,092)	-0.126 (0.104)
	30 à 39 ans	-0,203 (0,108)	-0,19 (0,114)	-0,152 (0,095)	-0.08 (0.099)
	40 à 49 ans	**-0,338** (0,105)	**-0,335** (0,11)	**-0,208** (0,092)	-0.153 (0.095)
	50 ans et plus	**-0,22** (0,106)	**-0,221** (0,107)	-0,061 (0,094)	0.002 (0.098)
Nombre d'heures de travail	16 à 29	**-0,211** (0,091)	-0,177 (0,096)	-0,043 (0,079)	-0.013 (0.085)
	30 à 40	**-0,273** (0,084)	**-0,216** (0,087)	-0,103 (0,076)	-0.092 (0.088)
	Plus de 40	**-0,451** (0,104)	**-0,392** (0,105)	-0,049 (0,095)	-0.04 (0.116)
Qualification	Supérieure	**0,176** (0,063)	**0,146** (0,067)	0,068 (0,056)	0.056 (0.057)
	Niveaux A	0,151 (0,082)	**0,161** (0,079)	0,037 (0,071)	0.069 (0.073)
	Niveaux O	**0,259** (0,074)	**0,258** (0,083)	0,113 (0,064)	0.106 (0.064)
	Autre	**0,344** (0,081)	**0,326** (0,077)	**0,251** (0,071)	**0.246** (0.078)
État matrimonial	Seul(e)	**-0,179** (0,056)	**-0,158** (0,066)	**-0,238** (0,051)	**-0.243** (0.053)
Classe sociale	Cadre	0,015 (0,101)	0,02 (0,114)	-0,069 (0,085)	-0.059 (0.098)
	Travailleur spécialisé	-0,048 (0,105)	-0,041 (0,114)	**-0,198** (0,089)	**-0.215** (0.105)
	Travailleur non spécialisé	-0,015 (0,113)	-0,018 (0,127)	-0,145 (0,095)	-0.152 (0.111)
État de santé	Bon	**-0,314** (0,05)	**-0,336** (0,053)	**-0,314** (0,045)	**-0.309** (0.043)
	Passable	**-0,71** (0,064)	**-0,747** (0,068)	**-0,643** (0,058)	**-0.637** (0.065)
	Mauvais	**-0,859** (0,096)	**-0,887** (0,095)	**-0,787** (0,088)	**-0.81** (0.088)
Nombre d'enfants		**-0.133** (0.026)	**-0,155** (0,03)	0,003 (0,024)	-0,011 (0,028)
Revenu		-0.022 (0.107)	-0,039 (0,116)	**0,537** (0,093)	**0,576** (0,107)

Les valeurs en **caractères gras** correspondent à des estimations significatives.

En ce qui concerne la *satisfaction matérielle*, nous pouvons conclure qu'en moyenne, en 2002, les femmes étaient significativement plus satisfaites que les hommes si l'on tient compte du plan de sondage. En outre, la *satisfaction matérielle* des personnes de 22 à 29 ans et de 40 à 49 ans était, en moyenne, plus faible que celle des personnes de la catégorie de référence (16 à 21 ans). Il convient toutefois de souligner que ces différences ne sont plus significatives si l'on ne tient pas compte du plan de sondage pour estimer le modèle.

Travailler de longues heures réduisait la *satisfaction à l'égard des loisirs* (moyenne en 2002). L'état de santé perçu et l'état matrimonial étaient des déterminants significatifs de la *satisfaction à l'égard des loisirs* ainsi que de la *satisfaction matérielle* en 2002. Un plus grand nombre d'enfants dans le ménage impliquait un niveau moyen plus faible de *satisfaction à l'égard des loisirs* en 2002, tandis que le niveau moyen de *satisfaction matérielle* augmentait de manière significative avec le revenu du ménage.

5. Discussion

Le présent article illustre la façon de modéliser les perceptions du bien-être subjectif en se servant d'un MCCL conditionnel. La modélisation en adoptant le cadre des MES nous donne la souplesse de pouvoir tenir compte de l'erreur de mesure. Les

résultats nous fournissent certaines preuves qu'il importe de tenir compte du plan d'échantillonnage complexe dans l'analyse des données longitudinales, même si l'on ajuste des modèles très perfectionnés. En effet, en omettant de tenir compte des caractéristiques du plan de sondage, nous aurions pu considérer erronément une série de catégories de covariables comme étant significatives, alors qu'en réalité elles ne devaient pas l'être (p. ex. les effets de « sexe féminin » et des groupes d'âge « 22 à 29 ans » et « 40 à 49 ans » sur l'ordonnée à l'origine aléatoire du facteur *satisfaction matérielle*). Nous avons par conséquent constaté, dans les analyses longitudinales que nous avons effectuées, que les plans d'échantillonnage complexes avaient des effets d'accroissement de la variance. Nos futurs travaux porteront sur la façon de traiter les données manquantes et d'inclure dans le modèle des covariables variant temporellement.

Bibliographie

Asparouhov, T. (2005). Sampling weights in latent variable modeling, *Structural Equation Modeling*, 12(3), pp. 411-434.

Berrington, A., Hu, Y., Smith, P.W.F et Sturgis, P. (2008). A graphical chain model for reciprocal relationships between women's gender role attitudes and labour force participation, *Journal of the Royal Statistical Society, series A,* 171(1), pp. 89-108.

Bollen, K. A. et Curran, P. J. (2006). *Latent Curve Models – A Structural Equation Perspective*, New Jersey: Wiley.

Duncan, T. E., Duncan, S. C. et Strycker, L. A. (2006). *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues an Applications*, 2nd ed., New Jersey: Lawrence Erlbaum Associates.

Muthén, B. O. (1998-2004). *Mplus Technical Appendices*, Los Angeles, CA: Muthén and Muthén.

Muthén, L. K. et Muthén, B. O. (1998-2007). *Mplus User's Guide* (5th ed.), Los Angeles, CA: Muthén and Muthén.

Peasgood, T. (2007). Does well-being depend upon our choice of measurement instrument?, paper presented at the British Household Panel Survey Conference*,* University of Essex, Colchester, UK.

Preacher, K. J., Wichman, A. L., MacCallum, R. C. et Briggs, N. E. (2008). *Latent Growth Curve Modelling*, Thousand Oaks: Sage.

Taylor, M.F. (Ed.), avec Brice, J., Buck, N. et Prentice-Lane, E. (2008). *British household panel survey user manual volume A: Introduction, technical report and aeppendices*. Colchester: University of Essex.

Remerciements

Les travaux de recherche de M. de Fátima Salgueiro ont été financés par la subvention PTDC/GES/72784/2006 de la Fundação para a Ciência e a Tecnologia.

Une méthode d'analyse de transition latente pour modéliser l'évolution de l'hétérogénéité non observée d'une population

Andy Ross [1]

Résumé

Cette communication porte sur l'utilité de l'analyse des transitions latentes aux fins de modélisation de l'hétérogénéité non observée de la population au fil du temps. Plus précisément, nous utilisons un cadre de classes latentes pour saisir des sous-groupes de jeunes à engagement/désengagement scolaire de 14 à 16 ans et nous modélisons les transitions dans ces divers groupes au fil du temps. Les recherches quantitatives précédentes dans ce domaine ont souvent utilisé des définitions réductrices et unidimensionnelles du désengagement scolaire, comme le décrochage ou les résultats scolaires décevants. Nous allons au-delà de ces limites dans notre étude, adoptant une approche statistique qui saisit la nature multidimensionnelle du désengagement en utilisant des renseignements tirés de toute une gamme de mesures. Les sous-groupes de jeunes engagés ou désengagés sont définis selon leurs réponses combinées aux questions mesurant les aspirations, les attitudes et les comportements.

Nous procédons à notre analyse par étapes. Dans un premier temps, nous estimons les sous-groupes latents pour trois cycles de données. Nous analysons ensuite les transitions latentes pour tester la stabilité de ces sous-groupes et mesurer les transitions au fil du temps. En dernier lieu, nous ajoutons les covariables (fixes ainsi que variant en fonction du temps) pour mesurer les caractéristiques des personnes et leurs expériences à la maison et à l'école afin de déterminer le moment et la raison du désengagement scolaire de certains jeunes. Les données de l'étude proviennent de la Longitudinal Study of Young People in England (LSYPE), une étude par panel récente de 15 000 jeunes de 9 à 11 ans et au cours des premières années suivant la fin de la fréquentation obligatoire de l'école.

[1] Andy Ross, National Centre for Social Research, Royaume-Uni (Andy.Ross@natcen.ac.uk)

Les modèles d'appartenances mixtes longitudinaux des données d'enquête sur l'invalidité

Daniel Manrique-Vallier et Stephen E. Fienberg[1]

Résumé

Lorsqu'on analyse des données longitudinales, il est nécessaire de trouver un juste équilibre entre l'obtention de renseignements sur la variabilité individuelle et la production de résumés utiles et intelligibles sur les tendances globales au sein de la population. Et ce, surtout quand on sait que la façon dont les membres de la population cible évoluent au cours du temps est hétérogène à cause de traits individuels non observés. D'autres complications surgissent lorsque les données sont discrètes et multivariées, si bien que les tableaux de contingence résultant sont quasi vides. Nous proposons, pour analyser ce genre de données, une nouvelle famille de modèles combinant les caractéristiques d'une version du modèle transversal de degré d'appartenance (*Grade of Membership;* Erosheva et coll., 2007) et du modèle longitudinal de trajectoires latentes multivariées (*Multivariate Latent Trajectory Model;* Connor, 2006) que nous appliquons ensuite aux données de la National Long Term Care Survey (NLTCS), une enquête longitudinale dont six vagues ont été réalisées, afin d'évaluer l'état et les caractéristiques d'invalidité chez les citoyens américains de 65 ans et plus. Ces modèles sont fondés sur l'hypothèse qu'il existe un petit nombre de classes « typiques » ou « extrêmes » d'individus et modélisent leur évolution au cours du temps. Nous considérons que les individus appartiennent à toutes les classes à des degrés divers en les traitant comme des combinaisons pondérées convexes de ces classes extrêmes. Nous pouvons ainsi décrire des tendances générales distinctes (les cas extrêmes) tout en tenant compte de la variabilité individuelle. Nous proposons une spécification entièrement bayésienne et des méthodes d'estimation basées sur l'échantillonnage Monte Carlo par chaîne de Markov. Nous illustrons nos méthodes au moyen de données tirées de la NLTCS.

Mots-clés : modèle hiérarchique bayésien, modèle de degré d'appartenance, trajectoires latentes, Monte Carlo par chaîne de Markov.

1. Introduction

Dans le présent article, nous proposons des modèles et des méthodes d'estimation permettant de traiter des données longitudinales multivariées discrètes obtenues auprès d'une population hétérogène. L'étude est motivée par l'analyse de données provenant de la National Long Term Care Survey (NLTCS), une enquête par panel longitudinale dont le but est d'évaluer l'invalidité chronique chez la population âgée des États-Unis. Grâce à l'analyse des données de la NLTCS, les chercheurs tentent de répondre à des questions importantes concernant le processus de vieillissement et la prévalence de l'invalidité aux États-Unis, telles que : Combien d'Américains âgés présenteront une invalidité? Quelle est la durée des périodes d'invalidité? Quel est l'âge de la manifestation de l'invalidité? Change-t-il chez les générations plus jeunes? (Voir, p. ex., Connor et coll., (2006)). Les réponses à ces questions sont d'une grande importance pour l'élaboration des politiques publiques, entre autres parce que les dépenses publiques et privées engagées au bénéfice des personnes handicapées augmentent comparativement à celles visant leurs homologues valides (Manton et coll., 1997).

Nombre de questions pertinentes en matière de politique publique auxquelles la NLTCS pourrait éventuellement fournir une réponse ont trait aux changements au cours du temps, c'est-à-dire les changements durant la vie d'un individu (« Comment cette personne vieillira-t-elle vraisemblablement? ») ou à des comparaisons de personnes appartenant à diverses générations (« Les personnes des générations les plus récentes acquièrent-elles des incapacités différemment de celles nées 20 ans plus tôt? »). Afin de répondre à ces questions, nous devons examiner les données longitudinalement. En outre, on sait qu'aux États-Unis, les personnes âgées forment une population hétérogène et que l'on ne peut s'attendre à ce que tout le monde vieillisse de la même façon. Donc, les modèles applicables aux données longitudinales sur l'invalidité doivent pouvoir représenter ce genre d'hétérogénéité.

Dans le présent article, nous présentons des modèles et des méthodes d'analyse de ces données dans lesquelles nous cherchons à saisir la nature longitudinale de la NLTCS de même que la complexité liée à l'hétérogénéité du processus de vieillissement

[1] Daniel Manrique-Vallier, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. USA; Stephen E. Fienberg, Department of Statistics and Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213. USA.

humain en combinant les notions d'appartenance mixte du modèle de degré d'appartenance (Woodbury et coll., 1978; Erosheva et coll., 2007) et les descriptions longitudinales du processus de vieillissement issues de la famille de modèles de trajectoires latentes (Nagin 1999; Connor, 2006). Nous illustrons les méthodes au moyen de données provenant des six vagues de la NLTCS.

2. Données – La National Long Term Care Survey

La National Long Term Care Survey (NLTCS) est une enquête par panel longitudinale ayant pour but d'évaluer l'invalidité chronique chez la population âgée des États-Unis. La population cible englobe les personnes de 65 ans et plus présentant des limitations fonctionnelles d'une durée, observée ou prévue, d'au moins 90 jours (White, 2008). Jusqu'à présent, six vagues de l'enquête ont été réalisées, en 1982, 1984, 1989, 1994, 1999 et 2004.

La base de sondage de la NLTCS est le système de dossiers de Medicare, qui offre une bonne représentation de la population âgée des États-Unis, puisque près de 90 % des Américains de 65 ans et plus y sont inclus (Corder et Manton, 1991). Après un tirage initial, conformément à un plan de sondage complexe, chaque personne incluse dans l'échantillon fait l'objet d'une présélection afin de déterminer si elle présente une limitation fonctionnelle. Celles qui sont sélectionnées pour participer sont alors priées de répondre à un questionnaire détaillé, puis sont réinterviewées durant chaque cycle de l'enquête jusqu'à leur décès. Celles qui sont rejetées font de nouveau l'objet d'une présélection au moment des vagues subséquentes afin de vérifier si leur état fonctionnel a changé. À chaque vague, une nouvelle cohorte (environ 5 000 personnes) sont échantillonnées pour remplacer celles qui sont décédées, afin que la taille de l'échantillon soit de l'ordre de 20 000 personnes à chaque vague (Clark, 1998). Jusqu'à présent, 45 009 personnes ont été interviewées, toutes vagues confondues.

L'évaluation de l'invalidité dans le cadre de la NLTCS s'appuie sur la mesure de la capacité qu'a chaque personne d'exécuter un ensemble de six « activités de la vie quotidienne » (AVQ), telles que manger, se laver ou s'habiller, ainsi que dix « activités instrumentales de la vie quotidienne » (AIVQ), telles que préparer les repas ou s'occuper de ses affaires financières. De manière générale, les AVQ ont pour but d'évaluer la capacité qu'a une personne de prendre soin d'elle-même à un niveau fondamental, tandis que les AIVQ mesurent la capacité de vivre de façon autonome au sein de la collectivité (Connor, 2006). Le questionnaire de l'enquête permet d'enregistrer ces mesures sous forme d'une série de réponses à des questions déclencheuses qui sont ensuite résumées en un ensemble de réponses binaires. Ces réponses binaires indiquent la présence ou l'absence de déficiences en ce qui concerne l'exécution de ce genre d'activités.

3. Méthodes

Le but de notre analyse est de caractériser les trajectoires types d'acquisition d'une invalidité au cours du temps, tout en tenant compte de l'hétérogénéité de la population et en caractérisant celle-ci. Pour cela, nous avons combiné deux méthodes utilisées antérieurement.

La première méthode s'appuie sur le modèle des trajectoires latentes (Nagin, 1999), qui convient particulièrement bien dans les applications où le chercheur désire comprendre l'évolution type au cours du temps et soupçonne que la population est hétérogène, mais qu'un petit nombre de classes homogènes pourraient exister. Connor (2006) a adapté cette technique à l'analyse de données discrètes multivariées et l'a appliquée à l'analyse des données de la NLTCS en identifiant les courbes latentes de probabilité d'acquérir une invalidité au cours du temps. Cet outil fournit un moyen souple et facile d'interpréter la représentation des données qui tiennent compte de l'hétérogénéité latente de la population en regroupant cette dernière en classes *exclusives*. Comme l'a formulé Connor, cette hypothèse signifie essentiellement que, dans une classe, chaque individu suit exactement le même processus sous-jacent de vieillissement. Par conséquent, toute la variabilité de réponse dans la classe est attribuable à des fluctuations aléatoires dans cette classe, sans tenir compte du fait que les classes ainsi définies sont des classes idéales auxquelles il est fort possible qu'aucun individu réel n'appartienne vraiment (Kreuter et Muthén, 2008).

La famille de modèles de degré d'appartenance (*Grade of Membership* ou *GoM*) (Woodbury et coll., 1978; Erosheva et coll., 2007) offre un moyen conceptuellement intéressant de relâcher cette hypothèse. Au lieu de forcer chaque individu à appartenir à une seule classe, le modèle GoM vise à déterminer des *types purs* ou profils extrêmes, puis suppose que chaque individu appartient à plus d'un de ces types à des degrés divers. De cette façon, le modèle retient le pouvoir d'interprétation tenant à la spécification d'un nombre réduit de profils « types » ou « extrêmes », mais ajoute une certaine souplesse en ne supposant pas que l'appartenance à ces profils est exclusive.

3.1 Notation et configuration de l'étude

Nous utiliserons la notation et la structure qui suivent :

1. il existe N sujets dans l'échantillon, indicés par i, et N_i mesures pour chaque sujet $i \in \{1,...,N\}$;
2. pour chaque personne, nous mesurons J variables binaires simultanément à l'occasion de chaque mesure. Le vecteur de réponses manifestes pour l'individu i et la question j est $y_{ij*} = (y_{ij1},...,y_{ijN_i})$ [2];
3. chaque individu possède un vecteur de covariables connexes $X_{i.}$. Dans cette application, nous ne considérerons qu'un vecteur de covariables temporellement variantes $X_{i*} = (X_{i1},...,X_{iN_i})$, mais des covariables temporellement invariantes et d'autres structures plus complexes de covariables peuvent également être envisagées.

3.2 Un modèle de trajectoires multivariées avec degrés d'appartenance

Nous commençons par modéliser la distribution marginale de la réponse à la question $j \in \{1,...,J\}$ au temps de mesure $t \in \{1,...,N_t\}$, y_{ijt}, pour un membre à part entière du profil extrême k (c.-à-d. une personne i telle que $g_{ik}=1$ et $g_{ik'}=0$ pour $k' \neq k$) sous forme d'une fonction de certaines covariables enregistrées au temps t, X_{it},

$$Pr\left(Y_{ijt} = y_{ijt} \mid g_{ik} = 1, X_i\right) = f_{\theta_{j|k}}\left(y_{ijt} \mid X_{it}\right)$$

et nous modélisons la même distribution marginale de la réponse pour une personne générique caractérisée par le vecteur d'appartenances $G_{i*} = (g_{i1},...,g_{iK})$ sous forme de la combinaison convexe,

$$Pr\left(Y_{ijt} = y_{ijt} \mid G_{i*} = (g_{i1},...,g_{iK}), X_{i*}\right) = \sum_{k=1}^{K} g_{ik} f_{\theta_{j|k}}(y_{ijt} \mid X_{it}).$$

Maintenant, en supposant que, conditionnellement au vecteur d'appartenances, g_i, et aux covariables, X_{i*}, les réponses sont indépendantes entre items et entre mesures, nous avons

$$Pr\left(Y_{i**} = y_{i**} \mid G_{i*} = (g_{i1},...,g_{iK}), X_{it}\right) = \prod_{j=1}^{J} \prod_{t=1}^{N_i} \sum_{k=1}^{K} g_{ik} f_{\theta_{j|k}}(y_{ijt} \mid X_{it})$$

qui, combinée avec l'hypothèse d'échantillonnage aléatoire, nous donne le modèle conjoint

$$Pr\left(Y_{***} = y_{***} \mid g_{**}, X_{**}\right) = \prod_{i=1}^{N} \prod_{j=1}^{J} \prod_{t=1}^{N_i} \sum_{k=1}^{K} g_{ik} f_{\theta_{j|k}}(y_{ijt} \mid X_{it})$$

Ce modèle est similaire au modèle conjoint de classes et de trajectoires latentes proposé par Connor (2006), dans lequel nous généralisons la mise en grappes pour passer de l'appartenance à part entière (c.-à-d. $g_{ik} = 1$ pour un certain k) à l'appartenance mixte.

Comme l'a fait Connor (2006) dans son application, nous choisissons la fonction de répartition $f_{\theta_{j|k}}(y \mid X_i)$ pour la réponse unique de l'individu de type pur appartenant au profil extrême k comme étant $f_{\theta_{j|k}}(y_{ijt} \mid X_{i*}) = \lambda_{jt|k}(X_{it})^{y_{ijt}} \left(1 - \lambda_{jt|k}(X_{it})\right)^{1-y_{ijt}}$ avec $\lambda_{jt|k}(X_{it}) = logit^{-1}\left(\beta_{0j|k} + \beta_{1j|k} Age_{it}\right)$, où Age_{it} est l'âge de la i^e personne au temps de mesure t. Notons que, sous cette spécification, $\lambda_{jt|k}$ est réellement une fonction qui dépend du temps. Cette spécification a l'avantage d'être relativement simple, contenant seulement $2 \times J$ paramètres par profil extrême, et de représenter la notion intuitivement valable selon laquelle la probabilité sous-jacente d'invalidité est une fonction monotone (croissante) de l'âge. D'autres spécifications sont certainement possibles.

---

[2] Tous au long de l'article, nous utiliserons cette notation quand nous faisons référence au vecteur obtenu lorsque l'on fixe un sous-ensemble des sous-incides d'une variable indicée, pendant qu'on laisse varier les autres. Par exemple, si nous avons la série de variables scalaires λ_{jk} avec $j \in \{1,..,J\}$ et $k \in \{1,...,K\}$, nous pouvons écrire les vecteurs $\lambda_{*k} = (\lambda_{1k}, \lambda_{2k},..., \lambda_{Jk})$, et $\lambda_{j*} = (\lambda_{j1}, \lambda_{j2},..., \lambda_{jK})$.

Nous considérons que les N vecteurs d'appartenances, g_{i*}, sont des réalisations iid d'une loi commune, G_α, avec pour appui le simplexe unitaire à K–1 dimensions, Δ_{K-1}. À l'instar d'Erosheva et coll., (2007), nous modélisons cette loi comme étant $g_{i*} \mid \alpha \sim^{iid} Dirichlet(\alpha)$.

Dans la présente configuration, la loi de Dirichlet possède des propriétés importantes. En premier lieu, elle est conjuguée à la loi multinomiale, ce qui facilite grandement les calculs en utilisant les échantillonneurs de Gibbs; en deuxième lieu, en adoptant la reparamétrisation $\alpha = (\alpha_0 \cdot \xi_1, ..., \alpha_0 \cdot \xi_K)$ avec $\alpha_0 > 0$, $\xi_k > 0$ et $\sum_k \xi_k = 1,$ nous pouvons interpréter le vecteur ξ_* comme étant la proportion moyenne de la population dans le k^{e} profil extrême et α_0 comme étant un paramètre régissant l'étalement de la distribution : à mesure que α_0 tend vers 0, les échantillons issus de G_α sont de plus en plus concentrés vers les sommets de Δ_{K-1} et à mesure que α_0 augmente, ils sont de plus en plus concentrés à proximité de la moyenne arithmétique de la distribution.

Comme le mentionnent Erosheva et coll. (2007) et Airoldi et coll. (2007) dans leurs discussions, fixer a priori les paramètres α de la loi de Dirichlet pourrait représenter une hypothèse trop forte pour permettre une modélisation réaliste. Nous estimons ces paramètres d'après les données en spécifiant des hyperpriors et en calculant les distribution a posteriori. Pour cela, nous utilisons pour α_0 et ξ_* des hyperpriors semblables à ceux de Erosheva (2002) et d'Erosheva et coll. (2007), à savoir $\alpha \sim Gamma(\tau,\eta)$, $\xi \sim Dirichlet((1,...,1)_k)$ (loi uniforme sur Δ_{K-1}), et complétons la spécification par les priors $\beta_{0j|k} \sim^{iid} N(\mu_0, \sigma_0^2)$ et $\beta_{1j|k} \sim^{iid} N(\mu_1, \sigma_1^2)$, avec β_0 indépendant β_1.

4. Résultats

Pour tester nos méthodes, nous les avons appliquées à des données tirées des six vagues de la NLTCS (N=45 009), en analysant les réponses pour les six AVQ[3] (J=6) en nous servant d'un total de K=3 profils extrêmes. Nous avons calculé l'estimation a posteriori en utilisant un algorithme Monte Carlo par chaîne de Markov (MCMC) spécial basé sur une représentation des classes latentes proposée pour la première fois par Erosheva (2002) et par Erosheva et coll. (2008). Pour la présente application, nous avons choisi les priors $\beta_{0j|k}, \beta_{1j|k} \sim^{iid} N(0,100)$ et $\alpha_0 \sim Gamma(1,5)$.

La figure 4-1 donne les courbes de la probabilité a posteriori estimée d'avoir une invalidité pour chaque AVQ (quatre seulement sont présentées) en fonction de l'âge, pour les membres *idéaux* des trois profils extrêmes correspondants à ceux définis par Connor (2006). Ces profils semblent refléter certaines caractéristiques souhaitables : d'une part, les pentes de tous les profils extrêmes sont positives, montrant la tendance croissante prévue de la probabilité d'avoir une invalidité à mesure que le temps passe et, d'autre part, les âges où les individus idéalisés des profils extrêmes atteignent une probabilité de 0,5 d'avoir une invalidité se situent dans des fourchettes raisonnables. En outre, les distributions a posteriori des paramètres sont relativement concentrées autour d'une valeur centrale (données non présentées), par opposition à leur spécification a priori. L'examen de la figure 1 nous permet de constater que la méthode permet de distinguer trois profils biens distincts qui reflètent des processus de vieillissement assez différents, à savoir une catégorie de personnes qui vivent en relativement bonne santé jusqu'à un âge très avancé (k=1); une catégorie de personnes qui demeurent en bonne santé jusqu'à environ 85 ans, âge où elles connaissent un accroissement soudain de probabilité d'avoir des invalidités (k=2), et une catégorie de personnes pour lesquelles la probabilité d'avoir une invalidité augmente à un âge précoce (k=3).

Le tableau 4-1 donne les estimations ponctuelles (moyenne a posteriori) des paramètres α_0 et ξ, qui reflètent la façon dont des individus particuliers vieillissent par opposition aux individus idéaux. L'estimation du paramètre ξ_* = (0,65, 0,25, 0,1) indique l'ordre d'importance relatif de chacun des trois profils extrêmes, montrant que le nombre de personnes s'approchant du profil k=1 est le plus élevé, suivi du nombre pour le profil k=2, puis pour k=3. L'estimation du paramètre α_0= 0,264 indique que la distribution de la population est relativement concentrée vers les sommets du simplexe Δ_{K-1}, quoique pas suffisamment pour que le modèle se comporte comme un modèle de mélange de lois ordinaire. La figure 4-2 illustre comment les trajectoires individuelles sont formées à partir des profils extrêmes. Comme le montre son examen, le modèle est assez souple; il permet

[3] Les six AVQ sont manger (j=1), se mettre au lit et se lever (j=2), mobilité à l'intérieur (j=3), s'habiller (j=4), prendre son bain (j=5) et faire sa toilette (j=6).

des trajectoires individuelles relativement variées, mais n'extrait que quelques courbes extrêmes simples qui sont faciles à caractériser et à interpréter.

Figure 4-1 : Trajectoires de la probabilité d'avoir une invalidité pour les quatre premières AVQ chez les membres idéaux de chaque profil extrême (*K*=3 profils extrêmes) au cours du temps. Le point d'intersection des droites est celui où la probabilité d'avoir une invalidité pour l'AVQ *j* atteint 50 %.

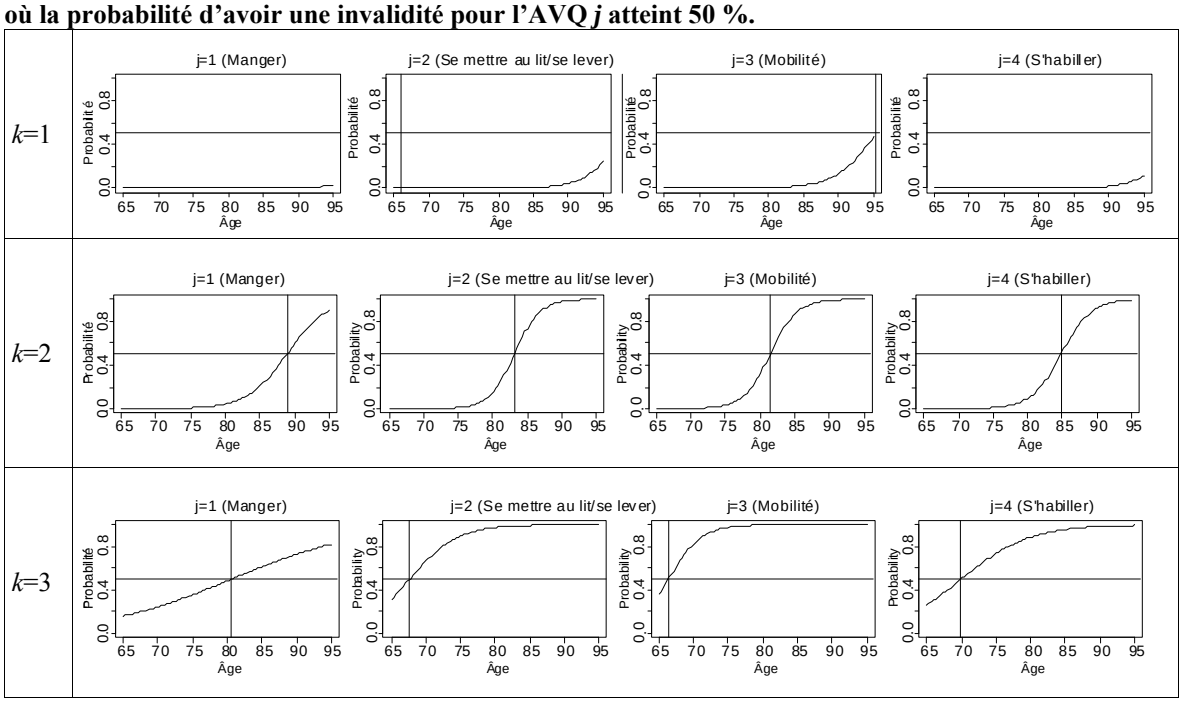


Figure 4-2 : Exemple de mélange au niveau individuel de trajectoires pour l'AVQ *j*=2 (se mettre au lit ou se lever). Dans le graphique de droite, les trois trajectoires extrêmes sont superposées et un échantillon de 100 trajectoires individuelles est présenté.

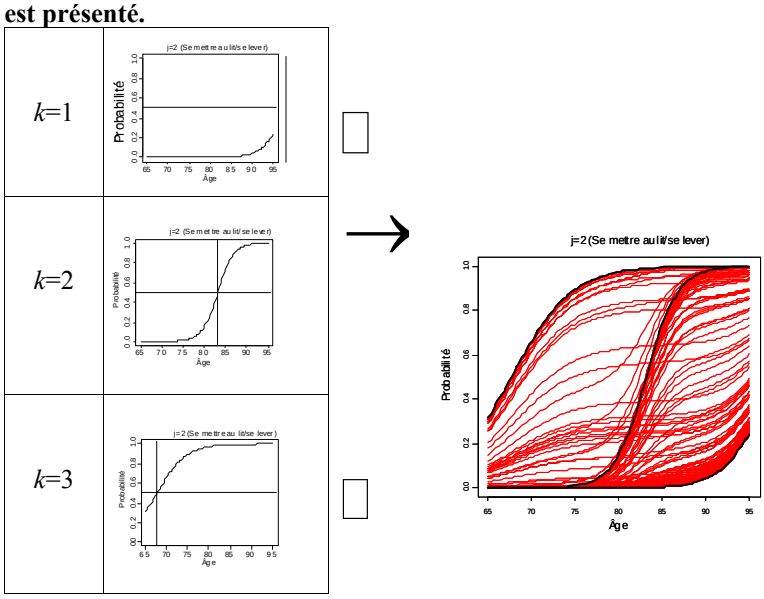


Tableau4-1 : Moyennes a posteriori des paramètres au niveau de la population pour le modèle avec K=3 profils extrêmes. Les chiffres entre crochets sont les écarts-type a posteriori

Paramètre	**Estimation [é-t]**
α_0	0,264 [0,00489]
(ξ_1, ξ_2, $_1\xi_3$)	(0,65 [0,004], 0,25 [0,003], 0,104 [0,002])

5. Discussion

Nos résultats préliminaires sont intéressants, parce qu'ils montrent les possibilités de nos méthodes. Dans notre application aux données longitudinales de la NLTCS, nous avons pu caractériser des profils extrêmes bien distincts et intuitivement valables qui peuvent être interprétés comme des modes types de vieillissement en même temps que l'hétérogénéité de la population à l'aide d'un outil très simple qui permet de construire des courbes individualisées à partir des profils extrêmes. Quoiqu'un peu plus compliquée que celle proposée par Connor (2006), cette façon de traiter l'hétérogénéité nous permet de maintenir les profils extrêmes peu nombreux et interprétables, tout en évitant d'introduire des hypothèses d'homogénéité intra-classe trop fortes.

Le modèle présenté ici représente une application élémentaire de l'idée générale consistant à combiner des trajectoires variant en fonction du temps complètes en utilisant un mécanisme d'appartenance mixte. Selon le problème étudié, un certain nombre d'extensions évidentes peuvent être exécutées, dont certaines sur lesquelles nous nous penchons à l'heure actuelle. Dans notre application à la NTLCS, certaines de ces extensions naturelles sont l'inclusion d'autres covariables au niveau de l'appartenance au groupe et au niveau du profil extrême, ainsi que la formulation conjointe avec des modèles de survie pour étudier la relation entre l'invalidité et la mortalité. Nombre de ces extensions seront incluses dans Manrique-Vallier (2010).

En guise d'illustration, nous avons choisi d'appliquer la méthode en utilisant K=3 profils extrêmes. De manière plus générale, nous devons intégrer une méthode permettant de choisir la valeur optimale de K. Nous avons effectué les calculs complets pour une série de valeurs de K allant de 2 à 5. Bien que la qualité de l'ajustement du modèle, mesurée par les réponses prédictives a posteriori, augmente avec K, nous avons constaté une moins bonne séparation des profils pour K=4 et K=5, et une interprétation moins satisfaisante de la forme et de la structure des profils. Le choix de la valeur appropriée de K reste une question ouverte qui sera abordée dans les travaux exposés dans Manrique-Vallier (2010).

6. Bibliographie

Airoldi, E., Fienberg, S., Joutard, C. et Love, T. (2007). Discovering Latent Patterns with Hierarchical Bayesian Mixed-Membership Models, *Data Mining Patterns: New Methods and Applications*, 240–275.

Clark, R. (1998). An Introduction to the National Long-Term Care Survey, Office of Disability, Agigng, and Long-Term Care Policy within the U.S. Department of Health and Human Services.

Connor, J., Fienberg, S., Erosheva, A. et White, T. (2006). Towards a restructuring of the national long term care survey: A longitudinal perspective, Tech. rep.

Connor, J. T. (2006). Multivariate Mixture Models to Describe Longitudinal Patterns of Frailty in American Seniors, Ph.D. thesis, Department of Statistics & H. John Heinz III School of Public Policy & Management. Carnegie Mellon University.

Corder, L. et Manton, K. (1991). National surveys and the health and functioning of the elderly: the effects of design and content, *Journal of the American Statistical Association*, 86, 513–525.

Erosheva, E. (2002). Grade of membership and latent structures with application to disability survey data, Ph.D. thesis, Department of Statistics. Carnegie Mellon University.

Erosheva, E., Fienberg, S. et Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data, *Annals of Applied Statistics*, 1, 502–537.

Kreuter, F. et Muthén, B. (2008). Analyzing criminal trajectory profiles: Bridging multilevel and group-based approaches using growth mixture modeling, *J Quant Criminol*, 24, 1–31.

Manrique-Vallier, D. (2010). Longitudinal Mixed Membership Models With Applications, Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University, expected May 2010.

Manton, K., Corder, L. et Stallard, E. (1997). Chronic disability trends in elderly United States populations: 1982-1994, *Proceedings of the National Academy of Sciences*, 94, 2593–2598.

Nagin, D. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach, *Psychological Methods*, 4, 139–157.

White, T. (2008). Extensions of Latent Class Transition Models with Application to Chronic Disability Survey Data, Ph.D. thesis, University of Washington.

Woodbury, M., Clive, J. et Garson Jr, A. (1978). Mathematical typology: A grade of membership technique for obtaining disease definition. *Computers in Biomedical Research*, 11, 277–98.

SESSION 17 – ERREURS DE MESURE

La non-réponse et l'erreur de mesure dans la recherche sur l'emploi

Frauke Kreuter, Gerrit Mueller et Mark Trappmann[1]

Résumé

Les méthodologistes d'enquête se préoccupent de plus en plus de l'interaction de plusieurs sources d'erreur, notamment de la non-réponse et de l'erreur de mesure. Selon une hypothèse souvent formulée, les membres de l'échantillon qui ne répondent à l'enquête qu'après le recours à des tentatives répétées et à d'autres stratégies de recrutement sont plus susceptibles de fournir des données de faible qualité (voir Groves et Couper, 1998). On évalue souvent la qualité des données à l'interne d'après les proportions de réponses manquantes, de « ne sait pas », etc. (voir Fricker, 2007). Dans ces études, on dispose rarement de données externes pour évaluer la qualité des réponses des répondants (voir Cannell et Fowler, 1963; Olsen, 2006).

L'étude par panel PASS (Trappmann et coll., 2009) est un ensemble de données de recherche original sur le marché du travail, l'État-providence et la pauvreté en Allemagne. Menée auprès de quelque 19 000 personnes interviewées dans plus de 12 500 ménages, la PASS est à l'heure actuelle l'une des enquêtes par panel les plus exhaustives réalisées en Allemagne. Le premier cycle de collecte de données a commencé en 2006. Les données d'enquête de la PASS sur les antécédents d'emploi et de chômage, le revenu et la scolarité des participants peuvent être couplées aux données correspondantes tirées des dossiers administratifs des répondants.

En nous fondant sur cette étude, nous évaluons la qualité des données en fonction de la possibilité de prendre contact et de la propension à répondre. Dans le cas de certaines variables seulement, l'erreur de mesure (variance ou biais) évaluée au moyen des dossiers administratifs est inversement proportionnelle à la possibilité de prendre contact et à la propension des personnes cibles à répondre. On le constate en particulier dans le cas des questions rétrospectives. Ici, l'écart de durée entre la date de l'interview et celle de l'événement explique en grande partie l'écart dans l'erreur de mesure selon que la propension des répondants à répondre est élevée ou faible.

[1] Frauke Kreuter, JPSM University of Maryland, É.-U.; Gerrit Mueller (Gerrit.Mueller@iab.de) et Mark Trappmann, IAB Institute for Employment Research, Allemagne.

Incohérences dans les caractéristiques d'emploi déclarées par les salariés qui conservent leur emploi : données tirées d'une série de panels à deux cycles de l'enquête italienne sur la population active, 1993-2003

Francesca Bassi, Ugo Trivellato et Alessandra Padoan[1]

Résumé

Dans cet exposé, nous abordons l'erreur de mesure et la distorsion qu'elle risque d'entraîner dans les renseignements sur le statut des personnes à l'égard de la branche d'activité et de la profession. Notre étude de cas porte sur des panels à deux cycles de données recueillies à intervalles d'un an par l'enquête trimestrielle italienne sur la population active pendant la période d'avril 1993 à avril 2003. Nous analysons l'incohérence des renseignements sur les caractéristiques de l'emploi (statut à l'égard de la branche d'activité et de la profession) tirés des matrices de transition annuelles des travailleurs ayant déclaré qu'ils avaient été occupés pendant toute l'année et qu'ils n'avaient pas changé d'emploi.

Nous commençons par calculer et commenter des indicateurs habituels de disparité. Nous constatons qu'il existe manifestement une erreur de mesure appréciable en ce qui concerne le statut à l'égard de la branche d'activité et de la profession. Puis, nous tentons de déterminer si la répétition des renseignements accroît la cohérence de façon significative lorsqu'on agrège le nombre de catégories. L'agrégation des catégories améliore la concordance. Dans le cas du statut professionnel, le meilleur niveau d'agrégation est le niveau binaire (salarié ou travailleur indépendant); dans le cas de la branche d'activité, deux classifications comportant cinq ou six catégories réduisent les incohérences au minimum. Nous approfondissons notre étude des profils d'incohérence entre les catégories de variables en mettant à l'essai plusieurs spécifications du modèle de quasi-indépendance de Goodman. Le modèle est presque toujours rejeté, ce qui prouve que même des renseignements transversaux risquent d'être faussés par l'erreur de mesure non aléatoire. Enfin, nous considérons et comparons d'autres classifications à quatre catégories obtenues en agrégeant le statut professionnel et la branche d'activité en une seule variable. Dans ce cas, le meilleur niveau d'agrégation est donné par une classification non standard à quatre catégories qui distingue, d'une part, les employés des services marchands et, d'autre part, ceux du secteur industriel et des services privés.

[1] Francesca Bassi and Ugo Trivellato (trivell@stat.unipd.it), University of Padova, Italie; Alessandra Padoan, ISTAT, Italie.

Le chevauchement des lisières dans l'enquête HILDA : problèmes et observations

Nicole Watson[1]

Résumé

Les effets de lisière sont particuliers aux enquêtes longitudinales. Ils se produisent lorsque, d'une interview à l'autre, les variations des données ont tendance à se concentrer inhabituellement dans des périodes adjacentes. L'une des composantes de la Household, Income and Labour Dynamics in Australia (HILDA) Survey qui est sujette à des effets de lisière est le calendrier de l'activité sur le marché du travail dans lequel on demande aux répondants de se rappeler les divers emplois qu'ils ont occupés au cours d'une période 14 à 18 mois. Comme les interviews sont menées annuellement, il se produit un chevauchement de deux à six mois dont nous pouvons tirer parti pour étudier les causes des erreurs de datation des événements, comme la durée et le type de la période, la durée du chevauchement des lisières, la capacité de remémoration du répondant et les caractéristiques de l'interview qui peuvent avoir une incidence sur la remémoration. Le chevauchement des lisières permet aussi d'étudier l'erreur de mesure au cours du temps, afin de déterminer si les mêmes personnes commettent continuellement les mêmes erreurs.

Mots clés : erreur de mesure, calendrier de l'emploi, qualité des données, enquêtes longitudinales.

1. Introduction

Les études longitudinales contiennent souvent des questions qui demandent au répondant de déclarer des activités qui ont eu lieu au cours de la période entre l'interview courante et l'interview précédente. Les incohérences entre les données dont ils se souviennent et celles recueillies à l'interview précédente donnent lieu à des effets de chevauchement des lisières. Ceux-ci se produisent lorsque, d'une interview à l'autre, les variations des données ont tendance à se concentrer inhabituellement dans des périodes adjacentes (Tourangeau et coll., 2000). Il n'est pas infréquent que la grandeur des transitions entre les périodes adjacentes (la lisière) soit de deux à huit fois supérieure à la grandeur des transitions qui ne se produisent pas à la lisière (par exemple, Burkhead et Coder, 1985; Tourangeau et coll., 2000; Lemaître, 1992).

Le présent article vise à mieux faire comprendre pourquoi les effets de lisière se produisent en se servant de données provenant du calendrier de l'activité sur le marché du travail de l'enquête HILDA. Le calendrier est conçu pour recueillir des renseignements à partir d'un point fixe l'année qui a précédé la date de l'interview, ce qui donne lieu à deux déclarations pour la même tranche de temps (phénomène appelé « chevauchement des lisières »). Ce chevauchement des lisières nous donne la possibilité d'étudier un certain nombre de méthodes d'appariement et d'examiner les différences entre les deux déclarations.

2. Facteurs produisant des déclarations de période non cohérentes

Au cours d'une interview, un certain nombre d'erreurs peuvent aboutir à des déclarations non cohérentes. Ces erreurs comprennent l'omission de certaines périodes, la classification incorrecte des périodes, l'avancement ou le recul de la période dans le temps (télescopage), surtout au début de la période de référence, la mauvaise compréhension de ce qui est demandé ou les inférences incorrectes quand il n'est pas possible de se remémorer clairement la période. Ces erreurs peuvent avoir lieu au moment de la première ou de la deuxième déclaration de la période, mais nous nous attendons à ce qu'elles soient plus nombreuses pour la deuxième déclaration, car, dans ce cas, la remémoration se fait à un moment plus éloigné de la période en question.

Les facteurs qui ont une incidence sur ces erreurs de remémoration des périodes peuvent être répartis en cinq catégories :

- *Caractéristiques de la période* – Les périodes plus courtes sont plus susceptibles d'être omises ou mal situées (Paull, 2002; Jäckle, 2008). Les événements dont les dates sont clairement définies sont déclarés avec plus d'exactitude que ceux dont les dates sont floues, tels que les transitions hors de la population active (Jäckle, 2008).

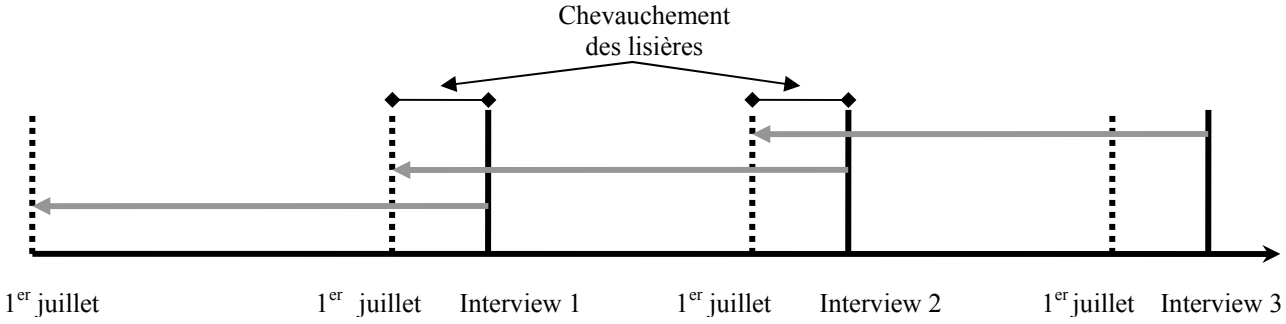
[1] Nicole Watson, Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Vic 3010, Australia (n.watson@unimelb.edu.au).

- *Complexité de la remémoration* – La difficulté de la tâche de remémoration et l'exactitude des stratégies d'inférence que les répondants utilisent quand ils ne peuvent pas se rappeler précisément ce qu'on leur demande ont une incidence sur la grandeur de l'effet de lisière (Callegaro et Belli, 2007).
- *Caractéristiques du répondant* – Même si les résultats des diverses études sont mixtes, la probabilité que des déclarations ne concordent pas a tendance à varier avec l'âge, le sexe et le niveau de scolarité (Hill, 1987; Jäckle et Lynn, 2007; Hill, 1987; Callegaro et Belli, 2007; Jäckle, 2008).
- *Caractéristiques de l'interview* – L'utilisation du même intervieweur aux diverses vagues peut réduire le nombre de déclarations non concordantes, grâce à l'application cohérente des techniques d'interview (Vick et Weidman, 1989) ou parce que la présence même de l'intervieweur déclenche la remémoration du répondant. La qualité des réponses du répondant dépend aussi de sa compréhension des questions et de sa volonté à puiser dans sa mémoire et à interpréter l'information requise (Tourangeau et coll., 2000; Burkhead et Coder, 1985).
- *Caractéristiques du processus d'enquête* – Les effets de lisière peuvent également résulter de changements ou d'erreurs dans le processus d'enquête. La forme, l'énoncé et l'ordre de la question jouent un rôle important dans la cohérence de l'interprétation au cours du temps (Burkhead et Coder, 1985; Jäckle, 2008; Callegaro et Belli, 2007). Des erreurs de transcription ou de saisie des données peuvent également se produire (Burkhead et Coder, 1985; Lemaître, 1992).

3. Le calendrier de HILDA

Nous examinons maintenant les incohérences relevées dans le calendrier des activités sur le marché du travail de l'enquête HILDA au cours du temps[2]. L'enquête HILDA est une enquête-ménages par panel à l'échelle de l'Australie dans le cadre de laquelle environ 7 000 ménages et 13 000 particuliers sont interviewés chaque année (Watson et Wooden, 2004). L'enquête a débuté en 2001 et, à chaque vague, on demande aux répondants de se rappeler les divers emplois qu'ils ont eus, le temps passé au chômage et le temps passé hors de la population active. En outre, des renseignements sont recueillis sur les périodes d'études à temps plein et à temps partiel (quoique nous ne discutions pas de celles-ci dans le présent article par souci de concision). Les dates de début et de fin de chaque période sont enregistrées en tenant compte du fait qu'elles se situent au début, au milieu ou à la fin de chaque mois. Le calendrier couvre une période de 12 à 18 mois, allant du 1er juillet de l'année précédente jusqu'à la date de l'interview courante (voir la figure 3-1)[3]. Toutes les périodes d'emploi sont incluses dans le calendrier (et pas simplement celles correspondant à l'emploi principal). Quand deux emplois ou plus ont lieu au tout début du calendrier, leurs dates de début sont également recueillies afin de faciliter l'appariement des données entre les vagues de l'enquête. Comme le calendrier est rempli à chaque vague, un chevauchement de deux à six mois se produit, selon le moment où le répondant a été interviewé (la moyenne étant de 2,9 mois).

Figure 3-1
Chevauchement des lisières dans HILDA



4. Méthodes d'appariement

Afin d'étudier la valeur du chevauchement des lisières, nous évaluons trois méthodes d'appariement :

i) appariement des emplois au 1er juillet;
ii) appariement des emplois à la date de la dernière interview (en accordant la priorité à l'information recueillie au moment le plus rapproché de la période remémorée);

[2] L'ensemble de données utilisé pour l'analyse présentée ici est l'ensemble de données nommé Confidence Release 7 HILDA.
[3] Les interviews ont lieu du milieu d'août à mars, quoique la dernière période du calendrier soit en décembre (96 % à 98 % des interviews sont alors achevées).

iii) rapprochement des périodes déclarées entre les deux versions des événements produites à environ un an d'intervalle.

Dans les deux premières méthodes d'appariement, la déclaration alternative d'une même période est ignorée. Lorsque plusieurs périodes déclarées durant une vague peuvent être appariées à une période d'une autre vague (comme cela se produit pour les périodes d'emploi), un appariement est choisi au hasard. L'appariement des périodes est effectué pour chaque type de période (emploi, chômage ou inactivité).

La troisième méthode d'appariement s'appuie sur les deux déclarations pour produire une version rapprochée de la période de chevauchement. La première déclaration (remémoration la plus rapprochée du moment où ont eu lieu les événements) a précédence sur la deuxième. Néanmoins, les périodes enregistrées dans la deuxième déclaration sont appariées à celles du même type figurant dans la première déclaration, afin de repérer (et d'éliminer) les périodes qui ont été mal situées. Cette méthode ne permet pas de résoudre les situations où une période mentionnée dans la première déclaration n'est pas remémorée dans la seconde ou inversement. Un score d'appariement a été créé pour déterminer quelles périodes devraient être appariées. Le score d'appariement est égal à la somme des cas où les débuts de période concordent exactement à l'intérieur du chevauchement des lisières, les débuts de période concordent à un mois près, les fins de période concordent exactement, les fins de période concordent à un mois près et trois fois le rapport de durée des périodes (en prenant la plus longue comme dénominateur). Le score d'appariement maximal est, par conséquent, égal à 7. Les périodes obtenant le score d'appariement le plus élevé sont appariées les premières. Les périodes restantes sont ensuite appariées de la même façon jusqu'à ce qu'il ne reste plus aucune période pouvant être appariée. Les périodes dont le score d'appariement est inférieur à 3 ne sont pas appariées.

5. Résultats

5.1 Grandeur de l'effet de lisière

Le nombre de transitions d'emploi à la lisière pour les périodes appariées au début de juillet est environ huit fois plus élevé que pour les transitions hors de la lisière. Pour les deux autres méthodes, la lisière se situe à la date d'interview de chaque répondant, donc s'étale de la mi-août à décembre, ce qui donne lieu à une augmentation du nombre de transitions d'emploi au cours de cette période. La figure 5-1 donne le nombre (non pondéré) de débuts d'emploi centré à la lisière pour les personnes interviewées à chacune des sept vagues. La ligne gris clair montre les résultats quand les périodes sont appariées à la date de la dernière interview (méthode DDI) et la ligne gris moyen (à peine visible) donne les résultats quand les périodes sont rapprochées (méthode CHEV, qui en majeure partie, se comporte de manière fort semblable à la méthode DDI). En outre, par souci de clarté, les nombres de débuts (et de fins) d'emploi à la lisière sont également présentés pour chaque méthode dans le tableau 5-1.

Plusieurs observations se dégagent de la figure et du tableau :

- les répondants ont tendance à déclarer un moins grand nombre de débuts d'emploi au commencement du calendrier (ce que montrent les résultats pour la méthode de début de l'année financière (DAF) pour les quelque premiers mois après la lisière) et, par conséquent, la méthode DDI donne un nombre légèrement plus élevé de débuts d'emploi à la lisière que la méthode DAF. Ces résultats suggèrent soit une diminution de la mémoire au cours du temps ou un télescopage arrière des événements à la frontière entre les périodes de référence (un effet semblable a été observé par Kalton et Miller (1991) dans une analyse de la Survey of Income and Program Participation des États-Unis). Quand les périodes sont tirées de la première déclaration des événements faite par le répondant (comme dans les méthodes DDI et CHEV), cette différence est éliminée.
- La méthode CHEV produit un plus grand nombre de débuts d'emploi à la lisère que la méthode DDI. Lorsque l'on compare les périodes dans les deux versions du chevauchement des lisières, il est évident que certaines périodes appariées dans la méthode DDI ne sont en réalité pas les mêmes, si bien qu'elles ne sont pas reliées dans la méthode CHEV, ce qui donne un plus grand nombre de fins d'emploi et de débuts d'emploi à la lisière.

Figure 5-1
Nombre de débuts d'emploi, centré à la lisière entre chaque vague

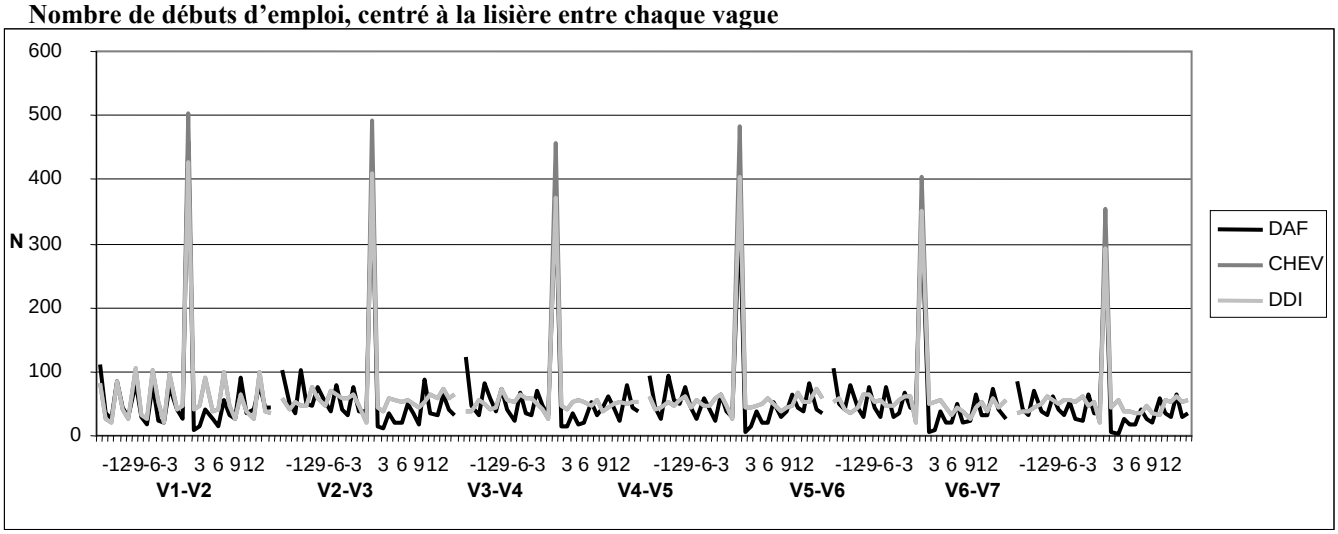


Tableau 5-1
Nombre de débuts et de fins d'emploi à la lisière, selon la vague et la méthode

	Débuts			*Fins*		
	Méthode DAF	*Méthode DDI*	*Méthode CHEV*	*Méthode DAF*	*Méthode DDI*	*Méthode CHEV*
Lisière V1-V2	459	426	502	485	432	526
Lisière V2-V3	407	410	493	465	416	512
Lisière V3-V4	368	372	458	467	387	474
Lisière V4-V5	347	405	482	437	389	469
Lisière V5-V6	342	350	404	512	423	495
Lisière V6-V7	298	293	353	480	421	495
Total	2 221	2 256	2 692	2 846	2 468	2 971

- Le nombre de débuts d'emploi à la lisière semble diminuer au cours du temps, mais il n'existe aucune baisse correspondante du nombre de fins d'emploi à la lisière. La cause de cette situation n'est pas claire.

Il est évident que les méthodes DDI et CHEV sont meilleures que la méthode DAF, car elles permettent d'éviter la baisse du nombre de périodes qui débutent ou se terminent au commencement de la période de référence. La méthode CHEV est supérieure à la méthode DDI, car elle aide à repérer les périodes qui ne devraient pas être appariées durant la période de chevauchement.

5.2 Caractéristiques associées aux erreurs de remémoration

La période de chevauchement est utilisée pour repérer les périodes omises (dans la première déclaration mais non dans la seconde), les périodes mal situées et les périodes ajoutées (qui figurent dans la deuxième déclaration mais non dans la première). L'ensemble de données au niveau de la période couvre les sept premières vagues de l'enquête HILDA et est limité aux personnes interviewées à chacune des vagues. Seules les périodes figurant dans les six chevauchements de lisières sont incluses dans l'analyse. Trois modèles de régression logistique distincts sont utilisés pour estimer la probabilité qu'une période soit omise, mal située ou ajoutée. L'ensemble de données groupées de la première déclaration contient 64 085 périodes ayant trait à 8 409 personnes et est utilisé pour étudier la probabilité d'omettre ou de mal situer des périodes. Un deuxième ensemble de données groupées ayant trait à la deuxième déclaration contient 59 981 périodes se rapportant à 8 409 personnes et est utilisé pour estimer la probabilité d'ajouter des périodes. Afin de tenir compte des mesures répétées sur les mêmes personnes, l'estimation des erreurs-types s'appuie sur l'hypothèse que les résultats des périodes sont corrélés entre les observations faites chez une même personne, mais sont indépendantes entre les personnes.

Comme à la section 2, nous supposons que la probabilité qu'une période soit omise, mal située ou ajoutée est une fonction de cinq groupes de caractéristiques. Les *caractéristiques de la période* comprennent la durée de la période et le fait que le type de période était le même à la première et à la deuxième interview. La *complexité de la remémoration* est

estimée par le nombre de périodes entre les dates d'interview et l'intervalle de temps entre les interviews. La longueur du chevauchement est incluse comme variable de contrôle. Les *caractéristiques du répondant* comprennent le sexe, l'âge (15 à 24 ans, 35 à 54 ans, 55 ans et plus), le niveau de scolarité, le revenu personnel durant l'année financière et le pays de naissance. Les *caractéristiques de l'interview* comprennent l'évaluation faite par l'intervieweur de la compréhension des questions par le répondant, le fait qu'il s'agissait ou non d'une interview téléphonique, le fait que le même intervieweur a effectué ou non les deux interviews et l'expérience de l'intervieweur (mesurée par le nombre d'interviews achevées durant la vague en question). Les *caractéristiques du processus d'enquête* comprennent le fait que la lisière était comprise ou non entre les vagues 1 et 2 (car une modification de la conception du calendrier a eu lieu entre ces vagues).

Les effets marginaux moyens estimés en se servant des modèles sont présentés au tableau 5-2 et donnent lieu aux observations suivantes :

i) Le facteur jouant le rôle le plus important dans la réduction de la probabilité de produire une déclaration non concordante est le fait que le même type de période avait lieu ou non au moment de la deuxième interview. Pour les périodes d'emploi, ce facteur réduit le risque d'omettre une période de 23 %, le risque de mal situer une période de 2 % et le risque d'ajouter une période de 10 % (toutes choses étant égales par ailleurs).

ii) Fait qui corrobore les travaux de recherche antérieurs, les périodes plus courtes sont plus susceptibles d'être omises, ajoutées ou mal situées.

iii) En général, plus la tâche de remémoration est complexe, plus le risque d'une déclaration non concordante est grand. Contrairement aux attentes, nous constatons que ce sont les périodes de chômage et non les périodes d'activité qui sont moins susceptibles d'être omises si de nombreuses périodes doivent être remémorées entre les dates d'interview.

iv) Certains effets d'âge et de sexe existent – les déclarations non concordantes sont plus probables pour des périodes inhabituelles, compte tenu de l'étape du cycle de vie atteinte par le répondant. Le niveau de scolarité et le pays de naissance n'ont souvent aucun effet.

v) Le revenu individuel a un effet mixte. Les personnes dont le revenu individuel est élevé sont moins susceptibles d'omettre ou d'ajouter des périodes d'emploi, mais plus susceptibles de le faire pour des périodes d'inactivité.

vi) Les résultats offrent un certain appui pour l'interview sur place et l'utilisation du même intervieweur en vue de réduire la probabilité que des déclarations non concordantes aient lieu. L'expérience de l'intervieweur n'a pas d'effet significatif.

vii) Les répondants qui semblent bien comprendre les questions posées pendant toute l'interview sont moins susceptibles que les autres de produire des déclarations non concordantes.

viii) Étonnamment, nous constatons que les périodes d'emploi étaient *moins* susceptibles d'être omises ou mal situées durant la période de chevauchement entre les vagues 1 et 2. La raison pour laquelle il en est ainsi n'est pas claire.

Les modèles susmentionnés ne nous permettent pas de déterminer si les répondants commettent les mêmes erreurs au fil du temps. Pour le vérifier, nous avons limité l'ensemble de données groupées à celles pour les chevauchements des lisières allant des vagues 2 et 3 au vagues 6 et 7, ainsi qu'une variable supplémentaire indiquant si le répondant a commis la même erreur (omettre, mal situer ou ajouter une période) au chevauchement des lisières précédent. Nous constatons que l'omission d'une période au chevauchement précédent augmente le risque d'omettre une période au chevauchement suivant — pour les périodes d'emploi, le risque d'omettre une période augmentait de 2 %, pour les périodes de chômage, il augmentait de 5 % et pour les périodes d'inactivité, il augmentait de 2 %. De même, si une période est ajoutée au chevauchement précédent, le risque d'ajouter une période au chevauchement suivant est plus élevé pour la plupart des types de périodes, mais mal situer une période durant un chevauchement ne semble pas influencer le fait que des périodes sont mal situées ou non à la période de chevauchement suivante.

Tableau 5-2
Effets marginaux moyens des caractéristiques associées à des déclarations non concordantes pour les périodes d'activité

	Périodes d'emploi			*Périodes de chômage*			*Périodes d'inactivité*		
Variable	*Omises*	*Mal situées*	*Ajoutées*	*Omises*	*Mal situées*	*Ajoutées*	*Omises*	*Mal situées*	*Ajoutées*
Caractéristiques des périodes									
Durée de la période (tiers du mois)	-0,003***	-0,003***	-0,001***	-0,005***	-0,006***	0,001	-0,002***	-0,002***	-0,001***
Même type de période à la date d'interview	-0,231***	-0,018***	-0,105***	-0,343***	0,051**	-0,152***	-0,476***	0,005*	-0,322***
Complexité de la remémoration									
Nombre de périodes entre les interviews	0,004***	0,003***	0,018***	-0,032***	0,010*	0,021*	-0,014***	0,003***	-0,002
Intervalle entre les dates d'interview précédente et courante	0,002***	-0,001**	0,001	-0,002	-0,002	0,003	0,001	0,000	0,002*
Durée du chevauchement	0,005***	0,004***	0,003***	0,001	0,005	0,018**	0,002	0,003***	0,003***
Caractéristiques du répondant									
Sexe et âge (base = jeunes hommes de 15 à 24 ans)									
Hommes dans la force de l'âge (25 à 54 ans)	-0,011*	0,009**	-0,004	-0,013	-0,005	-0,023	0,020	-0,002	-0,042
Hommes plus âgés (55 ans et plus)	0,019*	0,000	0,029***	0,084*	-0,045	0,021	-0,005	0,000	-0,069***
Jeunes femmes	0,007	0,009	-0,016	0,081*	-0,029	0,067	-0,010	0,030	-0,029
Femmes dans la force de l'âge	-0,005	0,003	0,007	0,066*	-0,006	0,071	-0,011	0,015	-0,086***
Femmes plus âgées	0,011	0,011	0,034***	0,183***	-0,055*	0,182***	-0,008	0,001	-0,070***
Niveau de scolarité (base = 11 années d'études ou moins)									
12e année	0,002	0,001	0,000	0,053	-0,018	0,054	0,003	0,003	0,002
Certificat	0,009*	0,001	-0,005	0,022	-0,032*	0,013	0,004	0,007***	0,009*
Diplôme	0,007	-0,003	-0,002	0,054	-0,045*	0,062	-0,001	0,001	-0,003
Diplôme d'études supérieures	0,015***	0,000	-0,001	-0,010	-0,021	0,037	0,003	0,001	0,011**
Revenu de l'année financière ($/10^5$)	-0,039***	0,005	-0,034***	-0,206	0,109	0,043	0,062***	-0,001	0,069***
Carré du revenu de l'année financière ($/10^{10}$)	0,004**	-0,002	0,003***	0,065	-0,037	0,008	-0,010***	-0,004	-0,015*
Pays de naissance (base = Australie)									
Pays anglophone principal	-0,009*	-0,001	-0,005	0,007	0,026	0,034	0,001	0,002	0,001
Autre pays	0,009	-0,005	0,007	0,004	0,014	-0,024	-0,003	0,006	0,000
Caractéristiques de l'interview									
Compréhension des questions, excellente ou bonne	-0,014	0,002	-0,028***	0,011	-0,017	0,001	-0,020***	0,004	0,003
Interview téléphonique	0,013*	-0,001	0,000	0,026	0,022	-0,024	0,008	-0,005	0,022*
Intervieweur, continuité	-0,007**	-0,004*	-0,002	-0,018	0,000	-0,030	-0,003	0,000	-0,009**
Expérience de l'intervieweur (calendriers achevés) ($/10^2$)	-0,003	0,003	-0,004	0,018	-0,002	-0,030	0,002	-0,001	-0,007
Caractéristiques du processus d'enquête									
Lisière entre les vagues 1 et 2	-0,012**	-0,010***	0,007	-0,046	0,001	0,067*	0,001	0,003	0,001
Pseudo R^2	0,134	0,357	0,120	0,123	0,137	0,039	0,518	0,362	0,355
Nombre de périodes	36 383	36 383	34 529	1 771	1 771	1 412	16 843	16 843	16 744
Proportion de périodes non concordantes	0,119	0,049	0,080	0,644	0,119	0,495	0,117	0,029	0,097

Nota : * significatif au seuil de 10 %, ** significatif au seuil de 5 %, *** significatif au seuil de 1 %.

6. Conclusion

Les données sur le chevauchement de lisières recueillies dans le calendrier de l'activité sur le marché du travail de l'enquête HILDA n'ont pas permis de résoudre le problème de l'effet de lisière. Si l'on apparie les périodes à la date de la dernière interview, chaque répondant possède effectivement sa propre lisière, ce qui étale la lisière sur toute la période d'interview. La présente analyse montre que l'appariement à la date de la dernière interview donne de meilleurs résultats que l'appariement au début de la dernière année financière, car cela évite le problème de la faible remémoration des débuts et des fins de période qui ont eu lieu durant les quelques premiers mois de la période couverte par le calendrier. La méthode qui vise à rapprocher les périodes se trouvant dans le chevauchement des lisières permet de repérer les périodes qui ne sont en fait pas les mêmes et, bien qu'elle produise un nombre plus élevé de périodes omises ou ajoutées, le fichier résultant de données sur les périodes est plus exact.

Les périodes les plus sujettes à des erreurs de remémoration sont celles qui sont différentes de celles déclarées à la date d'interview courante, les périodes courtes et les périodes qui font partie d'une bibliographie complexe. Les résultats de l'étude appuient de façon limitée la notion selon laquelle l'erreur de remémoration est réduite quand l'interview a lieu sur place et que l'intervieweur est le même d'une vague à l'autre. L'effet des caractéristiques du répondant sur la probabilité que des erreurs de remémoration aient lieu varie selon le type de période remémorée. Nous avons également dégagé des preuves que les répondants ont tendance à commettre les mêmes erreurs de remémoration au fil du temps en ce qui concerne l'omission ou l'ajout de périodes, mais pas leur situation incorrecte dans le temps.

La collecte de données sur les périodes chevauchantes permet d'analyser les déclarations non concordantes et d'aider les utilisateurs des données à mieux comprendre les limites de ces dernières. Cela n'aurait pas été possible si nous avions appliqué des méthodes d'interview avec rétroinformation qui donnent parfois à la cohérence la précédence sur l'exactitude.

Bibliographie

Burkhead, D. et Coder, J. (1985). Gross Changes in Income Recipiency from the Survey of Income and Program Participation, *Proceedings of the American Statistical Association, Social Statistics Section*, Washington, DC, pp. 351-356.

Callegaro, M. et Belli, R.F. (2007). Impact of Event History Calendar on Seam Effect in PSID: Lessons for SIPP, paper presented at Event History Calendar Methods Conference, Dec 5-6, Washington, DC.

Hill, D.H. (1987). Response Errors Around the Seam: Analysis of Change in a Panel with Overlapping Reference Periods, *Proceedings of the American Statistical Association*, Section on Survey Research Methods, Washington, DC, pp. 210-215.

Jäckle, A. (2008). The Causes of Seam Effects in Panel Surveys, Institute for Social and Economic Research Working Paper Series, No. 2008-14.

Jäckle, A. et Lynn, P. (2007). Dependent Interviewing and Seam Effects in Work History Data, *Journal of Official Statistics*, Vol. 23, No. 4, pp. 529-551.

Kalton, G. et Miller, M.E. (1991). The Seam Effect with Social Security Income in the Survey of Income and Program Participation, *Journal of Official Statistics*, Vol. 7, No. 2, pp. 235-245.

Lemaître, G. (1992). Dealing with the Seam Problem for the Survey of Labour and Income Dynamics, in SLID Research Paper Series*,* Statistics Canada, Ottawa.

Paull, G. (2002). Biases in the Reporting of Labour Market Dynamics, The Institute for Fiscal Studies Working Paper Series, No. 02/10.

Tourangeau, R., Rips, L.J. et Rasinski, K. (2000). *The Psychology of Survey Response*, Cambridge: Cambridge University Press.

Vick, S. et Weidman, L. (1989). Reporting on Income Recipiency by Self and Proxy Respondents in SIPP, *Proceedings of the American Statistical Association*, Section of Survey Research Methods, Washington, DC, pp. 315-319.

Watson, N. et Wooden, M. (2004). The HILDA Survey four years on, *Australian Economic Review*, Vol. 37, pp. 343-349.

SESSION 18 – IMPUTATION

Imputation par donneur et pondération en presence de non-réponse selon une approche fondée sur un modèle

Roberto Gismondi[1]

Résumé

Dans le présent article, nous formalisons une stratégie d'estimation qui repose à la fois sur l'imputation par donneur déterministe des cas de non-réponse et sur l'estimation fondée sur un modèle optimal dans une base de sondage. Nous développons l'erreur quadratique moyenne de l'estimateur fondé sur l'imputation par donneur et procédons à une comparaison théorique de l'efficacité de cet estimateur à celle de l'estimateur fondé sur la repondération des unités répondantes sans imputation. Nous discutons des conditions – qui dépendent à la fois du taux d'échantillonnage théorique et du taux de réponse – dans lesquelles l'imputation par donneur peut améliorer le modèle pseudo-optimal ordinaire fondé sur un prédicteur. Enfin, nous présentons les résultats d'une comparaison empirique entre l'imputation par donneur, la prédiction fondée sur un modèle et le calage en nous servant de données réelles sur le chiffre d'affaires.

Mots-clés : donneur, imputation, voisin le plus proche, non-réopnse, pondération.

1. Traitement de la non-réponse dans les enquêtes-entreprises

Dans une base de sondage, les ajustements pour traiter les cas de non-réponse ont pour but de réduire le *biais de non-réponse* éventuel (Billiet et coll., 2007). Ce biais dépend souvent d'une spécification incorrecte du modèle, par exemple parce que les répondants et les non-répondants ont des caractéristiques différentes. Des expériences récentes (Gismondi, 2008) ont montré que, dans de nombreux contextes réels d'enquêtes-entreprises, le biais de non-réponse n'est pas systématique, mais pourrait avoir lieu pour certaines vagues de l'enquête et/ou pour certains domaines seulement. Les résultats des stratégies classiques de réduction du biais de non-réponse sont souvent médiocres, par exemple parce qu'un trop petit nombre de variables auxiliaires sont disponibles à l'étape de l'estimation (Rizzo et coll., 1996). En particulier, la plupart des techniques d'imputation ne réduisent pas suffisamment le biais pour équilibrer l'accroissement de la variance due à l'imputation (Copeland et Valiant, 2007).

Dans ce contexte, sous une approche fondée sur un modèle (Valiant et coll., 2000), nous commençons par résumer les règles fondamentales en vue de prédire le total en présence de non-réponse par repondération des répondants, sous l'hypothèse que le biais de non-réponse peut être négligé (section 2). Ensuite (section 3), nous formalisons une stratégie d'estimation fondée à la fois sur l'imputation par donneur déterministe des cas de non-réponse et sur l'estimation optimale fondée sur un modèle par repondération des unités d'échantillonnage. Alors que l'imputation par donneur est utilisée fréquemment dans les sondages, très peu d'estimations de la variance ont été développées (Beaumont et Bocci, 2009). Nous calculons l'erreur quadratique moyenne de l'estimateur fondé sur l'imputation par donneur et procédons à une comparaison théorique de l'efficacité de cet estimateur à celle de l'estimateur fondé sur la repondération des unités répondantes sans imputation. Nous soulignons les conditions – qui varient selon les taux d'échantillonnage et les taux de réponse – dans lesquelles l'imputation par donneur peut améliorer la repondération pseudo-optimale ordinaire. Enfin, nous présentons les principaux résultats d'une étude empirique (section 4) comparant l'imputation par donneur, la prédiction fondée sur un modèle et le calage.

2. Imputation fondée sur un modèle et pondération

2.1 Estimation sans cas de non-réponse

Pour chaque unité i de la population U comprenant N unités, nous émettons l'hypothèse du modèle $y_i=\beta x_i+\varepsilon_i$ avec $E(\varepsilon_i)=0$, $Var(\varepsilon_i)=\sigma^2 v_i$, $Cov(\varepsilon_i,\varepsilon_j)=0$ si $i\neq j$, où β est inconnu, x est connu pour toutes les unités dans U et v doit être spécifié.
Si nous supposons également que nous avons un échantillon S de taille théorique n tiré de U, le meilleur prédicteur linéaire BLU qui minimise l'EQM : $E(T-N\bar{y})^2$ est (Cicchitelli et coll., 1992, p. 385 à 387) :

[1]Roberto Gismondi, ISTAT, Via Tuscolana 1788, 00173 Roma, Italy, gismondi@istat.it

$$T^* = N[f\,\bar{y}_S + (1-f)\,\bar{x}_{\bar{S}}\,\hat{\beta}^*] \qquad \text{où} \qquad \hat{\beta}^* = \left(\sum_S \frac{x_i y_i}{v_i}\right)\left(\sum_S \frac{x_i^2}{v_i}\right)^{-1} \tag{1}$$

où $f=n/N$ et $\bar{y}_S$, $\bar{x}_{\bar{S}}$ sont, respectivement, les moyennes d'échantillon de y et de x se rapportant aux unités non observées. Si nous désignons par j une unité générique non observée, l'*EQM* sous le modèle de T^* fondée sur n observations est :

$$EQM(T^*) = \sigma^2\left[\left(\sum_{\bar{S}} x_j\right)^2 \Big/ \left(\sum_S (x_i^2 / v_i)\right) + \sum_{\bar{S}} v_j\right]. \tag{2}$$

La stratégie d'estimation qui précède peut être généralisée, puisqu'elle équivaut à effectuer une imputation déterministe de chaque unité non observée en suivant la règle :

$$\hat{y}_j = \hat{\beta}\,x_j \qquad \text{avec} \qquad \hat{\beta} = \sum_S a_i y_i \qquad \text{pour chaque unité } j \in \bar{S} \tag{3}$$

où les n coefficients a_i doivent être spécifiés. Si X_Δ est la somme de x_i étendue au domaine Δ, la règle d'imputation définie par (3) équivaut à la règle de pondération $w_i = 1 + a_i X_{\bar{S}}$. Si nous posons que :

$$a_i = a_i^* = \frac{x_i}{v_i}\left(\sum_S x_i^2 / v_i\right)^{-1}, \tag{4}$$

nous revenons au prédicteur sans biais optimal (1) dont l'erreur quadratique moyenne est donnée par (2). L'utilisation de règles d'imputation (ou de systèmes de repondération particuliers) menant à d'autres prédicteurs que T^* est justifiée si ce dernier risque d'être biaisé, par exemple parce que a) toutes les unités non incluses dans l'échantillon observé, et uniquement ces unités, suivent un modèle différent; b) l'échantillon observé comprend certaines unités qui suivent le modèle supposé, mais aussi certaines unités qui suivent un autre modèle devant être spécifié. Nous pouvons définir un prédicteur général du total sous la forme :

$$T = \sum_S y_i + \sum_{\bar{S}} \hat{y}_j = \sum_S y_i + \sum_{\bar{S}} \hat{\beta}\,x_j\,. \tag{5}$$

Son erreur quadratique moyenne peut s'écrire :

$$EQM(T) = \sigma^2 X_{\bar{S}}^2 \sum_S a_i^2 v_i + \sigma^2 \sum_{\bar{S}} v_j + Biais^2(T) = \sigma^2 X_{\bar{S}}^2 \sum_S a_i^2 v_i + \sigma^2 \sum_{\bar{S}} v_j + \beta^2 X_{\bar{S}}^2 \left(\sum_S a_i x_i - 1\right)^2. \tag{6}$$

Sous une approche fondée sur un modèle, le cas des enquêtes par sondage sans non-réponse est formellement semblable à celui des enquêtes par recensement portant sur des populations comprenant N unités avec (N-n) non-réponse.

2.2 Estimation avec cas de non-réponse

Nous supposons que nous avons (n-n_R) cas de non-réponse. L'estimateur du total fondé sur les unités répondantes uniquement est donné par :

$$T_R = \sum_{S_R} y_i + \sum_{\bar{S}_R} \hat{y}_j = \sum_{S_R} y_i + \sum_{\bar{S}_R} \hat{\beta}_R x_j \qquad \text{avec} \qquad \hat{\beta}_R = \sum_{S_R} a_{Ri} y_i \quad \text{et} \quad \hat{y}_j = \hat{\beta}_R x_j\,. \tag{7}$$

L'étiquette R signifie que l'estimation est fondée sur les unités répondantes uniquement, sans aucune imputation préliminaire des non-répondants. En particulier, si $a_{Ri} = a_{Ri}^*$, où a_{Ri}^* est donné par (4) en substituant S_R à S, cela signifie aussi que $\hat{\beta}_R = \hat{\beta}_R^*$ et nous avons le prédicteur *BLU* fondé sur n_R unités répondantes, dont l'*EQM* est donnée par (2), en substituant respectivement S_R à S et $\bar{S}_R$ à $\bar{S}$. Un résultat fondamental est que, si a_i et a_{Ri} ont la même structure formelle, comme cela se produit s'ils sont dérivés du processus d'estimation *BLU* sous le modèle, l'imputation des unités non répondantes avant l'estimation finale n'est pas influente. Par ailleurs, elle peut être utile si le choix de a_{Ri} est fait selon une logique différente de celle qui dicte le choix de a_i. Par exemple, en présence de cas de non-réponse, la stratégie d'estimation pourrait être fondée sur le calage (Lundström et Särndal, 1999), dont l'utilisation est due au risque d'un biais découlant d'une spécification incorrecte du modèle. Il est possible d'éviter l'imputation en utilisant les coefficients a_{Ri} dérivés des poids calés, ou d'imputer pour les non-réponses sur la base de $\hat{\beta}_R$, puis d'utiliser les coefficients de calage a_i. Les résultats des deux stratégies seront différents (section 4). La forme générale de tout estimateur fondé sur l'imputation des cas de non-réponse et sur la pondération est :

$$T_w = \sum_{S_R} w_i y_i + \sum_{S\bar{R}} w_j \hat{y}_i\,. \tag{8}$$

3. Imputation par donneur déterministe et pondération

Nous analysons une stratégie d'estimation en deux étapes, à savoir 1) l'application d'une méthode d'imputation par donneur aux unités non répondantes, et 2) l'utilisation conjointe des données sur les unités répondantes et non répondantes afin de déterminer les unités n'appartenant pas à l'échantillon théorique sur la base des mêmes coefficients a_i définis en (3). En cas d'utilisation de l'imputation par donneur, pour chaque $j \in S_{\bar{R}}$, nous sélectionnons le donneur particulier désigné par $i(j)$ parmi les n_R répondants désignés par i. Nous disons aussi que $d_j=i$. Le processus d'imputation par donneur peut s'exprimer de la façon suivante :

$$\hat{y}_{jd} = \hat{\beta}_j x_j \quad \text{où} \quad \hat{\beta}_j = y_{i(j)} / x_{i(j)} = a_{i(j)} y_{i(j)} \quad \text{pour chaque } j \in S_{\bar{R}}. \tag{9}$$

La règle d'imputation (9) aboutit à la modélisation d'estimations sans biais. L'estimateur dérivé de (9) est donné par :

$$T_{I(d)} = \sum_{S_R} y_i + \sum_{S_{\bar{R}}} \hat{y}_{jd} + \sum_{\bar{S}} \hat{y}_j = \sum_{S_R} y_i + \sum_{S_{\bar{R}}} \hat{\beta}_j x_j + \sum_{\bar{S}} \hat{\beta}_{I(d)} x_j, \tag{10}$$

où $\hat{\beta}_j$ est défini par (9) et $\hat{\beta}_{I(d)} = \sum_{S_R} a_i y_i + \sum_{S_{\bar{R}}} a_j \hat{y}_{jd} = \sum_{S_R} a_i y_i + \sum_{S_{\bar{R}}} a_j \hat{\beta}_j x_j$. L'estimateur (10) est fondé sur l'imputation par donneur des (n-n_R) non-répondants et sur l'imputation des unités n'appartenant pas à S par la voie des coefficients originaux a_i. Il n'est *pas linéaire* si le choix des donneurs – et, par conséquent, des coefficients $a_{i(j)}$ – dépend de la connaissance des étiquettes dans S. Une technique fréquente de sélection des donneurs selon une approche déterministe est fondée sur la méthode du *plus proche voisin*, conformément à la règle :

$$D_{i(j),j} = \left| x_{i(j)} - x_j \right| = \underset{i \in S_R}{Min}(D_{i,j}) \qquad \text{pour chaque unité } j \in S_{\bar{R}} \tag{11}$$

où D est un opérateur de distance. En nous appuyant sur (11), nous pouvons supposer que $x_{i(j)} \approx x_j$ pour chaque unité $j \in S_{\bar{R}}$. Nous avons : $Var(\hat{y}_{jd}) = Var(y_{i(j)}) = \sigma^2 v_{i(j)}$. En outre, nous obtenons :

$$EQM(T_{I(d)}) = \sigma^2 \sum_{S_{\bar{R}}} v_{i(j)} + \sigma^2 \left(\sum_{\bar{S}} x_j \right)^2 \left\{ \sum_{S_R} a_i^2 v_i + \sum_{S_{\bar{R}}} a_j^2 x_j^2 a_{i(j)}^2 v_{i(j)} + 2 \sum_{i \in S_R} \left[a_i v_i \left(\sum_{j \in S_{\bar{R}} \cdot d_j = i} a_j a_{i(j)} x_j \right) \right] \right\} +$$
$$+ 2\sigma^2 \left(\sum_{\bar{S}} x_j \right) \left\{ \sum_{i \in S_R} \left[a_i v_i \left(\sum_{j \in S_{\bar{R}} \cdot d_j = i} a_{i(j)} x_j \right) \right] + \sum_{i \in S_R} \left[v_i \left(\sum_{j \in S_{\bar{R}} \cdot d_j = i} a_j a_{i(j)}^2 x_j^2 \right) \right] \right\} + \sigma^2 \sum_{\bar{S}_R} v_j \,. \tag{12}$$

La signification de chaque terme entre crochets est que, si l'on fixe l'étiquette i dans la première somme, le terme entre parenthèses est une somme comprenant un nombre de termes égal au nombre j d'unités non répondantes dont le donneur est donné par cette unité i particulière. Étant donné les formules (2), quand S=S_R et (12), nous pouvons vérifier sous quelle condition on pourrait avoir : $EQM(T_{I(d)}) \le EQM(T_R)$. En particulier, si $T_{I(d)}$ et T_R sont fondés, respectivement, sur les coefficients optimaux a_i^* et a_{Ri}^*, nous pouvons nous concentrer sur l'inégalité :

$$EQM(T^*_{I(d)}) \le EQM(T^*_R). \tag{13}$$

Même si T^*_R est le prédicteur *BLU* par construction, $T^*_{I(d)}$ pourrait avoir une *EQM* plus faible, parce qu'il est fondé sur un plus grand nombre d'unités échantillonnées, parmi lesquelles n_R sont des répondants réels et (n-n_R) ont été imputées.

Nous pouvons supposer que x_i=v_i pour chaque unité i de la population. Cela simplifie la formule (12), où nous pouvons poser que $v_{i(j)} = v_j = x_j$. De (2) quand S=S_R, il découle que :

$$EQM(T^*_{R/v=x}) = \sigma^2 (X^2_{\bar{S}_R} / X_{S_R} + X_{\bar{S}_R}). \tag{14}$$

Nous avons aussi $a_i^* = X_S^{-1}$, $a_{i(j)} = x_i^{-1}$. Compte tenu du fait que $S = S_R \cup S_{\bar{R}}$ et $\bar{S}_R = S_{\bar{R}} \cup \bar{S}$, nous obtenons :

$$EQM(T^*_{I(d)/v=x}) = \sigma^2 [X_{S_{\bar{R}}} (1 + 2X^2_{\bar{S}} X_S^{-2} + 4X_{\bar{S}} X_S^{-1}) + (X^2_{\bar{S}} X_S^{-1} + X_{\bar{S}_R})]. \tag{15}$$

En partant de (14) et (15), nous pouvons également vérifier que :

$$EQM(T^*_{I(d)/v=x}) \le EQM(T^*_{R/v=x}) \qquad \leftrightarrow \qquad (X_U / X_S)^2 \le (X_S / X_{S_R})(X_U / X_S - 0{,}5). \tag{16}$$

Par exemple, si X_U=100 et X_S=20, la condition (16) est satisfaite si $X_{S_R} \leq 3{,}6$. À partir de (16), en prenant $f_{xS} = X_S/X_U$ et $f_{xS_R} = X_{S_R}/X_U$, nous obtenons la condition pour laquelle la stratégie d'imputation par donneur peut améliorer T_R :

$$(X_{S_R}/X_S) \leq (X_U/X_S - 0{,}5)(X_U/X_S)^{-2} \qquad \leftrightarrow \qquad (f_{xS_R}/f_{xS}) \leq (f_{xS}^{-1} - 0{,}5) f_{xS}^2 . \tag{17}$$

Dans le tableau 3-1, nous avons calculé le ratio entre les erreurs quadratiques moyennes des estimateurs $T^*_{I(d)}$ et T^*_R pour certains niveaux des ratios X_S/X_U et X_{S_R}/X_S, quand *v*=*x*. Les ratios inférieurs à l'unité correspondent aux conditions dans lesquelles l'expression (17) est satisfaite (cases grisées). L'utilité de l'imputation par donneur est garantie si le taux d'échantillonnage pondéré est au moins de 60 % et que le taux de réponse pondéré n'est pas supérieur à 40 %, ce qui peut arriver dans les strates caractérisées par un taux d'échantillonnage élevé et un mauvais résultat d'enquête (les répondants représentent moins de la moitié de la taille de l'échantillon original). En particulier, dans une enquête par recensement, il est nécessaire que le taux de réponse sous échantillonnage pondéré soit inférieur à 50 %. Si le taux de réponse sous échantillonnage pondéré est faible (n'excède pas 40 %), dans la plupart des cas, l'estimation sans imputation par donneur doit être privilégiée; par exemple, avec un taux de réponse sous échantillonnage pondéré de 20 %, nous ne devrions recourir à l'imputation par donneur que si le taux d'échantillonnage pondéré original est au moins de 30 %. Pour les taux de réponse sous échantillonnage pondéré supérieurs à 40 %, nous devrions éviter l'imputation par donneur. Dans la dernière colonne du tableau, nous avons calculé le niveau du taux de réponse sous échantillonnage pondéré au-delà duquel l'imputation par donneur peut être utilisée; par exemple, quand le taux d'échantillonnage pondéré est égal à 40 %, l'imputation par donneur peut être privilégiée si le taux de réponse sous échantillonnage pondéré est inférieur à 32 %. Brièvement, un taux d'échantillonnage pondéré inférieur à 50 % est une condition *nécessaire*, mais non suffisante, pour que la stratégie d'imputation par donneur améliore celle fondée sur un modèle. Si *S*=S_R et que l'imputation par donneur est effectuée sur les unités appartenant à $\overline{S}$, alors la condition précédente sera *suffisante* également.

Tableau 3-1
Ratio des *EQM* des estimateurs $T_{I(d)}$ et T_R quand *v*=*x* pour divers taux d'échantillonnage pondérés et taux de réponse pondérés, et conditions dans lesquelles $T_{I(d)}$ améliore T_R (relation (17))

$100 \cdot f_{xS}$	$100 \cdot (f_{xS_R}/f_{xS}) = 100 \cdot F_{xS_R}$										$EQM(T^*_{I(d)}) \leq EQM(T$
	5,0	10,0	20,0	30,0	40,0	50,0	60,0	70,0	80,0	90,0	si $100 \cdot F_{xS_R} \leq$
5,0	1,03	2,05	4,10	6,15	8,21	10,26	12,31	14,36	16,41	18,46	4,88
10,0	0,53	1,05	2,11	3,16	4,21	5,26	6,32	7,37	8,42	9,47	9,50
20,0	0,28	0,56	1,11	1,67	2,22	2,78	3,33	3,89	4,44	5,00	18,00
30,0	0,20	0,39	0,78	1,18	1,57	1,96	2,35	2,75	3,14	3,53	25,50
40,0	0,16	0,31	0,63	0,94	1,25	1,56	1,88	2,19	2,50	2,81	32,00
50,0	0,13	0,27	0,53	0,80	1,07	1,33	1,60	1,87	2,13	2,40	37,50
60,0	0,12	0,24	0,48	0,71	0,95	1,19	1,43	1,67	1,90	2,14	42,00
70,0	0,11	0,22	0,44	0,66	0,88	1,10	1,32	1,54	1,76	1,98	45,50
80,0	0,10	0,21	0,42	0,63	0,83	1,04	1,25	1,46	1,67	1,88	48,00
90,0	0,10	0,20	0,40	0,61	0,81	1,01	1,21	1,41	1,62	1,82	49,50
100,0	0,10	0,20	0,40	0,60	0,80	1,00	1,20	1,40	1,60	1,80	50,00

4. Essai empirique et conclusion

Les données utilisées pour l'essai empirique proviennent de l'enquête par sondage trimestrielle sur le commerce de gros réalisée par ISTAT. Même si l'objectif premier de l'enquête est d'estimer des indices trimestriels de chiffre d'affaires avec pour année de base 2005=100 au moyen des données sur les chiffres d'affaires recueillies trimestriellement, ici, nous nous concentrons sur l'estimation du chiffre d'affaires annuel total disponible dans le registre des entreprises pour chaque entreprise en exploitation, pour la période de 2003 à 2007. Ce choix tient au fait qu'il est nécessaire d'utiliser la valeur totale réelle de la variable cible *y* comme valeur repère pour évaluer la précision des estimations obtenues en utilisant divers critères d'estimation. En outre, nous supposons que :

1) la population (de taille *N*) est donnée par l'échantillon théorique complet;
2) l'échantillon (de taille *n*) est donné par l'échantillon d'unités qui ont répondu dans les 180 jours après la fin du trimestre de référence (dans le contexte de l'enquête réelle, les estimations finales ne sont diffusées qu'après six mois);
3) l'échantillon de répondants (de taille n_R) comprend les unités qui répondent rapidement (réponse dans les 90 jours).

Pour chaque année, nous considérons comme étant des « répondants dans les 180 jours » (les n unités définies au point 2) les unités qui ont répondu dans les 180 jours à *chacun des quatre* trimestres, et comme des « répondants dans les 90 jours » (les unités définies au point 3) les unités qui ont répondu dans les 90 jours à *chacun* des quatre trimestres. Nous avons exécuté des analyses distinctes dans chacun des huit domaines suivants : 1) vente en gros sur une base libre ou contractuelle, 2) matières premières agricoles et cheptel vif, 3) aliments, boissons et tabac, 4) biens ménagers, 5) produits intermédiaires non agricoles, 6) machines, matériel et fournitures, 7) autres produits et 8) commerce de gros total. Si Y désigne l'année de référence courante, nous considérons comme étant la variable y le chiffre d'affaires total se rapportant à l'année (Y-1), puisque le registre des entreprises est mis à jour avec un an de retard, tandis que x est le chiffre d'affaires total de l'année (Y-2). En moyenne, de 2003 à 2007 (tableau 3), pour le commerce de gros total, nous avons N=7 572, n=5 636 et n_R=4 631, de sorte que 100 n/N=74,4 et 100 n_R/n=82,2. Selon les résultats théoriques dérivés du tableau 1, nous ne devrions pas obtenir de grands gains d'efficacité en utilisant l'imputation par donneur. Les huit stratégies d'estimation comparées sont décrites au tableau 4-1.

Tableau 4-1
Stratégies d'estimation comparées

Code	Modèle	Définition	Détails
I	v=1	Prédiction optimale – pas d'imputation	Prédicteur (7), β est donné par la deuxième formule de (1) avec S=S_R
I	v=x		
II	v=1	Imputation fondée sur un modèle des cas de non-réponse et calage	Prédicteur (8), non-réponses estimées par les deuxième et troisième formules de (7), β donné par la deuxième formule de (1) avec S=S_R, poids w provenant du calage
II	v=x		
III	v=1	Calage – pas d'imputation	Prédicteur (7), poids w provenant du calage
III	v=x		
IV	v=1	Imputation par donneur et prédiction optimale	Prédicteur défini par (10) où les coefficients a sont optimaux sous le modèle
IV	v=x		

Les principaux résultats sont résumés en examinant les erreurs absolues moyennes en pourcentage (*MAPE*, pour *mean of absolute percent estimation errors*). Les chiffres du tableau 4-2 sont les moyennes arithmétiques de 20 erreurs d'estimation trimestrielles couvrant la période de 2003 à 2007. L'utilisation des taux de réponse réels (partie supérieure du tableau) mène aux conclusions suivantes :

a) Dans le cas de la moyenne pour les huit domaines, la meilleure stratégie (caractères gras) est (1) avec v=x (*MAPE*=1,70 %), qui est également la meilleure pour six domaines individuels (tous sauf 4 et 6) et la deuxième par ordre décroissant (soulignement) pour le domaine 6. Ce résultat implique que l'imputation est à peine utile. Cette constatation est également confirmée par le deuxième meilleur résultat donné par la stratégie (III) avec v=x (*MAPE*=2,00 %), qui est fondée sur le calage sans imputation. Cette stratégie vient au deuxième rang pour quatre domaines également (1, 2, 7 et total).
b) Cependant, la stratégie (IV) – fondée sur l'imputation par donneur – avec v=x s'avère la meilleure pour le domaine 6 (*MAPE*=2,28 %) et occupe le deuxième rang pour les domaines 3 et 4. En outre, le modèle alternatif fondé sur l'imputation définie par la stratégie (II) avec v=x donne la *MAPE* la plus faible (1,64 %) pour le domaine 4.
c) Enfin, l'option v=1 est, en fait, presque toujours moins réaliste que l'option v=x, ce qui implique le refus du profil homoscédastique des variances sous le modèle unitaire.

Afin d'évaluer la stabilité des résultats qui précèdent pour un taux de réponse plus faible, une mise en blanc de 50 % des réponses a été répétée aléatoirement 1 000 fois. Les résultats résumés à la partie inférieure du tableau mènent à des conclusions semblables, à savoir que la seule exception significative est l'application du calage sans imputation (stratégie (III) avec v=x) par rapport à la stratégie (I) avec v=x, puisque la première devient la meilleure pour les domaines 1, 2 et 3, ainsi que dans le cas de la moyenne (*MAPE*=2,98 %), tandis que la seconde se classe au deuxième rang. Les résultats des stratégies fondées sur l'imputation par donneur ne présentent aucun changement pertinent, quoiqu'en moyenne, cette stratégie devrait quand même être préférée à l'imputation fondée sur un modèle (stratégie (II)).

Même si, selon les résultats qui précèdent, en présence de cas de non-réponse, l'imputation par donneur pourrait n'accroître la précision des estimations que dans quelques circonstances seulement, des travaux de recherche supplémentaires devraient être réalisés dans deux domaines, à savoir 1) les améliorations théoriques du mécanisme de sélection des donneurs (p. ex., sélection des donneurs éventuellement fondée sur plus d'une variable) et 2) d'autres essais empiriques portant sur des variables différentes (non continues, comme le nombre d'emplois vacants) et divers cadres caractérisés par des taux de réponse différents.

Tableau 4-2
MAPE [*] **par domaine et stratégie (données réelles et 1 000 répliques aléatoires, moyenne pour 2003 à 2007)**

Domaine	*N*	*n*	n_R	Stratégie (I) *v*=1	(I) *v*=*x*	(II) *v*=1	(II) *v*=*x*	(III) *v*=1	(III) *v*=*x*	(IV) *v*=1	(IV) *v*=*x*
				Taux de réponse réel							
1	1 205	789	605	1,95	**1,15**	3,70	1,96	2,89	1,31	2,68	1,55
2	570	417	331	1,67	**1,58**	2,11	1,85	2,13	1,62	2,43	1,65
3	961	743	605	2,39	**1,63**	4,19	3,18	3,60	1,96	2,87	1,75
4	2 227	1 671	1 387	3,15	1,95	**1,64**	1,72	1,82	2,02	2,23	1,72
5	1 248	997	853	3,31	**2,16**	3,11	3,22	3,49	2,90	2,84	3,17
6	863	671	570	3,29	2,60	3,29	3,03	3,44	3,14	4,66	**2,48**
7	498	348	281	1,68	**1,41**	3,17	1,86	2,40	1,50	2,25	2,71
Total	7 572	5 636	4 631	1,89	**1,16**	2,10	2,21	2,66	1,56	1,84	1,82
Moyenne				2,42	**1,70**	2,91	2,38	2,80	2,00	2,72	2,11
				1 000 répliques aléatoires avec un taux de réponse de 50 %							
1	1 205	789	405	4,48	3,16	6,07	4,85	4,75	**3,01**	4,97	3,34
2	570	417	214	5,73	3,41	4,77	4,22	4,44	**2,97**	5,09	4,08
3	961	743	385	3,98	2,50	5,95	4,81	5,20	**2,43**	5,09	2,71
4	2 227	1 671	857	5,51	3,69	3,98	4,87	4,06	3,57	5,54	**3,02**
5	1 248	997	512	4,68	**3,60**	4,60	4,20	4,94	3,78	5,37	3,78
6	863	671	339	4,84	**4,02**	6,19	5,32	7,12	4,23	4,88	4,14
7	498	348	175	2,10	**1,73**	4,17	3,57	4,37	1,87	3,59	2,78
Total	7 572	5 636	2 887	2,91	**1,91**	3,74	3,54	4,18	1,98	4,40	2,31
Moyenne				4,28	3,00	4,93	4,42	4,88	**2,98**	4,86	3,27

[*] MAPE = *mean absolute percentage errors* (erreurs absolues moyennes en pourcentage)

Bibliographie

Beaumont, J.F. et Bocci, C. (2009). Variance Estimation when Donor Imputation is Used to Fill in Missing Values, *Canadian Journal of Statistics*, on line publication at the address: http://www3.interscience.wiley.com/cgi-bin/fulltext/122466454/PDFSTART.

Billiet, J., Philippens, M., Fitzgerald, R. et Stoop, I. (2007). Estimations of Non-response Bias in the European Social Survey Using Information from Reluctant Respondents, *Journal of Official Statistics*, Vol.23, 2, pp.135-162.

Cicchitell,i G., Herzel, A. et Montanari, G.E. (1992). *Il campionamento statistico*, Bologna, Il Mulino.

Copeland, K.R. et Valliant, R. (2007). Imputing for Late Reporting in the U.S. Current Employment Statistics Survey, *Journal of Official Statistics*, Vol.23, 1, pp.69-90.

Gismondi, R. (2008). Reducing Revisions in Short-term Business Surveys, Statistica, anno LXVIII, 1, pp.85-116, Bologna, Clueb.

Lundström, S. et Särndal, C.E. (1999). Calibration as a Standard Method for Treatment of Non-response, *Journal of Official Statistics*, Vol.15, 2, pp.305-327.

Rizzo, L., Kalton, G. et Brick, M.J. (1996). Comparaison de quelques methodes de correction de la non-réponse d'un panel, Techniques d'enquête, 22, 1, pp. 43-53.

Valliant, R, Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference – A Prediction Approach*, New York, J.Wiley & Sons.

L'imputation aléatoire équilibrée dans les enquêtes

David Haziza, Guillaume Chauvet et Jean-Claude Deville[1]

Résumé

Les méthodes d'imputation aléatoire sont souvent utilisées dans la pratique parce qu'elles ont tendance à préserver la répartition de la variable imputée, caractéristique importante lorsqu'il s'agit d'estimer des quantiles. Une méthode d'imputation aléatoire particulière, l'imputation hot deck aléatoire, est souvent utilisée dans la pratique si la variable imputée est catégorique parce qu'elle élimine la possibilité de valeurs impossibles. Elle est utilisée également lorsqu'on souhaite imputer plus d'une variable à un moment donné parce que le même donneur peut être utilisé pour imputer toutes les valeurs manquantes, ce qui aide à préserver les relations entre variables. Toutefois, les méthodes d'imputation aléatoire introduisent une certaine variabilité supplémentaire, appelée variance d'imputation, en raison de la sélection aléatoire des résidus. Dans cet article, nous adaptons la méthode du cube (Deville et Tillé, 2004) qui permet de sélectionner des échantillons équilibrés et nous proposons une catégorie de méthodes d'imputation équilibrée aléatoire qui réduisent ou éliminent la variance due à l'imputation tout en préservant la répartition de la variable imputée. La catégorie proposée de méthodes d'imputation peut être appliquée aux variables tant catégoriques que continues. En outre, elle peut être utilisée dans tout plan d'échantillonnage. Nous présentons également les résultats d'une étude par simulation limitée.

[1] David Haziza, Université de Montréal, Canada (david.haziza@statcan.gc.ca); Guillaume Chauvet et Jean-Claude Deville, Laboratoire de Statistique d'Enquête (CREST/ENSAI), France.

La mise à l'essai de nouvelles méthodes d'imputation des gains dans la Survey of Income and Program Participation

Martha Stinson et Gary Benedetto, U.S. Census Bureau, U.S.A.[1]

Résumé

Cet article porte sur la faisabilité et l'efficacité de trois modifications importantes apportées aux méthodes standard utilisées par le Census Bureau pour imputer les gains dans la Survey of Income and Program Participation (SIPP). Actuellement, l'imputation est effectuée par stratification des données en fonction d'un ensemble de caractéristiques choisies par l'analyste, avec tri aléatoire dans chaque sous-groupe, et choix de l'enregistrement donneur qui est le plus proche voisin. Nous examinons la possibilité d'utiliser une approche fondée sur un modèle, en ajoutant des données administratives sur les gains aux données d'enquête sur l'emploi et les caractéristiques démographiques et en utilisant l'imputation multiple proposée par Rubin. Nous modélisons les gains mensuels du 2 janvier 2004 au 6 décembre 2005 en utilisant le panel de la SIPP de 2004 lié aux enregistrements fiscaux (W-2) tirés du Social Security Master Earnings File. Nous utilisons des techniques de régression linéaire pour estimer une loi prédictive a posteriori qui est la répartition des gains sachant toutes les caractéristiques observées (y compris les données administratives sur les gains). À partir de cette répartition, nous procédons à quatre tirages de manière à créer quatre valeurs imputées par cas pour lequel les gains sont manquants. Après avoir ainsi « complété » les données manquantes, nous comparons les résultats en utilisant les valeurs initiales par rapport aux nouvelles valeurs imputées de plusieurs analyses types afin d'évaluer l'impact de notre nouvelle méthode. Plus particulièrement, nous examinons les coefficients dans une régression classique des gains, les tendances de la variation des gains au fil du temps, les moments de la répartition transversale des gains au cours d'un mois donné et les niveaux de pauvreté selon le revenu de la famille, dont les gains sont une composante importante. Les quatre valeurs imputées nous permettent de calculer les estimations de la variance au moyen des formules de Rubin de la variance par imputation multiple ainsi que d'évaluer l'effet de l'imputation sur la signification des coefficients de régression, la forme de la répartition des gains et la marge d'erreur dans les estimations de la pauvreté.

[1] Martha Stinson (martha.stinson@census.gov) and Gary Benedetto (gary.linus.benedetto@census.gov), U.S. Census Bureau, É.-U.

SESSION 19 – VÉRIFICATION ET IMPUTATION

LE POINT SUR LE PROJET EU-SILC EN SLOVÉNIE

Rudi Seljak[1]

Résumé

L'Enquête sur le revenu et les conditions de vie (SILC) est une enquête européenne harmonisée qui vise à produire des données sur les conditions de vie des membres des ménages et de certaines personnes, ainsi que sur la façon dont ils s'intègrent dans la société. Étant donné que les résultats de l'enquête sont harmonisés, les variables de résultats sont celles prescrites par le règlement européen, tandis que la méthode de collecte des microdonnées d'entrée est essentiellement déterminée par chaque pays.

En Slovénie, les microdonnées pour l'enquête EU-SILC sont recueillies par l'entremise de trois sources : enquête « classique », autres sources statistiques et sources administratives. L'utilisation exhaustive de sources administratives comporte un avantage évident de réduction du fardeau de réponse, mais peut aussi entraîner certains inconvénients, dont le plus évident est l'augmentation du travail de vérification des données. Dans la présente communication, nous résumons « l'expérience de l'EU-SILC », échelonnée sur quatre ans, de la fusion de données provenant de différentes sources, et nous soulignons les principaux avantages et inconvénients de cette approche.

Mots clés : Enquête sur le revenu et les conditions de vie, vérification, données administratives.

1. Introduction

L'Enquête sur le revenu et les conditions de vie (EU-SILC) est un projet visant à établir une enquête européenne harmonisée pour la collecte de statistiques comparatives sur la distribution du revenu et l'exclusion sociale auprès d'États membres de l'UE, de la Norvège et de l'Islande. Le projet a été lancé en 2003 (à ce moment-là sur la base d'une entente verbale) dans six États européens membres, a été élargi en 2004, pour englober 12 « anciens » États membres, l'Estonie et l'Islande, puis en 2005, pour inclure tous les États membres (à ce moment-là), la Norvège et l'Islande.

À partir de 2004, l'enquête a été menée sur la base d'un règlement du Parlement européen. Ce règlement définit l'EU-SILC comme une enquête dont les résultats sont harmonisés, ce qui signifie que les États membres doivent fournir les variables de résultats prescrites dans le règlement et que la méthode de collecte des données doit être déterminée par chaque pays. Ainsi, certains pays continuent de recueillir toutes les données de la façon « classique », au moyen d'une enquête sur le terrain, tandis que d'autres utilisent une combinaison de données d'enquête « classiques » et de données tirées de différentes sources administratives. Comme les données de l'EU-SILC doivent servir à la fois à des analyses transversales et longitudinales, il est fortement recommandé de la concevoir comme une enquête par panel.

En Slovénie, l'EU-SILC a été menée pour la première fois en 2004, à titre d'enquête pilote, puis, en 2005, à titre d'enquête « régulière ». À l'étape de la planification et du lancement, nous avons tenté de suivre les recommandations d'Eurostat d'utiliser le plus grand nombre possible de sources de données existantes, afin de réduire le fardeau de réponse et d'augmenter par conséquent le taux de réponse. Ainsi, en Slovénie, les microdonnées de l'EU-SILC sont recueillies à partir de trois types de sources. La première partie des données est recueillie au moyen d'une enquête « classique »; la deuxième partie provient d'autres sources statistiques, et la troisième, de registres et de sources administratives. Toutes les variables liées au revenu, notamment, (qui sont habituellement considérées comme de nature très délicate) sont recueillies auprès des différentes sources administratives.

Même si l'utilisation exhaustive de sources administratives comporte de nombreux avantages, particulièrement en ce qui a trait au fardeau de réponse et à la réduction des coûts de l'enquête, une telle approche peut aussi entraîner certains inconvénients. Le principal est certainement l'augmentation du travail de vérification des données. Lorsque nous avons mené l'enquête pour la première fois, nous n'étions pas pleinement conscients de la complexité du processus de vérification. Par conséquent, à ce moment-là, toutes les procédures de vérification étaient plus ou moins constituées de programmes informatiques ponctuels, sans perspective systématique et à long terme. Étant donné que cette approche a entraîné un grand nombre de problèmes à

[1] Rudi Seljak, Statistical Office of the Republic of Slovenia, courriel: rudi.seljak@gov.si, Tel.:+386 1 2415 294

l'étape du traitement, nous avons commencé à élaborer un système plus générique facilitant la gestion et permettant un meilleur contrôle du traitement des données.

Dans la présente communication, nous décrivons les expériences liées au déroulement sur quatre ans de l'EU-SICL en Slovénie. Dans la première partie du document, nous fournissons certains renseignements généraux et nous décrivons l'application de traitement des données. Dans la deuxième partie, nous résumerons les avantages et les inconvénients principaux de l'utilisation exhaustive de données administratives comme source de données directe. À la fin du document, nous soulignons certaines conclusions et orientations pour le développement futur.

2. EU-SILC en Slovénie

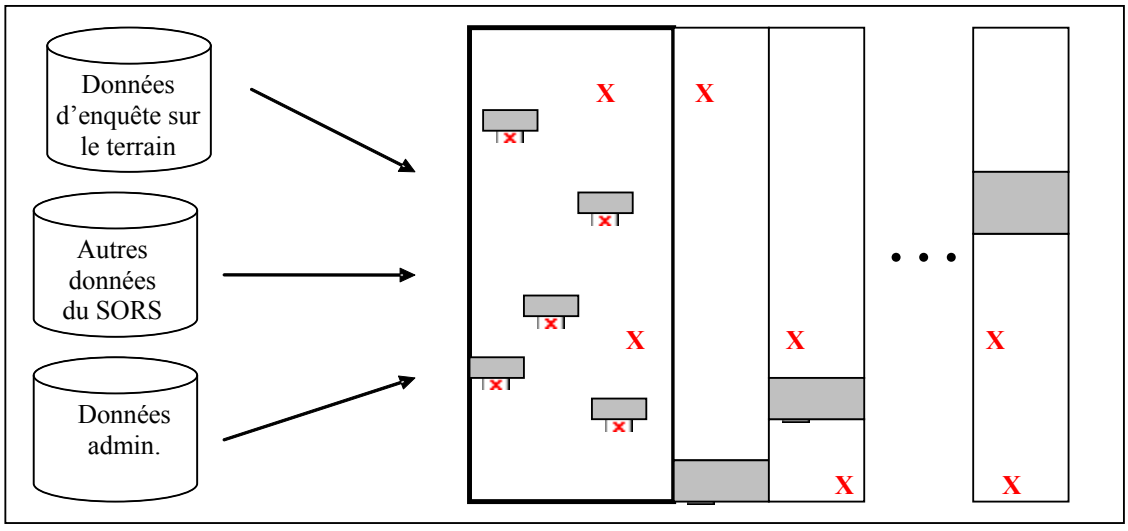
Comme il a été mentionné dans l'introduction, en Slovénie, les microdonnées pour l'EU-SILC sont recueillies par l'entremise de trois types de sources différentes :

- Une enquête sur le terrain classique utilisant des questionnaires par IPAO et ITAO. L'ensemble de données des questionnaires remplis par environ 9 000 ménages (environ 30 000 personnes) par année sert de base à l'intégration des données.
- D'autres sources à l'intérieur du bureau statistique. Les données du registre de l'emploi et de l'enquête sur les bourses d'études sont utilisées.
- Les données des institutions externes suivantes sont utilisées : administration fiscale; ministère du Travail, de la Famille et des Affaires sociales; institut des pensions et de l'assurance-invalidité; service d'emploi de Slovénie; institut de l'assurance-maladie; ministère de l'Intérieur.

Étant donné qu'en Slovénie, le registre des ménages et le registre des logements sont encore en voie d'élaboration, on ne les utilise pas pour le moment pour l'échantillonnage. Cela signifie que nous avons sélectionné un échantillon aléatoire de personnes, et que toutes les personnes vivant dans le même ménage que les personnes sélectionnées ont été interviewées. Comme nous ne sommes pas autorisés à obtenir le numéro d'identification personnel (NIP) administratif directement auprès du répondant, on dispose de ces numéros d'identification pour les personnes sélectionnées seulement. Par conséquent, avant l'étape de l'intégration, nous avons recours à l'approche du couplage indirect des enregistrements, à partir des données de base (nom, nom de famille, adresse, date de naissance, etc.) recueillies sur le terrain pour trouver le NIP correspondant dans le registre central de la population. Grâce à une combinaison d'applications informatiques et de travail manuel, les NIP d'environ 99 % des personnes sont déterminés, et seulement 1 % des NIP sont imputés.

Après avoir obtenu les NIP pour toutes les unités, on procède à l'intégration des données de différentes sources. Du fait de l'imperfection des sources, les données pour certaines unités sont manquantes, et certaines données deviennent incohérentes après l'intégration. La situation suivant l'étape de l'intégration est présentée à la figure 2-1, les parties en gris des données intégrées représentant des données manquantes et les croix en rouge, les données comportant une incohérence logique.

Figure 2-1
Processus d'intégration



3. Traitement des données

En 2005, la première année de l'exécution de l'enquête, tout le traitement a été effectué de façon « classique » au moyen de programmes informatiques personnalisés en SAS. En pratique, le personnel spécialisé et le personnel général de la méthodologie ont rédigé les instructions, et sur la base de ces dernières, le personnel des TI a préparé les programmes informatiques. Après la première exécution des programmes, on a déterminé les besoins éventuels de correction et d'amélioration, puis la nouvelle version du programme a été préparée. Du fait de la complexité des données de l'EU-SILC, une telle approche a pris du temps, a gaspillé une part importante du temps de travail de nombreux employés et a été qualifiée à juste titre de très inefficace. Par conséquent, nous avons commencé à examiner la possibilité d'élaborer un système nouveau, plus général et plus souple qui, en dernier ressort, servira aussi à d'autres enquêtes.

L'idée générale était de préparer une application axée sur les métadonnées, ce qui signifie que toute l'information déterminant les paramètres de l'exécution du traitement d'une enquête particulière et d'une période de référence particulière devait être fournie à l'extérieur du code machine de base. Les renseignements concernant l'exécution d'une enquête particulière ne devaient pas être intégrés dans le code programme, mais ils devaient être fournis par le personnel spécialisé, au moyen de tableaux spéciaux de métadonnées. Une description plus détaillée de ces métadonnées, dans le cas des trois processus particuliers, est fournie ci-après.

3.1 Vérifications logiques

Deux groupes de vérifications logiques ont pu être effectuées : transversales et longitudinales. Lorsqu'on exécute des vérifications transversales, seules les données de l'année en cours peuvent être contrôlées, tandis que dans le cas des vérifications longitudinales, les vérifications peuvent aussi inclure les données de l'année de référence précédente. L'utilisateur a aussi la possibilité de choisir la « version » des données à contrôler. Par « version », nous entendons ici les ensembles de données découlant des différentes parties du processus, comme les données brutes, les données suivant le transfert des données de l'année précédente, les données suivant les corrections logiques, etc.

Les métadonnées suivantes sont nécessaires pour exécuter le processus de vérification logique :
- dénotation de la vérification logique;
- nom du tableau dans la base de données;
- expression logique, qui devrait toujours être écrite sous la forme qui détermine l'erreur;
- commentaire ou description de la vérification (arbitraire).

Après l'exécution, l'utilisateur obtient immédiatement à l'écran le nombre d'enregistrements rejetés pour chacune des vérifications définies. L'ensemble d'unités rejetées est intégré dans le tableau uniforme (SAS et Access). Dans ce tableau, pour chacun des enregistrements rejetés, il existe aussi des renseignements sur la ou les vérifications qui ont donné lieu au rejet. Ces produits servent alors de base à la préparation des « métadonnées de correction », qui sont nécessaires plus tard dans le processus.

3.2 Corrections des données

Grâce à ce processus, les valeurs des variables qui ont été désignées auparavant comme erronées peuvent être corrigées. Il existe deux façons différentes de sélectionner les unités pour lesquelles la variable particulière devrait être corrigée, et trois façons différentes de déterminer la nouvelle valeur. La première façon de déterminer les unités dont la valeur devrait être corrigée consiste à fournir le numéro d'identification unique de l'unité. Nous appelons ces corrections des corrections individuelles. La deuxième façon consiste à corriger les valeurs de certaines variables pour toutes les unités qui respectent une condition en particulier. Nous appelons ces corrections des corrections systématiques.

Les nouvelles valeurs (corrigées) peuvent être déterminées de trois façons différentes. L'utilisateur peut fournir la nouvelle valeur exacte, l'expression arithmétique qui servira à déterminer la nouvelle valeur ou la limite inférieure et supérieure (fourchette) de la nouvelle valeur. La valeur dans ce dernier cas est estimée au moyen d'une procédure d'imputation Banff proc donorimputation, les contraintes de la vérification étant déterminées selon une fourchette donnée.

Les métadonnées qui doivent être fournies diffèrent légèrement selon que l'on choisit la méthode de la sélection d'unités ou de la détermination d'une nouvelle valeur :
- Nom de la variable.
- Nom du tableau dans la base de données.
- Identification de l'unité (dans le cas des corrections individuelles).

- Expression qui détermine l'ensemble d'unités pour lesquelles des corrections doivent être effectuées (dans le cas des corrections systématiques).
- Tableau dans la base de données à partir duquel l'expression qui précède peut être calculée. Il s'agit habituellement du même tableau que celui qui comprend la valeur de la variable, mais pas nécessairement.
- Valeur fixe si la première méthode de détermination de la valeur corrigée est choisie.
- Expression arithmétique si la deuxième méthode de détermination de la valeur corrigée est choisie.
- Limite inférieure et supérieure pour la nouvelle valeur si la troisième méthode de détermination de la valeur corrigée est choisie.

3.3 Imputation des données manquantes

Jusqu'à maintenant, sept méthodes d'imputation différentes peuvent être utilisées, et elles se répartissent en deux groupes différents : paramétriques et non paramétriques. La méthode paramétrique représente en fait un groupe de méthodes, et l'utilisateur peut utiliser le nombre arbitraire de paramétrages différents pour créer un nombre arbitraire de méthodes différentes. Par ailleurs, la méthode non paramétrique n'a pas besoin de paramétrage, ce qui signifie que l'utilisateur ne peut pas créer de méthode « personnalisée », mais qu'il doit quand même insérer les métadonnées requises.

Afin d'illustrer comment le système fonctionne, nous présentons ici la façon dont le groupe de méthodes hot-deck doit être géré. Lorsque l'on décide qu'une certaine variable sera imputée au moyen de la méthode hot-deck, une procédure en deux étapes est nécessaire. À la première étape, le paramétrage de la méthode doit être effectué. Dans le cas de la méthode hot-deck, cela signifie que l'utilisateur choisit jusqu'à cinq variables de stratification et la variable d'appariement. Ces variables doivent évidemment faire partie de l'ensemble de données entrantes. Lorsque les paramètres sont déterminés, l'utilisateur désigne la méthode au moyen de la dénotation, par exemple, HD1. Si le paramétrage approprié de la méthode hot-deck a déjà été défini pour une autre imputation de variable, la première étape de la procédure peut évidemment être omise. À la deuxième étape, les métadonnées de l'imputation de la variable particulière doivent être insérées.

Les métadonnées requises sont les suivantes :
- Nom de la variable à imputer.
- Nom du tableau dans la base de données.
- Dénotation de la méthode d'imputation à utiliser (dans notre cas, HD1).
- Expression logique qui détermine les unités pour lesquelles des imputations seront effectuées.
- Expression logique qui détermine les unités qui ont pu être traitées comme donneurs possibles.
- Étape (de 1 à n) pendant laquelle la méthode particulière pour cette variable particulière sera exécutée. Le système fournit la possibilité d'exécuter la procédure d'imputation en plusieurs étapes consécutives.

4. Utilisation des données administratives – avantages et inconvénients

Le Bureau statistique de la République de Slovénie (SORS) a une longue histoire d'utilisation réussie des données provenant de registres et d'autres sources administratives et s'est toujours défini comme un bureau de la statistique axé sur les registres. Par conséquent, la décision d'utiliser de façon exhaustive différentes sources administratives dans le cas de l'EU-SILC a constitué un choix naturel, dans une certaine mesure, à l'étape de la planification de la conception de l'enquête. Même si une telle utilisation comporte de nombreux avantages, elle a aussi certains « effets secondaires » qui rendent le processus statistique plus exigeant et pose tout un défi pour les concepteurs du processus statistique. La plupart de ces défis sont liés aux parties de l'intégration et de la vérification du processus. Ci-après, nous présentons les principaux avantages et inconvénients de l'utilisation des données administratives dans l'EU-SILC, qui ont été déterminés grâce à une discussion avec le personnel d'enquête.

Les principaux avantages suivants sont mentionnés :
- Du fait de l'utilisation de données administratives, le questionnaire utilisé sur le terrain est beaucoup plus court, et la durée de l'interview est raccourcie de façon significative. Nous estimons que, dans le cas du questionnaire « complet », la durée de l'interview devrait se situer à environ 60 minutes, la durée moyenne actuelle étant d'environ 25 minutes.
- Étant donné que le questionnaire est plus court, il est plus facile de maintenir un niveau raisonnablement faible de non-réponse partielle et de cas de non-réponse. Nous pouvons à juste titre présumer que, dans le cas du questionnaire « complet », il serait difficile d'obtenir un taux de cas de non-réponse de 25 % (approximativement) comme actuellement.
- Les questions les plus difficiles et délicates (liées au revenu) sont laissées de côté, parce que les données qu'elles servent à recueillir sont recueillies auprès de sources administratives. Par conséquent, le risque de non-réponse partielle est beaucoup plus faible, et un moins grand nombre d'imputations des valeurs manquantes sont requises.

- On assiste essentiellement à une réduction des coûts en raison de l'utilisation de données administratives. C'est donc dire que si toutes les données devaient être recueillies au moyen du questionnaire, aucune interview téléphonique ne serait possible, et les coûts de l'enquête seraient beaucoup plus élevés.

Les principaux inconvénients, qui correspondent aux défis liés au développement futur, sont les suivants :

- Comme les données proviennent de nombreuses sources différentes, le processus statistique est beaucoup plus exigeant et long. Lorsque les données proviennent de sources différentes, nous faisons inévitablement face à de nombreux problèmes de couplage et de cohérence, ce qui fait de l'étape de la vérification une étape cruciale du processus statistique.
- Comme c'est habituellement le cas pour l'utilisation des données administratives, il existe des problèmes d'incohérence dans les définitions des variables. Par conséquent, toutes les sources devraient être étudiées de façon soigneuse et constante, afin de déceler les différences conceptuelles possibles.
- Il existe un problème d'actualité, du fait de l'arrivée tardive de certaines sources de données (particulièrement de données fiscales). Les résultats sont diffusés environ neuf mois plus tard que pour une enquête complète sur le terrain.

5. Conclusions

L'EU-SILC est pour le SORS (et probablement pour la plupart des organismes statistiques européens) l'une des enquêtes les plus exigeantes, longues et coûteuses. Une somme importante de résultats statistiques (harmonisés) doit être produite et publiée au niveau national, ainsi qu'au niveau agrégé de l'Europe. Même si les résultats à la sortie de l'enquête sont prescrits entièrement par un règlement, le choix de la méthode de collecte des données requises est laissé à chaque pays. Dans le cas de la Slovénie, les microdonnées sont recueillies auprès de trois sources différentes : une partie des données sont recueillies au moyen d'une enquête sur le terrain, une partie, à partir d'autres sources statistiques, et une (grande) partie, à partir de différentes sources administratives.

Même si l'utilisation de différentes sources administratives comporte de nombreux avantages, elle a aussi certains « effets secondaires », qui rendent le processus statistique plus exigeant et présente tout un défi pour les concepteurs du processus statistique. La plupart de ces défis sont liés aux parties de l'intégration et de la vérification du processus. Afin de faciliter l'étape très exigeante du traitement des données, une nouvelle application générique, fondée sur les métadonnées, a été conçue. La principale modification en ce qui a trait à la pratique qui précède est que le personnel spécialisé a beaucoup plus d'influence sur les décisions relatives au traitement des données. Ce fait est habituellement mentionné comme l'un des principaux avantages du nouveau système.

Comme l'application a été élaborée récemment, il y a encore beaucoup de place pour effectuer des améliorations et des mises au point. Le principal objectif du développement futur est de mettre à jour l'application, de façon à ce qu'elle puisse être utilisée pour d'autres enquêtes. Pour atteindre cet objectif, nous devons d'abord élaborer un outil plus intégré, qui facilitera la gestion de l'ensemble du processus et représentera le principal défi pour les prochaines années.

Bibliographie

Banff Support Team. *Functional Description of the Banff System for Edit and Imputation System*, Statistics Canada, Quality Assurance and Generalized Systems Section, Technical Report.

European Parliament and Council of the European Union (2003). *Regulation of the European Parliament and of the Council of 16 June 2003 concerning Community Statistics on Income and Living Conditions (EU-SILC).*

Eurostat (2004). *Imputation procedures*, EU-SILC Documents 136/04, European Commission, Eurostat.

Eurostat (2004). *Description of Target Variables: Cross-Sectional and Longitudinal*, EU-SILC document 065/04.

Fellegi, I.P. et Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation, *Journal of the American Statistical Association*, March 1976, Volume 71, No. 353, 17-35.

LA VÉRIFICATION DES DONNÉES LONGITUDINALES DE L'ENQUÊTE ITALIENNE SUR LA POPULATION ACTIVE

Simona Rosati et Barbara Boschetto[1]

Résumé

Bien que le principal produit de l'Enquête sur la population active de l'Italie soit des données transversales, l'obtention de données longitudinales est un objectif de grande importance. Néanmoins, la dimension temporelle rend plus difficile l'élaboration d'une stratégie permettant de traiter la non-réponse partielle ainsi que la non-réponse totale. Le présent article est consacré principalement à la méthode d'imputation longitudinale appliquée pour corriger la non-réponse partielle à des questions individuelles. Nous accordons une attention particulière aux problèmes d'incohérence des données et aux difficultés que pose l'utilisation de données longitudinales. L'article se conclut par une discussion des principaux résultats. Nous donnons également une ample explication du processus de couplage des enregistrements. Tous les résultats sont présentés pour la période de 2007 (1) à 2008 (1).

Mots clés : non-réponse partielle, vérification des données, imputation longitudinale.

1. Introduction à l'Enquête sur la population active

L'Enquête sur la population active (EPA) de l'Italie est une enquête continue réalisée au cours de chaque semaine d'une année et conçue pour produire des estimations annuelles, trimestrielles et mensuelles des principaux indicateurs du marché du travail.

La structure du plan de sondage permet également de reconstruire des microdonnées longitudinales. Le plan de sondage est en fait fondé sur un échantillon rotatif et les ménages sont interviewés quatre fois sur une période de 15 mois selon un plan de rotation 2-2-2 : les ménages participent pendant deux trimestres consécutifs, puis ils sortent temporairement de l'échantillon pendant les deux trimestres suivants, et enfin reviennent dans l'échantillon pendant deux autres trimestres consécutifs, après quoi ils sortent définitivement de l'enquête. Après quatre interviews, le ménage est remplacé par un autre vivant dans la même région ou une région similaire.

En particulier, l'échantillon de ménages de chaque trimestre comprend quatre groupes de renouvellement ou panels, qui participent, respectivement, à la première, à la deuxième, à la troisième et à la quatrième interview. Chaque série d'interviews auprès d'un panel constitue une vague. Comme le montre l'exemple de la figure 1-1, les ménages échantillonnés qui sont interviewés pour la première fois au premier trimestre de 2007, désignés comme étant le groupe G1, sont de nouveau interviewés au deuxième trimestre de 2007 (G2), puis après une pause de deux trimestres, sont interviewés au premier et au deuxième trimestre de 2008 (G3 et G4, respectivement). Ce plan de rotation implique que la moitié des ménages interviewés durant un trimestre sont réinterviewés après trois mois, une moitié après douze mois et un quart après quinze mois.

[1]Simona Rosati et Barbara Boschetto, Italian National Institute of Statistics, Via Adolfo Ravà 150, Roma, Italy, 00142 (courriel: sirosati@istat.it; boschett@istat.it)

Figure 1-1
Plan de rotation du plan de sondage (EPA)

Trimestre	Panel								
2006(4)	A4	B3			E2	**F1**			
2007(1)		B4	C3			**F2**	**G1**		
2007(2)			C4	D3			**G2**	H1	
2007(3)				D4	E3			H2	I1
2007(4)					E4	**F3**			I2
2008(1)						**F4**	**G3**		
2008(2)							**G4**	H3	

Les interviews des mêmes personnes à différentes périodes représentent l'information longitudinale de l'enquête. Il est important de se souvenir que l'EPA n'a pas été conçue comme une enquête longitudinale, si bien que les personnes qui sortent des ménages sélectionnés ou les personnes qui quittent la municipalité, ne sont pas interviewées. Est toutefois exclue la partie de la population qui se déplace dans le territoire de l'analyse longitudinale du marché du travail. Par conséquent, la population longitudinale est définie comme étant la population qui réside dans la même municipalité pendant une période donnée.

Un autre facteur important est que le plan de sondage comporte le remplacement d'un ménage en cas de sortie de l'échantillon pour diverses raisons (refus, adresse incorrecte, etc.), de sorte que l'échantillon de ménages est effectivement composé du « ménage de base » et de ses trois « ménages de remplacement » (ledit « quartet »). Manifestement, en cas de remplacement d'un ménage, la population longitudinale sera plus faible que la population transversale.

Les interviews sont réalisées au moyen d'une combinaison de techniques d'interview assistée par ordinateur (IA) par un réseau d'intervieweurs professionnels. La règle générale est que l'interview sur place assistée par ordinateur (IPAO) est utilisée pour la première interview, tandis que l'interview téléphonique assistée par ordinateur (ITAO) est privilégiée pour les interviews suivantes, dans la mesure du possible.

Le recours à des techniques d'IAO permet de s'assurer que la collecte des données est de haute qualité et est il est de surcroît particulièrement utile pour appuyer la composante par panel de l'enquête. Le questionnaire électronique permet de gérer l'information fournie par les répondants aux interviews antérieures, de sorte qu'il est possible de mener les interviews qui suivent la première en demandant à la personne échantillonnée de confirmer ou non l'information donnée à la vague précédente (la question n'est alors posée de nouveau que s'il n'y a pas de confirmation).

Ce système offre de nombreux avantages (quand la situation d'activité ou d'autres caractéristiques du répondant n'ont pas changé beaucoup) : non seulement il réduit considérablement la durée de l'interview et la possibilité d'erreurs dues à l'intervieweur, mais il permet aussi la correction immédiate de certaines erreurs enregistrées éventuellement durant une interview précédente. Ce facteur, comme nous l'expliquerons plus loin, constitue un soutien important du processus de correction des données longitudinales.

En outre, l'utilisation du questionnaire électronique facilite grandement l'appariement des données sur une même personne pour construire les microdonnées longitudinales. L'attribution d'un identificateur clé unique à chaque membre du ménage permet de s'assurer que les données recueillies à diverses périodes sont effectivement couplées à la même personne. Le questionnaire électronique enregistre également tous les changements qui surviennent au sein de la famille, à cause de la sortie ou de l'entrée de certains membres entre une interview et la suivante. Cela donne directement le nombre de personnes absentes à chaque moment de l'enquête et permet d'en tenir compte durant le processus de couplage des enregistrements.

2. Le processus de couplage des enregistrements

Le système de rotation de l'échantillon de l'EPA implique que 50% de l'échantillon trimestriel est de nouveau interviewé après trois, puis douze mois. Même si l'objectif est d'obtenir les flux sur le marché du travail après douze mois, nous avons choisi de coupler les quatre vagues de l'enquête, pour chaque groupe de renouvellement de l'échantillon trimestriel, afin de pouvoir utiliser toute l'information disponible durant la correction, afin d'être certain d'obtenir l'information la plus cohérente possible. Cela signifie que la procédure de couplage des enregistrements a été appliquée au départ aux personnes qui avaient participé à deux vagues après douze mois, et que l'information recueillie à deux autres vagues (après trois et quinze mois) a été reliée plus tard.

La population longitudinale de l'EPA est définie comme étant la population qui réside dans la même municipalité pendant une période donnée. Pour cette raison, en ayant choisi la période de référence pour le couplage des enregistrements après douze mois (ici, le premier trimestre de 2007 et le premier trimestre de 2008), nous avons comparé deux ensembles de données sources pour identifier la population de référence du fichier longitudinal. En fait, la population initiale peut changer durant une année pour plusieurs raisons, dont :

- la non-réponse des ménages d'un « quartet » (c'est-à-dire toutes les familles incluses dans le « quartet »);
- le remplacement d'un ménage non répondant par une autre famille dans le même « quartet », de sorte que ce ménage ne sera pas présent à l'interview suivante;
- l'existence de personnes qui ne sont plus membres du ménage à cause d'un décès ou du transfert dans une autre municipalité;
- l'existence de nouveaux membres du ménage à cause d'une naissance ou d'un transfert en provenance d'une autre municipalité.

La présence d'un code d'identification individuel unique au sein de la famille (clé individuelle, composée de six chiffres), facilite le couplage des enregistrements, mais le code peut contenir des erreurs, de sorte qu'une procédure de vérification déterministe a été appliquée en se servant de variables particulières qui nous permettent de repérer, d'une part, les personnes qui sont appariées correctement et, d'autre part, les « cas faussement négatifs » (quand une personne possède deux clés individuelles différentes) et les « cas faussement positifs » (quand une même clé individuelle est associée à deux personnes différentes).

En ne prenant en considération que les ménages interviewés aux deux trimestres appariés, nous avons obtenu un fichier composé de personnes « appariables » (c'est-à-dire des personnes pouvant faire l'objet d'un couplage). Dans ce fichier, nous avons distingué les paires de personnes couplées des personnes non couplées au moyen de la clé individuelle.

Le couplage des enregistrements est contrôlé par un processus déterministe qui vérifie la concordance de certains identificateurs personnels importants, à savoir la date de naissance, le sexe et le nom. Après ce contrôle, un code particulier est attribué afin d'indiquer si la correspondance de ces variables est exacte ou non.

De cette façon, les personnes appariées correctement ont été identifiées immédiatement, parce que, après une année, elles possédaient la même clé individuelle et les mêmes identificateurs personnels.

L'attribution d'une clé individuelle par erreur est un événement possible, même s'il est extrêmement rare, de sorte que les personnes qui n'ont pas été appariées d'après la clé individuelle ou celles pour lesquelles les données personnelles recueillies aux deux trimestres ne concordaient pas ont été comparées les unes aux autres au sein de la même famille au moyen des identificateurs personnels (date de naissance, sexe et nom) uniquement.

Les personnes pour lesquelles la concordance était parfaite pour toutes les variables (sauf la clé individuelle) ont été recouvrées.

À ce stade, l'ensemble de données initial de personnes « appariables » (composé de 79 151 personnes dans les ménages interviewés en 2007 (1) et en 2008 (1)), a été divisé comme il suit (tableau 2-1) :

- les personnes entièrement couplées (76 985, représentant 97,3 %) englobant :
  - les personnes appariées correctement au moyen de la clé individuelle et dont les renseignements personnels étaient cohérents (75 854, représentant 95,8 % du total des personnes « appariables »);
  - les personnes dont la clé individuelle était incorrecte (1 131, soit 1,4 % du total) englobant :
    - les personnes couplées qui ont été recouvrées de la quantité initiale de « cas faussement positifs » (personnes ayant la même clé individuelle, mais des valeurs non concordantes des variables personnelles, extraites par échange des clés : 740 personnes, soit 0,9 % des cas);
    - les personnes non couplées ou « cas faussement négatifs » (personnes ayant des clés individuelles différentes, mais des valeurs concordantes des variables personnelles, extraites par échange des clés : 391 personnes, soit 0,5 % des cas);
- les personnes entièrement non appariées (2 166 représentant 2,7 % du total des personnes appariables) englobant :
  - les « cas faussement positifs » (personnes ayant la même clé individuelle, mais des données personnelles différentes : 567 personnes, soit 0,7 %);
  - les personnes présentes dans une seule vague de la période visée par l'analyse (1 599 personnes, soit 0,2 % du total).

Tableau 2-1
Taux de couplage des enregistrements

Personnes dans les ménages interviewées en 2007 (1) et en 2008 (1)	Nombre	Pour cent
Enregistrements appariables	79 151	100,0
- Enregistrements couplés	76 985	97,3
-- Enregistrements correctement appariés	75 854	95,8
-- Enregistrements contenant une clé individuelle incorrecte	1 131	1,4
---- Enregistrements couplés recouvrés parmi les cas faussement positifs par échange de clés	740	0,9
---- Enregistrements non couplés recouvrés parmi les cas faussement négatifs par échange de clés	391	0,5
- Total des enregistrements non appariés	2 166	2,7
-- Cas faussement positifs	567	0,7
-- Enregistrements individuels présents dans une seule vague de la période visée par l'analyse	1 599	2,0

Les résultats de cet appariement ont révélé en premier lieu le haut degré de fiabilité de l'attribution des clés individuelles (75 854 enregistrements appariés au moyen de la clé correcte, ce qui représente 98,5 % des 76 985 enregistrements individuels appariés). En deuxième lieu, les résultats ont montré l'efficacité de la procédure de contrôle déterministe et de couplage des enregistrements décrite plus haut. En effet, d'une part, elle permet de valider l'exactitude de la clé individuelle et, d'autre part, de recouvrer les enregistrements individuels possédant une clé incorrecte. Même si cette procédure n'est pas automatique, elle a l'avantage d'être subdivisée en phases distinctes et répétables, de sorte qu'elle peut être appliquée au couplage des enregistrements d'autres ensembles de données.

Afin de terminer la reconstruction du fichier longitudinal, la même procédure de couplage des enregistrements a été utilisée pour ajouter l'information provenant des autres vagues en vue d'obtenir toute l'information pour chacun des groupes de renouvellement et de l'utiliser à la phase de correction. Pour cela, nous avons divisé l'échantillon de personnes appariées après douze mois en deux sous-échantillons et nous avons associé chacun d'eux à l'information provenant de trimestres différents. Comme le montre en effet la figure 1-1, l'échantillon de personnes appariées après douze mois est composé de deux groupes de renouvellement, F et G : pour le groupe F, les données des deuxième et quatrième vagues sont appariées, tandis que pour le groupe G, ce sont celles des première et deuxième vagues. Pour chaque groupe, nous avons rattaché au groupe F la première interview qui a eu lieu au quatrième trimestre de 2006, et la troisième, qui a eu lieu au quatrième trimestre de 2007, et nous avons rattaché au groupe G la deuxième interview qui a eu lieu au deuxième trimestre de 2007, et la quatrième, qui a eu lieu au deuxième trimestre de 2008.

Après le processus de couplage des enregistrements, nous avons obtenu deux fichiers de données longitudinales contenant l'information provenant de chaque vague de l'enquête. Ces fichiers seront utilisés séparément à la phase de correction et, enfin, seront fusionnés en vue d'analyser les données recueillies sur toutes les personnes interviewées du premier trimestre de 2007 au premier trimestre de 2008.

3. Imputation longitudinale

3.1 Non-réponse longitudinale

Dans les enquêtes longitudinales, deux types de non-réponse peuvent se produire, à savoir la non-réponse à une question (non-réponse partielle) et la non-réponse à une vague (non-réponse totale). La non-réponse à une vague a lieu quand une unité omet de fournir des données pour une ou plusieurs vagues de l'enquête. La non-réponse partielle a lieu quand des éléments de données manquent à n'importe quelle vague; sont également incluses les valeurs logiquement incohérentes entre deux vagues pour deux ou plusieurs variables, ainsi que les réponses situées en dehors d'une fourchette acceptée de valeurs. En général, la correction de la non-réponse à une vague est effectuée par repondération et celle de la non-réponse partielle, par imputation. Néanmoins, la dimension temporelle rend difficile l'élaboration d'une stratégie permettant de traiter la non-réponse partielle ainsi que la non-réponse totale.

La présente étude porte sur l'imputation pour la correction de la non-réponse partielle dans les enquêtes qui recueillent des données longitudinales. Il importe de souligner que, dans ce cas, la non-réponse partielle longitudinale comprend seulement les données non concordantes entre deux vagues en raison de la stratégie d'imputation. En fait, le fichier de données longitudinales de l'EPA est obtenu en partant de fichiers de données transversales individuels qui ont chacun été corrigés antérieurement pour la non-réponse partielle dans le contexte transversal (Ceccarelli et Rosati, 2005). De cette façon, le processus d'imputation est

divisé en un plus grand nombre d'étapes afin de réduire le coût informatique et de s'assurer que les estimations soient produites dans les délais prévus.

En fonction du processus de couplage des enregistrements, les membres du panel peuvent être répartis entre trois groupes distincts, à savoir :

a) les cas de réponse complète, c'est-à-dire les unités qui ont fourni toutes les données à chacune des quatre vagues;
b) les non-répondants à une vague, c'est-à-dire les unités qui ont omis de fournir les données à l'une des quatre vagues;
c) les non-répondants à deux vagues, c'est-à-dire les unités qui ont omis de fournir les données à deux des quatre vagues.

Les unités qui ont omis de fournir des données pour au moins une vague de la période de référence de l'analyse ne sont pas incluses (dans la présente étude, 2,0 % du total des enregistrements « appariables », comme le montre le tableau 2-1).

Dans le tableau 3.1-1, sont présentés les profils de réponse admissibles de l'EPA résultant du couplage des enregistrements. Comme le montre le tableau, un nombre important de membres du panel ont répondu à chacune des quatre vagues de l'enquête (en pourcentage, les valeurs pour les groupes F et G sont de 81,3 et 85,9, respectivement). Les personnes qui ont répondu à trois des quatre vagues représentaient environ 17% du groupe F, et 11 % du groupe G, tandis qu'une faible proportion seulement n'a participé qu'à deux des quatre vagues (environ 2,4 % en tenant compte des deux groupes de renouvellement).

Tableau 3.1-1
Profils de réponse aux quatre vagues pour les groupes F et G

Situation de réponse	Vague 1	Vague 2	Vague 3	Vague 4	Nombre	Pour cent
Complète	F1	F2	F3	F4	31 936	81,3
Non-répondants à une vague	F1	F2	---	F4	2 589	6,6
	---	F2	F3	F4	4 058	10,3
Non-répondants à deux vagues	---	F2	---	F4	704	1,8
Complète	G1	G2	G3	G4	32 413	85,9
Non-répondants à une vague	G1	G2	G3	---	3 116	8,3
	G1	---	G3	G4	1 050	2,8
Non-répondants à deux vagues	G1	---	G3	---	1 119	3,0

--- Non-répondant

Il convient de noter que le taux de réponse à l'EPA est assez élevé (variant de 76 % à 83 % durant la période de référence), ce qui indique que les données longitudinales de l'enquête peuvent être considérées comme fiables pour l'analyse, en particulier si celle-ci porte sur les données recueillies pour trois ou douze mois, ce qui représente environ 50 % des données de l'échantillon trimestriel.

3.2 Stratégie d'imputation

Plusieurs scénarios de classification sont proposés dans la littérature pour les méthodes d'imputation. Toutefois, comme l'exposent Kalton et Kasprzyk (1982), on peut distinguer deux grandes catégories d'imputation : déterministe et probabiliste. L'*imputation déterministe* consiste à attribuer une seule valeur, déterminée a priori, en se fondant sur d'autres valeurs considérées comme « réelles » par les experts. Au contraire, l'*imputation probabiliste* consiste à attribuer une valeur en s'appuyant sur un modèle stochastique (c'est-à-dire un modèle de régression) ou en utilisant une unité semblable à celle qui doit être imputée. Néanmoins, comme l'ont fait remarquer les mêmes auteurs, la plupart des méthodes d'imputation entrent dans le cadre général de la régression multiple (Kalton et Kasprzyk, 1986).

Il est reconnu que de nombreux éléments doivent être examinés afin de choisir une méthode d'imputation appropriée. Pour les besoins de notre étude, nous avons décidé pour plusieurs raisons d'élaborer une méthode déterministe longitudinale. Étant donné que l'interview peut être réalisée auprès du répondant proprement dit ou auprès d'un autre membre répondant du ménage, la situation d'interview (par personne interposée ou en personne) peut être utilisée comme variable auxiliaire pour évaluer la fiabilité des réponses non concordantes entre les vagues. Deuxièmement, sachant que l'ITAO est une méthode d'interview avec rétro-information, c'est-à-dire que les réponses données à la vague précédente sont utilisées dans l'énoncé de la question afin de rappeler au répondant les réponses antérieures ou de demander des éclaircissements au sujet d'incohérences

entre les variables, la variable de la confirmation de l'information antérieure a également été intégrée comme variable auxiliaire. De surcroît, les antécédents professionnels complets de chaque personne, de la première vague à la quatrième, ont été examinés afin de dégager des règles déterministes générales d'imputation longitudinale. Ces règles nous ont permis de déduire la valeur imputée d'après les données disponibles à la vague courante ou précédente sur la base de la relation entre les variables auxiliaires et la non-réponse partielle.

Plus précisément, la stratégie d'imputation permet de corriger les erreurs associées aux deux premières vagues en exécutant un algorithme qui impute la variable incorrecte en modifiant la valeur enregistrée à la première ou à la deuxième vague, conformément aux règles hiérarchiques suivantes :
1) si la méthode d'ITAO a été utilisée à la deuxième vague, donc qu'une réponse a été donnée à la question de « confirmation », la valeur de la variable à la première vague est imputée en se fondant sur la réponse donnée à la deuxième vague si le répondant rejette la réponse donnée à la première vague;
2) si la méthode d'ITAO n'a pas été utilisée à la deuxième vague, la valeur de la variable à la première vague est imputée en utilisant l'information fournie en personne par le répondant à la deuxième vague; inversement, la valeur de la variable à la deuxième vague est imputée au moyen de la valeur fournie en personne par le répondant à la première vague;
3) si ni l'une ni l'autre des conditions qui précèdent n'a pu être appliquée, la valeur de la variable à la première vague est imputée en attribuant la valeur de la variable correspondante à la deuxième vague. Cette règle a été déduite de l'examen de la cohérence des réponses à une question répétée à chacune des quatre vagues.

Aux vagues subséquentes, l'imputation longitudinale consiste à utiliser les données provenant de la vague précédente pour imputer les donnes non concordantes à la vague courante (*imputation conditionnelle*). Quoique cette règle ajoute des contraintes supplémentaires au processus complet, elle fait en sorte qu'une exigence fondamentale de l'EPA soit satisfaite, c'est-à-dire l'obtention d'estimations finales n'ayant pas besoin d'être révisées quand les données d'une nouvelle vague sont disponibles.

3.3 Résultats

L'effet de l'imputation est résumé brièvement au tableau 3.3-1. Ici, seuls les résultats concernant le groupe de renouvellement F sont présentés. Comme le montre le tableau, le taux d'imputation associé aux deux premières vagues d'interview est égal à 9,8 %. Autrement dit, les règles déterministes énoncées plus haut ont produit 3 838 des 39 287 enregistrements dans lesquels au moins la valeur d'une variable a été modifiée à la première ou à la deuxième vague. En ce qui concerne l'imputation conditionnelle, deux calculs distincts doivent être faits. L'un a trait aux enregistrements qui représentent des répondants ayant participé à trois vagues ou plus, et le résultat est de 9,3 % (vague 3, conditionnellement à la participation à la vague 2); l'autre concerne les enregistrements des cas de réponse complète (c'est-à-dire les personnes qui ont répondu à chacune des quatre vagues) et le résultat était de 13,8 % (vague 4, conditionnellement à la participation à la vague 3).

Il convient de mentionner que nombre de variables visées par l'imputation ont trait à des questions rétrospectives (par exemple, mois et année où l'emploi a débuté, mois et année où l'emploi s'est terminé), qui sont les plus sujettes à des erreurs, telles que des problèmes de remémoration et des problèmes de conditionnement du panel, au fil du temps (Kalton et coll., 1989).

Tableau 3.3-1
Enregistrements selon le type d'imputation (groupe F)

Type d'imputation	Nombre	Pour cent
Vague 1 ou vague 2	3 838	9,8
Vague 3, conditionnellement à la vague 2	3 592	9,3
Vague 4, conditionnellement à la vague 3	4 415	13,8

3.4 Conclusion

La méthode d'imputation que nous avons élaborée a révélé certaines bonnes propriétés, mais d'autres travaux de recherche seront nécessaires pour évaluer les effets de l'imputation. Une analyse plus approfondie doit être effectuée afin de déterminer l'incidence de l'imputation sur la distribution des données.

Même si notre méthode est fondée sur l'approche déterministe, elle semble convenir pour l'EPA courante. En effet, nous avons utilisé des variables auxiliaires pour déduire les règles d'imputation appropriées et pour réduire le biais d'imputation. En outre, la méthode est utilisable en cas de non-réponse partielle à une vague pour une grande gamme de variables ainsi que pour divers

profils de répondants. Néanmoins, l'imputation conditionnelle est nécessaire pour s'assurer d'obtenir des estimations longitudinales définitives pour chaque trimestre.

Comme prévu, les variables modifiées le plus fortement par l'imputation sont celles liées aux dates. Par conséquent, nous proposons d'introduire un ensemble de règles de vérification longitudinale dans le questionnaire en vue de réduire les incohérences durant l'interview.

Il convient également de mentionner que deux autres stratégies d'imputation longitudinale ont été mises en œuvre dans le cadre de l'ancienne EPA qui a été réalisée jusqu'au premier trimestre de 2004. L'une d'elles comportait une combinaison d'imputation déterministe et probabiliste (Rosati, 2004). L'expérience a révélé que les méthodes généralisées peuvent être adaptées afin de corriger les données longitudinales, bien qu'elles puissent nécessiter d'importantes ressources informatiques. Cependant, cette solution ne semblait pas convenir pour la nouvelle EPA.

Nous pouvons conclure que notre méthode représente un outil utile pour l'imputation des données longitudinales dans les grandes enquêtes en vue de rendre les microdonnées intérieurement cohérentes. Bien que l'entièreté du système d'imputation, y compris le processus de couplage des enregistrements qui en fait partie intégrante, soit entièrement déterministe, il a l'avantage d'être reproductible pour tout ensemble de données de l'EPA et, par conséquent, pour tout ensemble de données d'enquête semblable.

Bibliographie

Ceccarelli, C. et Rosati, S. (2006). Data Editing for the Italian Labour Force Survey, Statistical Data Editing: Impact on Data Quality, Vol. No. 3, *United Nations Economic Commission for Europe Work Sessions on Statistical Data Editing*, pp. 291-300.

Kalton, G. et Kasprzyk, D. (1982). Imputing for Missing Survey Responses, *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, pp. 22-31.

Kalton, G. et Kasprzyk, D. (1986). The Treatment of Missing Survey Data, *Survey Methodology*, 12, pp. 1-16.

Kalton, G., Kasprzyk, D. et Mc Millen, D.B. (1989). Nonsampling Errors in Panel Surveys, in *Panel Surveys*, New York: Wiley, pp. 249-270.

Rosati, S. (2004). Longitudinal Imputation for the Quarterly Labour Force Survey, *Atti della XLII Riunione Scientifica*, *Socieà Italiana di Statistica*, pp. 335-338.

Imputation et dérivation de registres longitudinaux : le cas des ménages

Laan, D. J. van der et Kuijvenhoven, L.[1]

Résumé

Les registres peuvent être une riche source de données pour les analyses longitudinales. Toutefois, nombre de stratégies de vérification et d'imputation sont axées sur des analyses transversales, ce qui crée des incohérences longitudinales à un microniveau. À Statistics Netherlands, la composition des ménages est établie d'après les registres de population municipaux. Cependant, la procédure d'établissement de la composition et d'imputation ne s'appuie que sur les données pour la période considérée, ce qui rend les analyses longitudinales impossibles. Nous présentons les modifications apportées à la méthode actuelle afin que les ménages puissent faire l'objet d'analyses longitudinales. Nous veillons tout particulièrement à ce que les estimations transversales et les estimations des variations soient exactes.

Mots-clés : registres, longitudinal, imputation, ménages.

1. Introduction

Les registres peuvent être une riche source de données pour les analyses longitudinales. Toutefois, nombre de stratégies de vérification et d'imputation sont axées sur des analyses transversales, ce qui crée des incohérences longitudinales à un microniveau. Afin que les données conviennent pour les analyses longitudinales, il est nécessaire de tenir compte des renseignements recueillis pour d'autres périodes lorsqu'il faut dériver et imputer les variables.

Wallgren et Wallgren (2007) distinguent trois attributs de la qualité des registres statistiques, à savoir la qualité transversale, la qualité chronologique et la qualité longitudinale. Si un registre possède une qualité transversale, des comparaisons peuvent être faites à l'intérieur du registre. S'il possède une qualité chronologique, des comparaisons peuvent être faites dans le temps à un niveau agrégé. S'il possède une qualité longitudinale, des comparaisons peuvent être faites dans le temps à un microniveau.

À Statistics Netherlands, la composition des ménages est établie d'après les registres de population municipaux depuis 2000. Les données sur les relations (familiales) qui figurent dans ces registres permettent de déterminer spécifiquement la composition des ménages pour environ 93 % des adresses. Pour les 7 % d'adresses restants, la composition des ménages est imputée en se servant d'un modèle d'imputation stochastique qui tient compte des caractéristiques des personnes habitant aux adresses en question. Néanmoins, comme la procédure d'établissement de la composition et d'imputation ne tient compte que des données pour la période considérée, les ménages ainsi dérivés ne conviennent pas pour les analyses longitudinales. Nous présentons les modifications apportées à la méthode actuelle pour que les données sur les ménages puissent faire l'objet d'analyses longitudinales. Nous veillons tout particulièrement à ce que les estimations transversales et les estimations des variations soient exactes.

Le fait que les bureaux de la statistique reçoivent périodiquement de nouvelles données, mais qu'ils sont tenus de respecter un échéancier de publication constitue un problème particulier, car les nouvelles données pourraient contenir de nouveaux ou meilleurs renseignements au sujet des données dérivées antérieurement. Nous montrons qu'en général, il est impossible d'obtenir un registre longitudinal cohérent sans apporter de corrections aux données dérivées antérieurement. Nous exposons la façon dont il est prévu de traiter cette situation à Statistics Netherlands.

[1]D.J. van der Laan, Statistics Netherlands, Henri Faasdreef 312, 2492 JP The Hague, The Netherlands, (dj.vanderlaan@cbs.nl); L. Kuijvenhoven, Statistics Netherlands, Henri Faasdreef 312, 2492 JP The Hague, The Netherlands, (l.kuijvenhoven@cbs.nl)

2. Établissement déterministe de la composition des ménages

Les chiffres sur les ménages publiés par Statistics Netherlands sont basés sur la définition du maintien du ménage. Cette définition implique que seules les personnes qui résident à la même adresse et subviennent à leurs besoins quotidiens sont considérées comme formant un ménage.

Les statistiques sur la population et les ménages des Pays-Bas produites par Statistics Netherlands s'appuient sur des registres de population municipaux automatisés. Ce système d'enregistrement est désigné par l'acronyme GBA, Gemeentelijke Basis Administratie persoonsgegevens, le système d'enregistrement municipal de base des données sur la population. Le terme « de base » fait référence au fait que le GBA sert de registre de base des données sur la population au sein d'un système de registres locaux. Pour en savoir plus sur le système GBA, consulter Prins (2000).

Le GBA contient des renseignements sur les liens familiaux. Chaque enregistrement personnel contient des renseignements sur le ou les parents et sur tous les enfants, quel que soit leur lieu de résidence actuel. Il contient également des renseignements sur le partenaire de la personne. Conjugués à l'information détaillée sur l'adresse, ces renseignements permettent d'identifier toutes les familles nucléaires traditionnelles. Cela donne les règles déterministes suivantes :

D1 Si deux personnes appartiennent à la même famille et vivent à la même adresse, elles appartiennent au même ménage.
D2 Si deux personnes sont apparentées et vivent à la même adresse, elles appartiennent au même ménage.
D3 Si deux personnes déménagent à une nouvelle adresse à la même date, elles appartiennent au même ménage.
D4 Si une personne de 15 ans ou moins sans parent identifiable vit à une adresse avec un autre ménage, elle appartient à ce ménage.

3. Utilisation de sources additionnelles

L'une des modifications apportées à la méthode courante d'établissement de la composition des ménages est l'utilisation de sources additionnelles afin de dériver un plus grand nombre de ménages de manière déterministe. Les sources les plus importantes proviennent du ministère des Finances. Ce dernier fournit des renseignements sur l'utilisation 1) des régimes de soutien fiscal pour les soins médicaux, 2) des régimes de soutien au logement et 3) des régimes d'impôt généraux. Le point commun de ces trois systèmes est que le demandeur indique avec qui il ou elle forme un ménage. L'information peut être couplée à celle des registres de population. Une règle supplémentaire peut donc être créée maintenant :

D5 Si une autre source que le GBA indique que deux personnes vivant à la même adresse forment un couple, elles appartiennent au même ménage.

L'utilisation des règles D1 à D4 avec les registres de population permet d'établir la position et la composition des ménages pour environ 92,6 % des adresses. L'utilisation des sources d'information supplémentaires et l'application de la règle D5 permettent de déterminer directement la composition de 94,0 % des ménages (voir le tableau 3-1). En particulier, l'utilisation de ces sources additionnelles permet d'améliorer considérablement les données sur le nombre de couples dans les catégories « deux personnes hors famille » et « famille monoparentale + personne hors famille ».

Tableau 3-1
Réduction du nombre d'adresses nécessitant une imputation quand sont utilisées des sources additionnelles (données pour 2006)

Source	Nombre de personnes dans la source	Nombre à imputer	Réduction
Situation initiale	-	7,4%	-
Soutien pour soins de santé	2 075 460	6,6%	11,3%
Soutien au loyer	886 802	7,3%	2,0%
Remboursement des impôts prévisionnels	3 065 367	6,7%	9,7%
Total des trois sources	-	6,0%	19,5%

4. Imputation stochastique

Comme nous l'avons mentionné à la section précédente, pour 7,4 % (6,0 % en utilisant les sources additionnelles) des adresses, il n'est pas possible d'établir complètement la composition des ménages au moyen de règles déterministes. Cela signifie qu'à la

fin de l'établissement déterministe de la composition, il reste plus d'un « ménage partiellement déterminé » à ces adresses et que nous devons décider si nous voulons ou non continuer de combiner ces ménages pour en former de plus grands.

Nous avons décidé d'imputer la composition des ménages vivant à ces adresses. Malheureusement, il s'agit d'un groupe d'adresses spécial. Par exemple, bon nombre de ces adresses sont celles de couples non mariés. Nous ne pouvons donc pas utiliser l'information venant des adresses pour lesquelles la composition du ménage est entièrement établie et il nous faut d'autres données sur les adresses restantes afin d'estimer le modèle choisi pour l'imputation. Nous utilisons pour cela l'Enquête sur la population active (EPA), qui porte sur un échantillon aléatoire d'adresses. Les intervieweurs prennent contact avec tous les ménages vivant à une adresse donnée. Ainsi, en principe, la composition complète des ménages est connue pour ces adresses. Les données de l'EPA sont couplées aux adresses restantes afin de déterminer la composition des ménages vivant à ces adresses.

Pour l'imputation, nous estimons un modèle qui prédit la composition du ménage en s'appuyant sur les caractéristiques de base de l'adresse (urbanisation) et des personnes vivant à cette adresse (âge, sexe, ethnicité). Des modèles distincts sont estimés pour chacun des groupes suivants : deux personnes, famille et une personne, parent seul et une personne, trois personnes et quatre à neuf personnes. Ces groupes sont suffisamment grands pour permettre une modélisation précise et on peut s'attendre à ce que les probabilités diffèrent pour les compositions qu'il peut y avoir.

Dans la plupart des cas, il existe deux compositions possibles : soit les personnes forment un seul ménage, soit elles forment deux ménages distincts. Par conséquent, nous utilisons un modèle de régression logistique pour prédire les probabilités pour chacune des deux compositions possibles. Lorsque trois personnes vivent à une même adresse, nous utilisons deux modèles logistiques, l'un prédisant la probabilité qu'il existe deux ou trois ménages comparativement à un ménage et l'autre prédisant la probabilité qu'il existe trois ménages comparativement à un ou deux ménages. Le groupe de quatre à neuf personnes vivant à une même adresse est trop petit pour estimer un modèle fondé sur les caractéristiques de base. Nous utilisons donc les fréquences observées comme probabilités.

En nous servant des modèles estimés, nous pouvons calculer, pour chaque adresse nécessitant une imputation, les probabilités de chacune des compositions possibles du ménage. Ces probabilités sont utilisées pour imputer stochastiquement la composition des ménages à ces adresses. L'imputation stochastique est préférée à l'imputation déterministe afin d'être certain que les agrégats soient sans biais, quoique l'imputation déterministe (imputation de la composition la plus probable) produirait une imputation plus correcte à un microniveau. Pour l'imputation stochastique, nous utilisons la méthode de Fellegi (1975) qui donne les mêmes totaux que ceux découlant du modèle.

5. Cohérence longitudinale

La méthode actuelle décrite plus haut pose certaines difficultés quand les compositions des ménages établies d'après les données de diverses périodes sont utilisées pour effectuer des analyses longitudinales. Comme la méthode est appliquée séparément à chaque période, les compositions imputées à diverses périodes ne concordent pas nécessairement les unes avec les autres. Par exemple, avec la méthode courante, il est possible que deux personnes vivant à une adresse soient imputées une année comme deux ménages d'une seule personne et l'année suivante, comme un couple, ce qui donne lieu à une surestimation des variations lorsqu'on utilise les données pour des analyses longitudinales.

Afin d'éviter les incohérences longitudinales, il est nécessaire de combiner les données provenant de plusieurs périodes pour l'établissement de la composition et l'imputation. Il est raisonnable de supposer que la relation entre deux personnes (appartenant ou n'appartenant pas à un seul ménage) ne change pas aussi longtemps que ces personnes vivent à la même adresse. Cette hypothèse mène aux règles longitudinales suivantes :

L1 Si deux personnes, A et B, appartiennent au même ménage au temps t_j, et qu'elles vivent à la même adresse au temps t_i ($i=j$-1 ou $i=j$+1), alors A et B appartiennent au même ménage au temps t_i.
L2 Si deux personnes, A et B, n'appartiennent pas au même ménage au temps t_j et qu'elles vivent à la même adresse au temps t_i ($i=j$-1 ou $i=j$+1), alors A et B n'appartiennent pas au même ménage au temps t_i.

Supposons que nous ayons un ensemble complet de ménages présentant une cohérence longitudinale pour les périodes t_1 et t_2. Si nous appliquons les règles longitudinales, l'algorithme pour l'établissement de la composition du ménage pour la période t_3 est alors le suivant :

1. Appliquer les règles déterministes (voir la section 2) pour la période t_3.

2. Transférer de manière déterministe l'information dérivée de la période t_2 à la période t_3 (=en avant) en appliquant la règle longitudinale L1.
3. Retransférer cette information dérivée de manière déterministe à la période t_2, puis à la période t_1 (=vers l'arrière) en appliquant la règle longitudinale L1.
4. Corriger les imputations stochastiques faites à la période t_1 et transférer l'information imputée de manière stochastique de la période t_1 à la période t_2 en appliquant les règles L1 et L2.
5. Corriger les imputations stochastiques faites à la période t_2 et transférer l'information imputée de manière stochastique de la période t_2 à la période t_3 en appliquant les règles L1 et L2.
6. Imputer les données pour les adresses restantes.

Dans l'algorithme, nous utilisons également la nouvelle information pour corriger les données dérivées à la période précédente. Cette correction est nécessaire afin de maintenir la cohérence longitudinale. Supposons, par exemple, que deux personnes vivant à une adresse pour laquelle les données imputées indiquent deux ménages d'une seule personne se marient. À partir du moment où elles se marient, nous savons qu'elles forment un seul ménage. Afin de maintenir la cohérence longitudinale avec les périodes antérieures, il est nécessaire de modifier l'imputation de deux ménages d'une seule personne. Il est important de noter qu'aux étapes 4 et 5, les modèles de régression logistique sont de nouveau estimés avec les données nouvellement requises. Ces modèles sont ensuite utilisés pour les nouvelles imputations.

Un autre avantage de l'utilisation de données provenant de multiples périodes est l'augmentation du nombre de ménages dont la composition est déterminée de manière déterministe. Le tableau 5-1 montre la réduction du nombre d'adresses qui requièrent une imputation quand est utilisée l'information provenant d'autres périodes. Par exemple, le nombre d'adresses pour lesquelles la composition du ménage doit être imputée pour 2006 est réduite de 2,5 % quand nous utilisons l'information provenant de 2002 à 2005, et le nombre d'adresses qui requièrent une imputation en 2002 est réduit de 23,4 % quand nous utilisons l'information provenant de 2003 à 2006. Il s'agit des résultats obtenus quand nous n'exigeons pas que les personnes A et B mentionnées dans les règles L1 et L2 vivent à la même adresse aux deux périodes. Si nous imposons la contrainte que les adresses soient les mêmes, la réduction est plus faible. Les résultats montrent clairement que les nouvelles données contiennent plus d'information au sujet des périodes précédentes que l'inverse.

Tableau 5-1
Réduction du nombre d'adresses nécessitant une imputation quand sont appliquées les règles longitudinales (données couvrant la période de 2002 à 2006)

	Transfert avant [a]		Transfert arrière [a]	
	Non contraint [b]	Contraint [b]	Non contraint [b]	Contraint [b]
2002			23,4%	10,5%
2003	1,1%	0,7%	21,7%	10,0%
2004	1,8%	1,1%	18,1%	8,4%
2005	2,2%	1,3%	12,2%	5,9%
2006	2,5%	1,4%		

[a] Transfert avant/arrière : l'information déterministe provenant d'une période est utilisée aux périodes (ultérieures/antérieures) pour déterminer de manière déterministe la composition des ménages.
[b] Non contraint/contraint : quand l'information sur les personnes est transférée à d'autres périodes, les changements d'adresse sont (permis/non permis).

6. Discussion

Nous avons présenté une nouvelle approche en vue de déterminer des compositions de ménages longitudinalement cohérentes. Cette nouvelle approche est basée sur la méthode actuelle qui possède une qualité chronologique mais non une qualité longitudinale. Nous prévoyons étudier d'autres sources additionnelles que les trois examinées ici afin d'accroître le nombre de ménages dont la composition est établie de manière déterministe. L'accroissement du nombre de ménages dont la composition est dérivée de manière déterministe a pour inconvénient de réduire la quantité de données utilisables pour estimer les modèles servant à l'imputation stochastique, ce qui les rend moins précis. Cette question devra être étudiée plus en profondeur. Elle est liée à la méthode d'imputation stochastique utilisée, qui, à l'heure actuelle, est basée sur l'analyse transversale, mais qui pourrait être fondée sur des méthodes d'analyse longitudinale. Cet aspect doit également être étudié plus en détail.

Un autre question importante est celle de savoir jusqu'où il faut remonter lors du transfert des nouvelles données aux périodes antérieures, ce transfert modifiant les données dérivées antérieurement (déjà publiées). Outre les complications concernant les

publications (il existe alors diverses versions des données pour la même période), ce report en arrière complique l'entreposage des données et le contrôle des versions. Nous n'avons encore pris aucune décision à ce sujet.

Bibliographie

Fellegi, I. P. (1975). Controlled Random Rounding. *Survey Methodology,* 1, pp. 123–133.

Van der Laan, D.J. et Kuijvenhoven, L. (2009). Improving the Derivation of Households. (Vernieuwing Afleiding Huishoudens), Internal Report, CBS, The Hague.

Prins, C.J.M. (2000). Dutch population statistics based on population register data, *Maandstatistiek van de bevolking*, 48(2), CBS, Heerlen/Voorburg.

Wallgren, A. et Wallgren, B. (2007). *Register-based Statistics: Administrative Data for Statistical Purposes*, Chichester: Wiley.

SESSION 20 – APPLICATION : ANALYSE LONGITUDINALE DE DONNÉES SUR LA SANTÉ ET SUR LES ENTREPRISES

Les enfants des femmes plus âgées qui sont mères pour la première fois au Canada : une analyse longitudinale de leur santé et de leur développement

Tracey Bushnik et Rochelle Garner[1]

Résumé

À partir d'un échantillon de 3 382 enfants premiers nés tiré de l'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ), la présente étude examine la relation entre la maternité tardive (c.-à-d. avoir un premier enfant à l'âge de 35 ans ou plus tard) et la santé et le développement des enfants de 0 à 5 ans. Plusieurs questions méthodologiques ont été définies et examinées, y compris la diminution de la taille des échantillons pour les groupes plus âgés; la sélection de poids de sondage appropriés; les répercussions de la sélection de l'échantillon de l'étude et les effets de cohorte. Les résultats montrent que, même si les enfants des mères plus âgées sont plus susceptibles d'être exposés à des facteurs de risque au chapitre du développement, comme une naissance avant terme, sauf pour quelques exceptions, leurs résultats ne diffèrent pas de façon significative de ceux des enfants du groupe de référence.

Mots clés : analyse longitudinale, maternité tardive, santé des enfants, développement des enfants.

1. Introduction et objectifs

Au Canada, il est de plus en plus courant pour les femmes de reporter le moment d'avoir des enfants. Non seulement un plus grand nombre de femmes de plus de 30 ans donnent naissance à un enfant, mais la proportion de premières naissances chez les femmes de plus de 30 ans a augmenté de façon constante au cours des 20 dernières années (Statistique Canada, 2007).

Il est important de comprendre les répercussions possibles de ces tendances sur la santé humaine et le développement. Même si de nombreuses études ont porté sur les résultats de la grossesse et les résultats en matière de santé liés à la maternité tardive, nous en savons moins sur les conséquences possibles pour les enfants.

Afin d'aider à combler cette lacune, la présente étude repose sur des données de l'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ) et examine la relation entre la maternité tardive chez les femmes qui sont mères pour la première fois et trois aspects du développement de l'enfant : (i) santé et développement physique, (ii) comportement, et (iii) développement cognitif. La maternité tardive est définie de la façon suivante : donner naissance à un premier enfant à l'âge ou après l'âge de 35 ans.

Le présent document est fondé sur un rapport plus exhaustif intitulé « Les enfants des femmes plus âgées au Canada qui sont mères pour la première fois : leur santé et leur développement », qui a été publié par Statistique Canada en septembre 2008 (Bushnik et Garner, 2008). Le principal objectif du rapport de 2008 était de présenter les résultats d'une analyse détaillée de la relation entre la maternité tardive et une liste exhaustive des résultats des enfants. Le présent document sera axé principalement sur les défis méthodologiques de l'analyse – en fournissant des détails qui n'étaient pas disponibles dans le rapport de 2008 – et abordera brièvement seulement les résultats analytiques.

2. Méthodes

2.1 Source des données et participants

L'Enquête longitudinale nationale sur les enfants et les jeunes (ELNEJ) est une étude à long terme des enfants canadiens, qui suit leur développement de la naissance au début de l'âge adulte. L'ELNEJ a commencé en 1994 et est menée par Statistique Canada et parrainée par Ressources humaines et Développement des compétences Canada. L'enquête couvre une vaste gamme de sujets, y compris la santé, le développement physique, l'apprentissage, le comportement et l'environnement social. Dans le cas des enfants de moins de 16 ans, la majeure partie des données de l'enquête sont fournies par la personne la mieux

[1]Tracey Bushnik, Division de la statistique de la santé, Statistique Canada, K1A 0T6; Rochelle Garner, Division de la statistique de la santé, Statistique Canada, K1A 0T6.

renseignée au sujet de l'enfant (appelée PMR), c'est-à-dire habituellement la mère. Elle fournit des renseignements sur elle-même, le ménage et la famille, ainsi que l'enfant. Des mesures directes des capacités des enfants peuvent aussi être recueillies, selon l'âge de l'enfant.

L'échantillon complet de l'ELNEJ est constitué des membres de la cohorte longitudinale originale, qui sont suivis tous les deux ans depuis l'âge de 0 à 11 ans en 1994, et plusieurs membres de cohortes de développement de la petite enfance, qui sont interviewés à 0 et 1 an et suivis tous les deux ans jusqu'à 4 et 5 ans. Pour la présente étude, les enfants qui faisaient partie des cohortes de développement de la petite enfance qui ont été recrutés au cycle 3 à 6 inclusivement (1998 à 2005 inclusivement) ont été sélectionnés, ce qui a donné lieu à un échantillon de 18 907 enfants de 0 à 1 an. À partir de cet échantillon regroupé, seuls les enfants premiers nés dont la mère biologique était la personne qui avait été interviewée à chaque cycle ont été retenus pour l'étude. Cela a donné lieu à un échantillon d'étude de 3 382 enfants de 0 à 1 an. Parmi ces enfants, 2 365 ont été interviewés à nouveau à 2 et 3 ans, et 1 705, une troisième fois, à 4 et 5 ans. Le tableau 2.1-1 montre que, même si toutes les cohortes des enfants premiers nés de la présente étude comportent des données pour les caractéristiques mesurées de 0 à 1 an, seules les cohortes des cycles 3 à 5 inclusivement comportaient des données de 2 à 3 ans, et seules les cohortes des cycles 3 et 4 comportaient des données de 4 à 5 ans.

Tableau 2.1-1
Âge de l'enfant à chaque cycle de l'interview de l'ELNEJ

Cohorte d'entrée dans l'échantillon	Âge de l'enfant au cycle de l'interview			
	Cycle 3	Cycle 4	Cycle 5	Cycle 6
Cycle 3	0 à 1	2 à 3	4 à 5	
Cycle 4	...	0 à 1	2 à 3	4 à 5
Cycle 5	...	...	0 à 1	2 à 3
Cycle 6	...	...	...	0 à 1

… sans objet

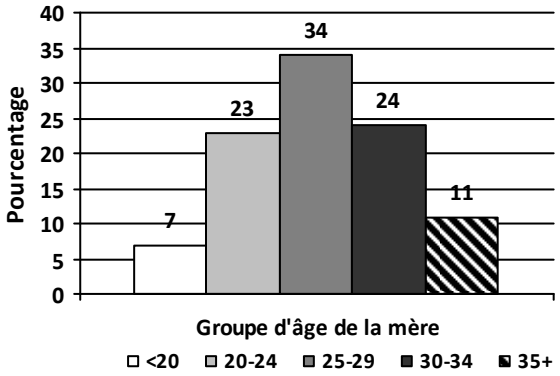
2.2 Mesures

2.2.1 Variables explicatives

Le groupe d'âge de la mère était la variable explicative d'intérêt et était fondé sur l'âge de la femme au moment d'avoir son premier enfant. Les groupes suivants de femmes et leurs enfants ont été définis : (i) *mères adolescentes* (moins de 20 ans au moment de la première naissance); (ii) *jeunes mères* (20 à 24 ans au moment de la première naissance); (iii) *mères du groupe de référence* (25 à 29 ans au moment de la première naissance; (iv) *mères d'âge intermédiaire* (de 30 à 34 ans au moment de la première naissance); et (v) *mères plus âgées* (35 ans et plus à la première naissance). La distribution pondérée des enfants selon le groupe d'âge de la mère est présentée dans la figure 2-1.

Parmi les variables explicatives additionnelles figurent les suivantes : caractéristiques sociodémographiques (situation de faible revenu du ménage, niveau de scolarité de la mère, nombre de parents dans le ménage); mesures périnatales et postnatales (âge gestationnel de l'enfant, poids à la naissance de l'enfant, durée de l'allaitement maternel); et mesures parentales, fonctionnement de la famille et santé mentale de la mère.

Figure 2.2-1
Distribution des enfants de 0 à 1 an selon le groupe d'âge de la mère



2.2.2 Variables de résultat

Les variables de résultat d'intérêt comprenaient les mesures de la santé à divers âges (soins spécialisés à la naissance; état de santé général; prévalence de l'asthme diagnostiqué) et du développement (étape du développement; capacités motrices et compétences sociales), comportement de l'âge de 2 à 5 ans (agressivité physique, troubles émotifs et anxiété, hyperactivité et inattention, comportement positif), et mesures du développement cognitif de 4 à 5 ans.

2.3 Questions méthodologiques

Compte tenu de la stratégie de sélection de l'échantillon de l'étude, plusieurs questions ont dû être abordées, y compris la diminution de la taille des échantillons (moins grand nombre de cohortes) dans les groupes plus âgés; la détermination des poids de sondage appropriés; l'évaluation des répercussions de la sélection de l'échantillon de l'étude; et les effets de cohorte.

2.3.1 Diminution de la taille des échantillons

Le regroupement des quatre cohortes a créé un échantillon suffisamment grand pour permettre d'examiner les caractéristiques des enfants au fil du temps, selon le groupe d'âge de la mère. Toutefois, les données n'ont pas été mesurées à tous les moments pour toutes les cohortes. Même si on a mesuré les données de 0 à 1 an pour toutes les cohortes, la taille de l'échantillon et le nombre de cohortes ont diminué à des âges plus avancés (tableau 2.1-1). Cela a eu deux conséquences principales. Tout d'abord, on a noté des variances accrues des estimations aux âges plus avancés, ce qui a affecté la signification statistique de certaines différences. En deuxième lieu, les inférences concernant les 2 à 3 ans ou les 4 à 5 ans concernaient une population cible plus limitée. De façon plus particulière, les 2 à 3 ans de la présente étude représentaient la population des enfants de 0 à 1 an en janvier des années 1999, 2001 et 2003. Les 4 à 5 ans représentaient la population des enfants de 0 à 1 an en janvier des années 1999 et 2001.

2.3.2 Sélection des poids de sondage

Un seul poids de sondage a été sélectionné par répondant. Tous les poids ont été tirés du dernier cycle auquel l'enfant a répondu (tableau 2.3.2-1). Cela a permis de faire des inférences concernant la population des enfants canadiens représentés par l'échantillon qui ont commencé à participer à l'enquête entre 0 et 1 an. Les poids bootstrap correspondants ont été appliqués pour l'estimation de la variance et les tests d'hypothèse.

Tableau 2.3.2-1
Poids de sondage sélectionnés

Cohorte d'entrée dans l'échantillon	Poids de sondage
Cycle 3	Poids longitudinal tiré du cycle 6
Cycle 4	Poids longitudinal tiré du cycle 5
Cycle 5	Poids longitudinal tiré du cycle 5
Cycle 6	Poids transversal tiré du cycle 6

2.3.3 Évaluation de l'incidence de la sélection de l'échantillon de l'étude

Les critères de sélection de l'échantillon de l'étude ont donné lieu à l'exclusion de certains répondants. Dans les faits, ces enfants exclus sont devenus des non-répondants. Étant donné que la non-réponse peut biaiser les résultats si les non-répondants affichent des caractéristiques significativement différentes des répondants, il était important d'évaluer l'incidence des critères de sélection sur l'analyse.

Au total, 625 enfants premiers nés dont la mère biologique était la PMR au cours de la première interview ont été perdus en raison de l'attrition et, par conséquent, exclus de l'étude. En outre, 396 enfants premiers nés dont la mère biologique était la PMR au moment de la première interview, mais avait changé au cours des interviews subséquentes, ont été exclus de l'étude. Le premier groupe d'enfants a été qualifié de « perdu en raison de l'attrition », tandis que pour le dernier, on a parlé de « PMR ayant changé au fil du temps ».

Les caractéristiques des enfants de ces deux sous-groupes ont été comparées à celles des 3 382 enfants gardés dans l'étude. Ces caractéristiques ont été mesurées de 0 à 1 an et comprenaient les caractéristiques sociodémographiques, prénatales, périnatales et de santé représentant un intérêt pour l'analyse principale.

L'analyse de la non-réponse a montré que les enfants exclus de l'étude parce que la PMR avait changé au fil du temps ne différaient pas de façon significative de l'échantillon final pour la plupart des caractéristiques examinées. Il y avait toutefois plusieurs différences significatives entre les enfants perdus en raison de l'attrition et les enfants de l'échantillon de l'étude. La majeure partie de ces différences rendent compte du fait que les enfants qui abandonnent l'ELNEJ au fil du temps sont plus susceptibles d'avoir un faible statut socioéconomique (SSE) que ceux qui continuent à participer. Il existe deux raisons principales pour lesquelles le biais possible découlant de la perte de ces enfants n'a pas constitué un problème pour l'étude principale. Tout d'abord, les poids longitudinaux fournis avec les données ont été corrigés pour tenir compte de ce type d'attrition. En deuxième lieu, les enfants des mères plus âgées avaient, en moyenne, un SSE relativement élevé et étaient par conséquent moins susceptibles d'abandonner l'enquête. Le biais aurait peut-être posé davantage de problèmes si le présent document avait été axé sur les enfants de mères adolescentes, plutôt que sur ceux des mères plus âgées. Compte tenu de leur faible SSE, les premiers avaient une plus grande probabilité d'être sous-représentés dans l'échantillon sélectionné.

2.3.4 Effets de cohorte

Étant donné le nombre de cohortes regroupées pour l'analyse, il a été nécessaire de vérifier les effets de cohorte. Les caractéristiques et les résultats ont été comparés d'une cohorte à l'autre selon le groupe d'âge de la mère. Les résultats laissent supposer une faible variation systématique selon la cohorte. La variation limitée qui a été observée a fait l'objet d'une analyse plus poussée, afin de déterminer si les différences dans les distributions sous-jacentes d'âges en mois entre les cohortes étaient responsables des différences mineures. Cela ne semble pas avoir été le cas. Néanmoins, une variable fictive de cohorte a été incluse dans l'ensemble de l'analyse multidimensionnelle pour contrôler les effets de cohorte possibles.

3. Analyse et résultats

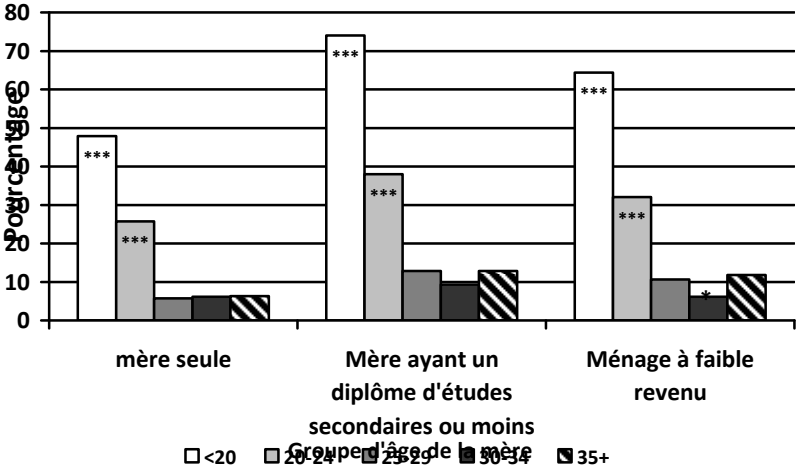
Pour chaque mesure de résultat, des moyennes ou des proportions ont été produites globalement et selon le groupe d'âge de la mère. Dans toutes les analyses, les enfants dont la mère était âgée de 25 à 29 ans au moment de leur naissance ont constitué la catégorie de référence (appelée *mères du groupe de référence*). Par la suite, des modèles de régression multidimensionnels ont été produits pour chaque résultat. Dans chaque modèle, on a inclus l'âge de la mère au moment de la naissance de l'enfant et un groupe de caractéristiques sociodémographiques. En outre, selon le résultat examiné, certains comportements ou caractéristiques de la mère liés à ce résultat particulier dans les ouvrages publiés ont été inclus dans le modèle. Comme il est souligné précédemment, une variable de contrôle de la cohorte a aussi été incluse dans chaque modèle.

Les modèles logistiques ont été ajustés en fonction de variables dichotomiques, tandis que les modèles linéaires ont été ajustés en fonction de variables continues. Toutes les analyses ont été pondérées (voir la section 2.3.2), en vue d'être généralisées à la population canadienne. Afin de tenir compte du plan de sondage complexe de l'ELNEJ, des poids bootstrap ont servi à produire des estimations de la variance. Toutes les analyses ont été effectuées au moyen de SUDAAN en SAS.

3.1 Caractéristiques sociodémographiques

Les enfants des mères plus âgées avaient un profil sociodémographique similaire à celui des enfants des mères du groupe de référence (figure 3.1-1). Les enfants nés de mères adolescentes et de jeunes mères, toutefois, étaient beaucoup plus susceptibles d'appartenir à un ménage monoparental à faible revenu ayant à sa tête une femme, et d'avoir une mère ayant fait des études secondaires seulement, comparativement aux enfants des mères du groupe de référence.

Figure 3.1-1
Caractéristiques sociodémographiques de 0 à 1 an, selon le groupe d'âge de la mère

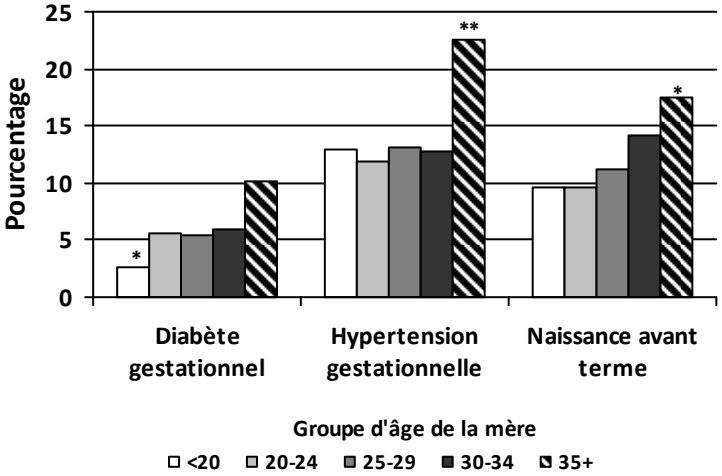


* significativement différent du groupe de référence à p<0,05
***significativement différent du groupe de référence à p<0,001

3.2 Facteurs de risque prénatal et caractéristiques liées à la naissance

Les enfants des mères plus âgées étaient plus susceptibles d'avoir une mère souffrant d'hypertension gestationnelle et d'être nés avant terme (figure 3.2-1) que les enfants des mères du groupe de référence. Les naissances par césarienne étaient plus répandues chez les enfants des mères plus âgées : environ 40 % des enfants des mères plus âgées sont nés par césarienne, comparativement à 23 % des enfants des mères du groupe de référence.

Figure 3.2-1
Facteurs de risque prénatal et naissance avant terme selon le groupe d'âge de la mère



* significativement différent du groupe de référence à p<0,05
**significativement différent du groupe de référence à p<0,01

3.3 Maternité tardive et résultats des enfants

Même s'ils ne sont pas présentés ici, les résultats de l'analyse ont montré que les enfants des mères plus âgées ne différaient pas des enfants des mères du groupe de référence pour de nombreux résultats en matière de développement. Par exemple, les enfants des mères plus âgées étaient aussi susceptibles que les enfants des mères du groupe de référence d'avoir reçu des soins spéciaux à la naissance, d'avoir une santé excellente ou très bonne pendant la petite enfance ou de recevoir un diagnostic d'asthme. Ils obtenaient des résultats semblables en ce qui a trait au moment où ils ont dit leurs premiers mots et fait leurs premiers pas, ainsi que des scores moyens similaires au chapitre de l'agressivité physique, des troubles émotifs et de l'anxiété,

ainsi que de l'hyperactivité et de l'inattention. Les enfants des mères plus âgés obtenaient aussi des scores similaires pour le développement cognitif que les enfants des mères du groupe de référence.

Toutefois, l'âge avancé de la mère comportait un lien significatif avec d'autres résultats. Même après avoir contrôlé un certain nombre de caractéristiques, une proportion plus élevée d'enfants des mères plus âgées étaient considérés comme ayant des retards pour ce qui est de s'asseoir seuls par rapport aux enfants des mères du groupe de référence. En outre, les enfants des mères plus âgées obtenaient des scores plus faibles sur l'échelle du développement moteur et social entre 0 et 1 an et entre 2 et 3 ans. Ils obtenaient aussi des scores plus faibles au chapitre du comportement positif entre 4 et 5 ans.

4. Discussion

À partir de l'ELNEJ, la présente étude a permis d'évaluer la relation entre la maternité tardive et les résultats des enfants au fil du temps. Selon les résultats, même si les enfants des mères plus âgés sont plus susceptibles d'être exposés à des facteurs de risque au chapitre du développement, comme la naissance avant terme, sauf pour quelques exceptions, les résultats ne diffèrent pas de façon significative de ceux des enfants du groupe de référence.

L'analyse a comporté son lot de défis méthodologiques. L'attrition et son effet sur la taille des échantillons et l'inférence, l'utilisation appropriée des poids de sondage et la possibilité d'effets de cohorte ont dû être reconnues et prises en compte.

L'étude comportait certaines limites, y compris : un biais possible en faveur des enfants en santé, parce que l'ELNEJ est une enquête auprès des ménages dont l'échantillon ne comprend pas la population vivant en établissement; l'utilisation des données déclarées par la mère pour certains résultats; et la période relativement courte de développement des enfants (0 à 5 ans) qui a été examinée.

En dépit de ces limites, la présente étude donne un nouvel aperçu du rapport entre la maternité tardive et le développement de l'enfant dans le contexte canadien.

Bibliographie

Bushnik, T. et Garner, R. (2008). *Les enfants des femmes plus âgées qui sont mères pour la première fois au Canada : leur santé et leur développement*, Ottawa: Statistique Canada. Catalogue 89-599-M, no. 005.

Statistique Canada (2007). *Naissances: Naissances vivantes 2005*. Statistique Canada Catalogue no. 84F0210XIF. Ottawa.

L'évolution de l'IMC et de la taille au cours du cycle de vie : comparaison de deux cohortes de naissances britanniques

Leah Li, Rebecca Hardy, Diana Kuh et Chris Power[1]

Résumé

L'obésité ne cesse de croître dans le monde entier. Les trajectoires d'IMC ont peut-être évolué chez les générations récentes touchées par l'épidémie d'obésité à différentes étapes de la vie. D'autres composantes du développement physique ont peut-être évolué également. Nous comparons les trajectoires de croissance de l'enfance à l'âge adulte dans deux cohortes de naissance britanniques, soit de personnes nées en 1946 (n=5 300) et en 1958 (n=17 000), qui font l'objet d'un suivi jusqu'à l'âge de 53 ans et de 45 ans, respectivement.

Les personnes nées en 1958 n'avaient pas un poids plus élevé à la naissance que les membres de la cohorte de 1946, mais elles avaient un centimètre de plus de taille au début de l'enfance, elles ont grandi plus rapidement et, adolescentes, elles avaient entre trois et quatre centimètres de plus de taille. Les membres de la cohorte de 1958 ont atteint leur taille adulte plus tôt et ils mesuraient un centimètre de plus, augmentation entièrement attribuable à leurs jambes plus longues. Nous avons appliqué des modèles splines linéaires à des mesures de l'IMC répétées (à 7, 11, 15, 20, 26, 36, 43 et 53 ans pour la cohorte de 1946 et à 7, 11, 16, 23, 33 et 45 ans pour la cohorte de 1958) correspondant à des trajectoires d'IMC distinctes pour « l'enfance » et « l'âge adulte ». Les trajectoires d'IMC divergent à partir du début de l'âge adulte, la cohorte de 1958 affichant un taux de croissance plus rapide que la cohorte de 1946, même si l'IMC moyen à sept ans et le taux d'accroissement durant l'enfance n'a pas augmenté dans une cohorte par rapport à l'autre. Au milieu de l'âge adulte, la cohorte de 1958 a un IMC plus élevé (1-2 kg/m2), un plus grand tour de taille (7-8 cm), un plus grand tour de hanches (5 cm) et une prévalence plus élevée de l'obésité (24 % par rapport à 12 %). Ces changements survenus au cours d'une période relativement courte de 12 ans donnent à penser qu'il est probable que des tendances opposées influent sur les risques de maladie plus tard dans la vie dans ces populations.

[1] Leah Li (l.li@ich.ucl.ac.uk), Rebecca Hardy, Diana Kuh et Chris Power, University College London, Royaume-Uni.

L'impact de la formation sur la productivité des entreprises canadiennes dans un contexte longitudinal : la comparaison d'un modèle additif et d'un modèle interactif

Amélie Bernier et Jean-Michel Cousineau[1]

Résumé

À partir des données des employeurs de l'Enquête sur les milieux de travail et les employés de 1999 à 2005, nous estimons les effets des dépenses en formation sur la productivité des entreprises. Nos résultats montrent que les investissements en formation ont des effets positifs qui s'étalent sur une période de 3 ans[2]. L'interaction entre les investissements en capital physique et en formation permet de vérifier l'hypothèse à l'effet que les investissements en capital physique et en capital humain soient complémentaires et se supportent mutuellement.

Mots Clés : étude longitudinale, modèle récursif, formation, productivité, entreprises.

1. Introduction

Les milieux de travail canadiens présentent des visages diversifiés et en pleine mutation. En raison de facteurs tels que les nouvelles technologies, la mondialisation, l'économie du savoir ou encore l'évolution démographique, la transformation des entreprises canadiennes passe par une main-d'œuvre qualifiée, adaptable et de qualité. Si l'on observe l'offre de formation structurée des entreprises canadiennes pour la période de 1999 à 2005 au sein de notre échantillon, on peut voir qu'en moyenne 32,5 % des entreprises offrent ce type de formation. Plusieurs entreprises semblent se questionner sur la nécessité et l'intérêt de participer à la formation continue (Goldenberg, 2006). Serait -ce parce qu'elles doutent de l'influence de leurs investissements en formation sur la productivité à court terme, et de leur pertinence à moyen terme?

La présente étude s'appuie sur les données de l'Enquête sur les milieux de travail et des employés (EMTE) en vue d'estimer les effets des dépenses en formation sur la productivité à court et à moyen terme. Notre étude est, aux meilleurs de nos connaissances, la seule qui estime les retours sur les investissements en formation structurée auprès de données d'entreprises canadiennes en utilisant un modèle récursif. Nos contributions sont les suivantes. D'abord, afin de s'inscrire dans le courant de recherche actuel sur les retours des investissements en formation, il est essentiel d'utiliser le caractère longitudinal de données, afin d'en mesurer les effets possibles au-delà de l'année courante des investissements réalisés par les entreprises. Ensuite, pour mesurer adéquatement les retours sur les investissements en formation au sein des entreprises, il faut au préalable se pencher, sur des problèmes techniques liés aux caractéristiques idiosyncrasiques des firmes et à la simultanéité de la formation avec la variable dépendante retenue. Notre modèle à effets retardés permet de capturer l'essentiel du problème d'endogénéité. Enfin, il demeure pertinent de valider empiriquement, l'hypothèse de complémentarité des investissements en formation avec d'autres types d'investissements au sein des firmes, puisque les résultats des études longitudinales consultées ne sont pas toujours concluants (Zwick, 2006).

2. La problématique

D'abord, lorsqu'une entreprise investit dans la formation, elle espère obtenir des gains monétaires ou en termes d'amélioration de la qualité de sa main-d'œuvre. En se référant à la théorie du capital humain et au modèle de partage des coûts de Becker (1964), une formation supplémentaire pour les travailleurs constitue un moyen d'accroître potentiellement la productivité de l'entreprise, par la valeur de productivité marginale après la formation ainsi que les gains monétaires des employés eux-mêmes. Bien que la théorie du capital humain ait largement démontré son application empirique dans l'explication des rendements de la scolarité post-secondaire, nous sommes portés à croire, qu'elle peut apporter un éclairage simple et convainquant dans

[1] Amélie Bernier, Candidate au doctorat (Ph.D.), École de relations industrielles, Université de Montréal, Canada, courriel : amelie.bernier@umontreal.ca ; Jean-Michel Cousineau, professeur titulaire, École de relations industrielles, Université de Montréal, Canada, courriel : jean.michel.cousineau@umontreal.ca

[2] Bien que la recherche et les analyses soient fondées sur des données de Statistique Canada, les opinions exprimées ne représentent que celles des auteurs.

l'explication des effets différés de la formation, notamment par sa traduction au niveau de la méthode, d'un modèle récursif appliqué à des données au niveau des entreprises.

Par ailleurs, cette approche ne propose pas de solutions clairement identifiées pour traiter empiriquement du problème potentiel du caractère endogène[3] de la formation. Ce qui a pour effet d'engendrer un vaste débat au niveau empirique. En effet, la recension des écrits (Colombo et Stanca, 2008; Zwick, 2006) nous a permis de constater que le caractère possiblement endogène de la formation est fréquemment invoqué pour expliquer les différences entre les résultats d'estimation.

Cet exposé ouvre la voie également à l'approfondissement des liens possibles entre différents types d'investissements permettant de vérifier l'hypothèse à l'effet que certains facteurs de production soient complémentaires et se supportent mutuellement pour l'amélioration de la productivité.

Empiriquement, les enjeux sont multiples. Parmi l'ensemble des travaux consultés, il ne semble pas y avoir de relation clairement définie entre les coûts et les bénéfices de la formation et il ne semble pas exister de consensus quant aux retours sur les investissements (Ballot, 2006). En effet, les liens entre la formation et la productivité ne sont pas si évidents comme nous le montrerons. Bien qu'un nombre croissant de travaux ont tenté de mesurer l'impact de la formation sur certains indicateurs de performance de l'organisation à partir de données représentatives au niveau de l'entreprise (Colombo et Stanca, 2008; Dostie et Pelletier, 2007), les résultats ne sont pas toujours concluants. L'ensemble des études recensées ne permet pas d'obtenir l'unanimité au niveau : 1) de la mesure elle-même de la variable « formation », 2) du type de formation à retenir pour fins d'analyse, 3) des résultats obtenus et 4) des techniques d'estimation utilisées. Ce qui manque aussi, à notre avis, c'est un moyen de faire le lien entre les investissements passés en formation et la productivité actuelle des entreprises. Nous devons faire ce lien si nous voulons comprendre les retombées des investissements en formation pour les entreprises canadiennes.

3. Données et méthode

Pour estimer nos modèles d'impact de la formation sur la productivité des entreprises canadiennes, nous travaillons avec un échantillon composé de 1621 firmes et de 11347 observations. Les données utilisées sont issues du questionnaire des employeurs de l'EMTE pour la période de 1999 à 2005. En résumé, notre échantillon est composé de l'ensemble des firmes à but lucratif ayant au moins un employé, qui rapportaient des rendements financiers positifs et pour lesquelles, nous avons de l'information quant à leurs décisions de formation. Nous présenterons dans la prochaine section les résultats obtenus.

Pour estimer les rendements de la formation au niveau organisationnel, nous favorisons une fonction Cobb-Douglas[4] présentée comme un processus de production (Q_{it}) permettant de relier la valeur ajoutée[5] aux dépenses de formation au sein des organisations. L'utilisation d'une fonction Cobb-Douglas avec la production (Q_{it}) comme variable dépendante permet d'étudier l'impact de la formation sur la productivité puisque l'emploi (ou le facteur travail) est une variable indépendante avec des rendements constants, ce qui impose, selon nous, une solution plus flexible et moins contraignante dans l'estimation de l'effet des dépenses en formation sur la productivité.

À l'instar des travaux de Almeida et Carneiro (2006) et de Barrett et O'Connell (2001), notre modèle de base, caractérisé par une firme *i* à l'année *t* se présente de la façon suivante :

$$\ln Q_{it} = \ln A_{it} + \alpha \ln L_{it} + \beta \ln I_{it} + \gamma \ln F_{it} + \delta T_{it} + \eta X_{it} + \varepsilon_{it} \qquad (1)$$

Où i = 1,…, N; t = 1,…, T; α; β; γ; δ; η sont des paramètres à estimer.

Q_{it} représente la valeur ajoutée[6] qui est fonction de trois principaux facteurs (L_{it}; I_{it}; H_{it}) ainsi que d'un paramètre d'échelle (A_{it}). D'abord, la main-d'œuvre (L_{it}) est mesurée par l'effectif total au sein de la firme. Ensuite, nous utilisons les

[3] Lorsque l'on parle d'endogénéité, on réfère aux variables dites indépendantes qui sont corrélées avec le terme d'erreur.
[4] La caractéristique principale de la fonction Cobb-Douglas réside dans le fait que l'élasticité de substitution est égale à 1 et demeure constante le long de l'isoquant. Une variation du taux marginal de substitution entraîne une variation proportionnelle du rapport des quantités de facteurs. Si l'on réfère à la présente analyse, cette élasticité de substitution égale à 1 suppose que les facteurs sont autant des compléments que des substituts.
[5] La valeur ajoutée est déterminée par la valeur brute de la production de laquelle nous avons soustrait le coût des intrants (Dostie et Pelletier, 2007).
[6] Pour tenir compte de l'effet de l'inflation sur les variables nominales telles que les recettes d'exploitation brutes, les dépenses en équipement et les dépenses en formation structurée, nous avons déflaté les valeurs par l'indice annuel des prix à la consommation pour le Canada (où 1992=100), tel que reporté par Statistique Canada (Tableau CANSIM no 326-0002).

investissements en capital physique (I_{it}) mesurés par le total des dépenses en équipements[7] au sein des organisations (Barett et O'Connell, 2001; Colombo et Stanca, 2008). Les investissements en capital humain (F_{it}) sont représentés par un ratio des dépenses en formation structurée sur l'effectif total au sein de la firme[8]. Le questionnaire des employeurs de l'EMTE permet de recueillir des données sur deux grandes catégories de formation : la formation structurée et la formation en cours d'emploi. Ainsi, que la formation soit à caractère générale ou spécifique, l'information disponible dans l'EMTE suggère plutôt une différentiation quant au caractère prédéterminé du contenu et au niveau des objectifs qui peuvent être évalués plutôt que la division prescrite théoriquement.

Bien qu'imparfaite, une variable de technologie (T_{it}) représentant la proportion d'employés utilisant un ordinateur dans leur travail quotidien a été ajoutée au modèle. Finalement, l'équation estimée à partir de la spécification (1) inclut aussi un ensemble de variables de contrôle représentées par un vecteur (X_{it}).

Pour avoir une vue d'ensemble des effets possibles de la formation sur la performance des entreprises, nous devons d'abord estimer l'équation (1) par la technique des moindres carrés ordinaires (MCO). Par ailleurs, sans remettre en cause complètement l'idée que les différences de résultats, dans les études longitudinales, proviennent du fait que la formation soit endogène à la variable dépendante, nous proposons de vérifier le caractère prédéterminé ou endogène de la formation dans le cadre de nos estimation. Dans ce cas, parmi les options envisagées pour traiter du non respect des hypothèses sous adjacentes aux MCO, on trouve l'usage de variables instrumentales (IV), des effets aléatoires (RE) ou fixes (FE), la correction pour l'auto-corrélation de premier ordre (AR1), le développement de modèles dynamiques utilisant la méthode des moments généralisés et le recours à l'analyse de données longitudinales.

Comme nous l'avons déjà signalé, le retard d'une année (ou plus) est nécessaire puisqu'un problème d'endogénéité se poserait si les décisions liées aux investissements en formation et à la productivité des organisations étaient déterminées simultanément. Ainsi, puisque nous supposons que la décision d'investir en formation et son retour sur l'investissement en termes de croissance de la productivité peut comporter un retard considérable, nous proposons donc un modèle dans lequel la longueur du retard est spécifiée (t = 4). La variable d'investissements en capital physique (I_{it}) sera également retardée dans le temps. Pour appliquer ces délais, nous devons réécrire l'équation (1) comme suit :

$$\ln Q_{it} = \ln A + \alpha \ln L_{it} + \sum_{j=1}^{4} \beta_j \ln I_{it-j} + \sum_{j=1}^{4} \gamma_j \ln F_{it-j} + \delta T_{it} + \eta X_{it} + \varepsilon_{it} \quad (2)$$

Où j varie de 1 à 4 années; $b_1 = \sum_{j=1}^{4} \beta_j$; $b_2 = \sum_{j=1}^{4} \gamma_j$

Par ailleurs, afin de valider l'hypothèse de complémentarité des investissements en formation (F_{it}) et des investissements en capital physique (I_{it}) sur la productivité, nous incluons un terme multiplicatif représenté par un symbole (*) entre F_{it} et I_{it} que nous pouvons voir à l'équation (3):

$$\ln Q_{it} = \ln A_{it} + \alpha \ln L_{it} + \sum_{j=1}^{4} \beta_j \ln I_{it-j} + \sum_{j=1}^{4} \phi_j \ln I_{it-j} * \ln F_{it-j} + \delta T_{it} + \eta X_{it} + \varepsilon_{it} \quad (3)$$

Où j varie de 1 à 4 années; $b_1 = \sum_{j=1}^{4} \beta_j$; $b_3 = \sum_{j=1}^{4} \phi_j$

La section suivante présente les résultats obtenus pour l'ensemble de nos estimations.

[7] La variable dépenses en équipement a été construite à partir des dépenses en exploitation brutes desquelles nous avons soustrait la masse salariale brute, les dépenses en avantages sociaux et les dépenses en formation structurée.

[8] Bien que les dépenses en formation soient peu utilisées dans la littérature empirique antérieure, elles constituent un indicateur fiable de l'intensité de la formation offerte au sein des emplacements. La fiabilité de nos données s'appuie notamment sur la méthodologie éprouvée de Statistique Canada, ce qui permet d'obtenir des taux de réponses élevés dans le cadre de l'enquête et d'assurer la qualité des données. À titre d'exemple, le taux de réponse des employeurs en 2005 s'élevait à 77,7 %. Une autre façon de mesurer la fiabilité des données est de calculer le coefficient de variation. Si la valeur obtenue est inférieure à 16,5 %[8], l'estimation de la variable F_{it} sera considérée fiable. Dans notre cas, le coefficient de variation est de 3,57 % ce qui représente une homogénéité dans les données utilisées.

4. Les résultats

4.1 Le caractère endogène de la formation

La réalisation du test d'endogénéité de Nakamura-Nakamura (Nakamura et Nakamura, 1998) montre que les résultats sont à la marge du seuil d'acceptation de l'hypothèse nulle (absence d'endogénéité). Un second estimateur qui s'offre à nous pour tester la présence d'endogénéité est l'utilisation du test de Hausman. Les résultats obtenus suggèrent encore une fois de rejeter l'hypothèse nulle d'exogénéité des variables explicatives, ce qui laisse la possibilité que la variable de formation soit endogène à la productivité[9]. Pour corriger ce biais potentiel, la formation devrait être estimée sur une longue période de temps, pour en documenter l'ensemble des retours possibles, d'où l'intérêt d'utiliser un modèle récursif pour mesurer les retours sur les investissements en formation avec plus d'une année de délais. Par exemple, le modèle 2 (tableau 4.1-1) montre que la partie exogène de la variable de formation exerce un effet positif et significatif sur la productivité. Ainsi, on peut voir qu'une augmentation de 10 % dans les dépenses de formation structurée par employé engendre une augmentation de 1,7 % de la productivité de l'entreprise pour l'année suivante.

Tableau 4.1-1
Résultats des estimations de l'impact des dépenses en formation sur la productivité des entreprises au Canada de 1999 à 2005

Variable dépendante : ln Production (valeur ajoutée)	**Modèle 1**	**Modèle 2**	**Modèle 3**	**Modèle4**	**Modèle 5**
	OLS	**IV**	**RE**	**FE**	**AR1, RE**
Variables explicatives					
Ln (nombre total d'employés)	0,8306***	1,0081***	0,8069***	0,5611***	0,8096***
	(0,0253)	(0,0183)	(0,0295)	(0,0592)	(0,0289)
Proportion d'employés utilisant l'ordinateur (en pourcentage)	0,0040***	0,0025***	0,0025***	0,0012*	0,0026***
	(0,0007)	(0,0004)	(0,0006)	(0,0006)	(0,0006)
ln (investissement en capital physique)		0,0346***			
		(0,0117)			
ln (dépenses de formation structurée/nombre total d'employés) $_{t-1}$	-0,0119	**0,1656*** **	-0,0006	-0,0081	0,0002
	(0,0253)	**(0,0203)**	(0,0136)	(0,0157)	(0,0138)
ln (dépenses de formation structurée/nombre total d'employés) $_{t-2}$	0,0047		0,0175	0,0029	0,0178
	(0,0283)		(0,0131)	(0,0151)	(0,0132)
ln (dépenses de formation structurée/nombre total d'employés) $_{t-3}$	**0,0340**		**0,0374*** **	**0,0206**	**0,0380*** **
	(0,0270)		**(0,0124)**	**(0,0141)**	**(0,0126)**
ln (dépenses de formation structurée/nombre total d'employés) $_{t-4}$	0,0341		0,0094	-0,0086	0,0107
	(0,0238)		(0,0124)	(0,0137)	(0,0126)
ln (investissement en capital physique) $_{t-1}$	0,0177		0,0659***	0,0282	0,0683***
	(0,0304)		(0,0165)	(0,0205)	(0,0167)
ln (investissement en capital physique) $_{t-2}$	0,0538*		0,0510***	0,0051	0,0508***
	(0,0330)		(0,0159)	(0,0190)	(0,0161)
ln (investissement en capital physique) $_{t-3}$	0,0397		0,0338***	-0,0049	0,0338***
	(0,0286)		(0,0128)	(0,0146)	(0,0130)
ln (investissement en capital physique) $_{t-4}$	0,0542**		0,0260**	-0,0007	0,0279***
	(0,0218)		(0,0109)	(0,0116)	(0,0111)
Nombre total d'observations (N)	1555	4477	1555	1555	1555
Nombre total de firmes (n)		1080	620	620	620
R carré du modèle	0,7814	0,7649	0,7728	0,6839	0,7738

Notes :
1. Écarts types robustes entre parenthèses. *** significatif à 1 % ** significatif à 5 % * significatif à 10 %
2. Une constante, des variables dichotomiques pour l'industrie (14), le temps, ainsi des variables de contrôle sont également incluses dans les modèles. Les résultats complets sont disponibles sur demande.

4.2 Les effets différés de la formation sur la productivité

Dans le cadre des modèles à effets différés la structure de retards est le temps t-1 à t-4. Le temps « t » pour l'année courante n'est pas retenu de façon à éviter les problèmes potentiels de simultanéité. Lorsque l'on s'intéresse aux effets différés de la formation, les résultats suggèrent aussi que les délais plus courts montrent un faible impact, confirmant l'hypothèse de la théorie du capital humain, pour ensuite connaître une croissance dans le temps. Dans les modèles 1, 3, 4 et 5 (tableau 4.1-1), le

[9] Les résultats de ces estimations sont disponibles sur demande.

coefficient estimé à t-3 suggère qu'il y une période d'ajustement entre la dépense de formation et la rentabilité de l'investissement. Cette explication associée aux retours sur les investissements en formation demeure pour l'ensemble des modèles récursifs estimés. Aussi, les résultats présentés dans les modèles 3 à 5 montrent que les résultats pour la variable de formation présentent une structure de rentabilité qui diffère de celles des investissements en capital physique. La structure des retards (t-1 à t-4) pour expliquer l'impact de la formation sur la productivité des entreprises canadiennes a une forme en U inversé : elle connaît d'abord une période de croissance, atteint un sommet et décline par la suite tandis qu'elle est strictement décroissante pour l'investissement en capital physique. L'ajout des variables de contrôle ne change en rien ce constat. La conception usuelle qui s'applique aux investissements physiques ne s'applique pas forcément aux investissements en capital humain.

4.3 La complémentarité des facteurs de production

Les résultats des estimations tenant compte de l'interaction entre les investissements en capital et ceux réalisés en formation sont rapportés au tableau 4.3-1. La variable d'interaction $\ln F_{it}$* $\ln I_{it}$ a un coefficient positif et significatif pour les modèles 6 et 7, indiquant des complémentarités entre F_{it} et I_{it} pour la productivité des firmes. À titre d'exemple, l'interprétation de l'effet global des I_{it} et de l'interaction entre F_{it} et I_{it} dans le modèle 6 signifie qu'une hausse de 10 % dans les investissements en capital physique bonifié des dépenses en formation structurée par employé, entraînera en moyenne une augmentation de 0,6 % de la productivité de l'entreprise l'année suivante, comparativement à une entreprise qui n'a pas intégré de pratiques complémentaires à la formation, tel que l'investissements en capital physique.

Tableau 4.3-1
Résultats des estimations de l'effet joint des dépenses en formation et des investissements en capital physique sur la productivité des entreprises au Canada de 1999 à 2005

Variable dépendante : ln Production (valeur ajoutée)	**Modèle 6**	**Modèle 7**
	AR1, RE	**AR1, RE**
Variables explicatives	**(t-1)**	**(t-4)**
Ln (nombre total d'employés)	0,9080***	0,8091***
	(0,0167)	(0,0289)
Proportion d'employés utilisant l'ordinateur	0,0026***	0,0026***
(en pourcentage)	(0,0004)	(0,0006)
ln (investissement en capital physique) $_{t-1}$	0,0598***	0,0691***
	(0,0090)	(0,0173)
ln (investissement en capital physique) $_{t-2}$		0,0441***
		(0,0167)
ln (investissement en capital physique) $_{t-3}$		0,0205
		(0,0136)
ln (investissement en capital physique) $_{t-4}$		0,0244**
		(0,0120)
ln Formation * ln Investissements en capital physique $_{t-1}$	**0,0013****	-0,0000
	(0,0006)	(0,0008)
ln Formation * ln Investissements en capital physique $_{t-2}$		0,0012
		(0,0008)
ln Formation * ln Investissements en capital physique $_{t-3}$		**0,0022*****
		(0,0008)
ln Formation * ln Investissements en capital physique $_{t-4}$		0,0006
		(0,0008)
Nombre total d'observations (N)	4487	1555
Nombre total de firmes (n)	1088	620
R carré du modèle	0,7738	0,7736

Notes :
1. Écarts types robustes entre parenthèses. *** significatif à 1 % ** significatif à 5 % * significatif à 10 %
2. Une constante, des variables dichotomiques pour l'industrie (14), le temps, ainsi des variables de contrôle sont également incluses dans les modèles. Les résultats complets sont disponibles sur demande.

Tout comme certaines études empiriques précédentes (Zwick, 2006) nous ne pouvons conclure, dans le cadre de nos estimations, à la supériorité de l'interaction entre les investissements en capital physique et les dépenses en formation structurée par employé. La comparaison des résultats obtenus (tels que la valeur des R carré) dans le modèles additif 5 avec les effets combinés de F_{it} et I_{it}, dans le modèles 7 permettent de tirer ces conclusions. Par contre, ces informations sont complémentaires. Ainsi, dans le modèle 5 par exemple, si nous additionnons les effets retardés pour les investissements en capital physique avec ceux de la variable $F_{it,}$ nous obtenons un effet global pour les quatre années de retards de l'ordre de 0,2475. Ce qui signifie qu'une augmentation de 10 % des investissements totaux pour une année, entraînera une hausse totale

d'environ 2,5 % de la productivité de l'entreprise sur une période de quatre ans. Par ailleurs, considérant que l'effet combiné global de F_{it} et I_{it}, dans le modèle 7 est de 0,1621, nous pouvons nous attendre à un impact moindre de l'effet d'interaction entre F_{it} et I_{it} sur la productivité des entreprises. En d'autres mots, une hausse de 10 % du produit des investissements en capital physique et dans les dépenses en formation structurée par employé, pour une année, entraînera au total sur une période de quatre ans, une croissance de la productivité d'environ 1,6 %.

5. Conclusion

Nos résultats justifient pertinemment l'usage des données longitudinales dans l'étude des retours sur les investissements en formation. Ainsi, il existe plusieurs avantages à exploiter une banque de données longitudinales, comme c'est le cas avec l'Enquête sur les milieux de travail et les employés de Statistique Canada. Le fait d'utiliser un modèle avec des données longitudinales permet d'être plus aptes à réaliser une recherche dynamique. Ainsi, l'étude de la formation au sein des entreprises peut être mieux documentée notamment en termes d'effet causal mais permet aussi de déceler des effets qui ne sont normalement pas détectables avec l'utilisation de données transversales dont des effets ambivalents à court terme, mais largement significatifs à moyen terme.

Malgré le support empirique des retours sur les investissements en formation à moyen terme pour les firmes canadiennes, nous pouvons nous demander pourquoi elles ne forment pas davantage si les retours sur les investissements en formation sont substantiels ? L'une des raisons qui peut expliquer cette réticence est la crainte de voir les employés quitter vers de meilleures opportunités à l'extérieur de l'entreprise. Dans le cadre de travaux complémentaires, nous tenterons d'apporter un éclairage additionnel sur la question de l'impact des investissements en formation structurée par employé sur différentes facettes du taux de roulement au sein des firmes canadiennes.

Références

Almeida, R. et Carneiro, P. (2006). The return to the Firm Investment in Human Capital, *Discussion Paper Series*, no 1937, Institute for the Study of Labor (IZA), 24 pages.

Ballot, G., Fakhfakh, F. et Taymaz, E. (2006). How Benefits from Training and R&D, the Firm or the Workers?, *British Journal of Industrial Relations,* vol. 44, no 3, p. 473-495.

Barney, J. (1991). Firm Resources and Sustained Competitive Advantage, *Journal of Management,* vol. 17, no 1, p. 99-120.

Bassi, L., Harrison, P., Ludwig, J. et McMurrer, D. (2001). Human Capital Investments and Firm Performance, Working Paper, Human Capital Dynamics, Washington.

Becker, G. (1964). Human capital : A theoretical and empirical analysis, with special reference to education, New York, National Bureau of Economic Research.

Colombo, E. et Stanca, L. (2008). The impact of Training on Productivity : Evidence from a Large Panel of Firms, Working Papers 134, University of Milano-Bicocca, Department of Economics.

Dostie, B. et Pelletier, M.-P. (2007). Les rendements de la formation en entreprise, *Canadian Public Policy/Analyse des Politiques.* Vol. XXXIII, no 1, 21-40.

Goldenberg, M. (2006). *Investissements des employeurs dans l'apprentissage en milieu de travail au Canada* , Document de recherche préparé par les Réseaux canadiens de recherche en politiques publiques au nom du Conseil canadien sur l'apprentissage, 67 pages.

Nakamura, A. et Nakamura, M. (1998). Model specification and endogeneity, *Journal of Econometrics*, vol. 83, 213-237.

Zwick, T. (2006). The impact of training intensity on establishment productivity, *Industrial Relations*, vol. 45, no. 1, 26-46.

Recueil du Symposium 2009 de Statistique Canada
Les enquêtes longitudinales : de la conception à l'analyse

Mobilité des travailleurs : sommaire et nouveaux résultats de l'Enquête sur le milieu de travail et les employés

Yves J. Decady[1]

Résumé

Dans le présent article, les données provenant du volet Employés de l'Enquête sur le milieu de travail et les employés (EMTE) sont d'abord utilisées pour décrire la prévalence de la mobilité de l'emploi. Ensuite, l'estimation par variables instrumentales est combinée au calcul des différences premières pour produire des estimations du rendement de la mobilité de l'emploi. Cette approche aide à éliminer les effets non observés des caractéristiques du travailleur, d'une part, et de tenir compte de l'endogénéité de la variable de mobilité de l'emploi, d'autre part. L'effet de sélection de l'échantillon sur les estimations est également examiné. Les résultats de l'étude indiquent que le rendement de la mobilité de l'emploi est positif. Ils étayent aussi la thèse selon laquelle l'effet de la mobilité des travailleurs sur la rémunération dépend de la position du travailleur sur la courbe de distribution des salaires.

Mots clés : différence première, effets fixes, variable instrumentale, sélection de l'échantillon, mobilité de l'emploi.

1. Introduction

1.1 Cadre théorique

Dans le domaine de l'économie du travail, la relation entre la mobilité et la rémunération des travailleurs a été étudiée sous des angles théoriques variés. Quatre thèses concurrentes se dégagent de la lecture de la littérature pertinente.

Divers modèles de la mobilité de l'emploi s'appuient sur l'opinion prédominante que la mobilité de l'emploi est volontaire et a donc un effet positif sur la rémunération. Par exemple, les modèles de concordance entre les emplois et les travailleurs (*job matching*) prédisent que la mobilité de l'emploi a un effet positif sur la rémunération, puisque les travailleurs décident de quitter leur emploi afin de trouver une meilleure concordance (Jovanovic, 1979). Dans la même veine, la théorie de l'expérimentation des emplois (*job shopping*) (Stigler, 1962) et l'approche de la formation décrite dans Mortensen (1988) expliquent la mobilité de l'emploi. Quand les compétences sont transférables d'une profession à l'autre, la mobilité professionnelle entre emplois et entre employeurs aboutit aussi à des augmentations salariales plus importantes pour les employeurs qui changent d'emploi que pour ceux qui ne le font pas (Sicherman et Galor, 1990).

Un deuxième ensemble de modèles repose sur l'idée que la mobilité de l'emploi a un effet négatif sur la rémunération (Blumen et coll., 1955). Plusieurs hypothèses ancrées dans la théorie du capital humain indiquent que les employés déplacés peuvent accuser une perte salariale liée à la mobilité de l'emploi. S'ils considèrent l'expérience de chômage passée comme un indice de productivité (Vishwanath, 1989), les employeurs pourraient offrir une rémunération plus faible aux travailleurs qui retournent sur le marché du travail après des périodes de chômage (Pissarides, 1992). Les théoriciens de la segmentation du marché du travail insistent sur le fait que les travailleurs qui passent du marché du travail interne/principal au marché du travail externe/secondaire subissent de baisses de salaire. En revanche, la mobilité entre les emplois du segment principal devrait donner lieu à un accroissement de la rémunération.

Selon la troisième thèse concernant la relation entre la rémunération et la mobilité de l'emploi, cette dernière n'est pas une variable explicative importante. Pour ne nommer qu'une seule école de pensée, la théorie du capital humain, qui remonte aux travaux fondamentaux de Becker (1962), soutient que la rémunération augmente parallèlement à l'accumulation de capital humain particulier. Si le salaire rémunère la productivité du travailleur, on peut soutenir que la mobilité de l'emploi n'est pas une variable explicative importante.

Enfin, certains chercheurs dont les travaux s'appuient sur des modèles de recherche d'emploi supposent que les travailleurs se servent de la mobilité de l'emploi et de la mobilité professionnelle comme un moyen de maximiser leurs revenus au cours de la

[1]Yves J. Decady, Statistique Canada, 170 Tunney's Pasture, Ottawa, Canada, K1A 0T6 (yves.decady@statcan.gc.ca)

vie. Selon ces modèles, la mobilité volontaire suscitera une progression positive de la rémunération (Burdett, 1978). La mobilité de l'emploi entraînera la mobilité des salaires, créant ainsi une relation bidirectionnelle entre la mobilité de l'emploi et la rémunération. Par conséquent, la mobilité de l'emploi peut être considérée comme une variable endogène.

Étant donné que chaque thèse offre une justification théorique plausible de la relation entre la rémunération et la mobilité de l'emploi, il va sans dire que le signe attendu de l'effet net de la mobilité de l'emploi sur la rémunération est, dans l'ensemble, ambigu. D'où l'objectif principal du présent article qui sera d'obtenir une estimation cohérente du rendement de la mobilité de l'emploi et de contribuer à la discussion en offrant des preuves empiriques supplémentaires.

La présentation de l'article est la suivante. Les données, les définitions de base ainsi que certains résultats descriptifs sont présentés à la section qui suit. La stratégie d'estimation est élaborée à la section 3. Les résultats sont discutés à la section 4. Enfin, certaines remarques concernant un futur plan de recherche sont exposées à la section 5.

2. Données, définitions de base et résultats descriptifs

2.1 Les données d'enquête de l'EMTE

L'Enquête sur le milieu de travail et les employés (EMTE), qui a été lancée en 1999, est une enquête annuelle longitudinale réalisée auprès d'un échantillon d'employeurs canadiens et d'un échantillon de leurs employés. L'EMTE suit les employés pendant une période de deux ans et les employeurs, pendant une période de huit ans. Au cours du cycle d'enquête, trois panels d'employés d'une durée de deux ans ont été sélectionnés, à savoir les panels de 1999-2000, de 2001-2002, et de 2003-2004. Si l'on compare la situation d'emploi la première année à cette situation la deuxième année, le taux de mobilité d'emploi est défini comme étant le nombre d'employés qui ont changé d'emploi en changeant d'employeur ou en restant chez le même employeur entre les deux années du panel, divisé par le nombre total de personnes occupées durant la première année du panel.

En 2003, 20 834 employés ont été interviewés, et en 2004, 16 804 de ces employés l'ont été de nouveau, dont 990 n'ayant plus d'emploi salarié. De ces 990 employés, 113 qui étaient des travailleurs autonomes ont été exclus. Nous obtenons ainsi un échantillon final de 15 814 travailleurs, dont 14 938 avaient un emploi salarié en 2003 et en 2004. Il semble que la rémunération offerte n'a pas été observée pour 877 personnes en 2004, soit parce qu'elles étaient chômeuses ou inactives. Dans toutes les analyses présentées ici, les poids de sondage et les poids bootstrap sont utilisés pour tenir compte correctement de la nature complexe du plan de sondage.

2.2 Analyse descriptive

Dans l'ensemble, la mobilité de l'emploi a diminué de 1999 à 2004. La proportion de travailleurs qui ont changé d'emploi a baissé de 5 points de pourcentage dans le panel d'employés de 2003-2004, comparativement à celui de 1999-2000 (14,6 % contre 19,7 %). La comparaison des trois panels révèle une diminution particulièrement sensible de la proportion de travailleurs qui ont changé d'emploi en restant chez leur employeur. Se chiffrant à 10,9 % dans le panel d'employés de 1999-2000, la mobilité interne avait diminué de près de 5 points de pourcentage dans le panel de 2003-2004 (10,9 % contre 6,1 %). De plus, il convient de noter que le flux de sortie de l'emploi rémunéré était sensiblement plus important chez les travailleuses que chez leurs homologues masculins dans le troisième panel d'employés.

Dans le panel de 2003-2004, les travailleurs qui étaient restés chez le même employeur et avaient conservé le même emploi avaient obtenu une hausse moins forte de leur rémunération horaire (5,9 %) que ceux qui avaient changé d'emploi en restant chez le même employeur (4,0 %), ou ceux qui avaient changé d'employeur (12,2 %). La mobilité de l'emploi était donc effectivement associée à une prime salariale, qu'elle soit interne ou qu'elle comporte un changement d'employeur. Toutefois, la mobilité n'était pas toujours rémunératrice.

L'analyse empirique portera sur les questions suivantes. Premièrement, nous déterminons dans quelle mesure le rendement salarial est plus élevé chez travailleurs entrant dans le panel qui changent d'employeur l'année suivante que chez les travailleurs non mobiles. Deuxièmement, nous accordons une attention particulière aux différences entre les primes salariales liées à la mobilité des hommes et des femmes. Enfin, nous étudions l'effet éventuel de sélection de l'échantillon sur la prime salariale associée à la mobilité de l'emploi. Nous utilisons pour cela les données recueillies auprès du panel d'employés de 2003-2004 de l'EMTE.

3. Stratégie d'estimation

3.1 Trois grands problèmes méthodologiques

Comme il est mentionné plus haut, l'objectif principal de la présente étude est d'obtenir des mesures cohérentes de l'effet de la mobilité de l'emploi sur la rémunération des membres du panel d'employés de 2003-2004. Cependant, trois grands problèmes méthodologiques qui menacent la cohérence des résultats doivent être abordés, c'est-à-dire l'hétérogénéité non observée, l'endogénéité de la mobilité de l'emploi et l'éventuel biais de sélection dans l'échantillon.

La mobilité de l'emploi est vraisemblablement corrélée à des caractéristiques inobservables, telles que l'ambition, les aptitudes et la motivation, qui ont aussi une influence sur la rémunération. Par conséquent, l'utilisation de la méthode des moindres carrés ordinaires (MCO) produira des estimations biaisées du rendement de la mobilité de l'emploi. Souvent, les facteurs non observés ou l'hétérogénéité non observée ont été traités par estimation des effets fixes (EF) (Light et McGarry, 1998; Arulampalamm 2001; Gregory et Roberts, 2001; Munasinghe et Sigman, 2004). Dans le présent article, nous utilisons la méthode des différences premières (DP). Cette méthode fournit des estimations et des inférences identiques à celles obtenues par la méthode des effets fixes si le nombre de périodes observées est égal à deux. Le calcul des différences élimine les effets propres à la personne, ainsi que les caractéristiques temporellement invariantes des travailleurs et des emplois.

En ce qui concerne le traitement du problème de l'endogénéité, un petit nombre de stratégies ont été avancées, notamment des modèles structurels (Flinn, 1986; Antel, 1991) et des modèles de changement endogène. Une méthode utilisée fréquemment est celle des variables instrumentales (VI). À titre d'exemples d'articles décrivant l'utilisation de variables instrumentales, nous mentionnerons Topel (1991) ainsi que Altonji et Williams (2005). La stratégie d'estimation décrite dans le présent article s'appuie sur l'estimation par variables instrumentales (Angrist et Krueger, 2001) afin de tenir compte de l'endogénéité de la variable binaire de mobilité de l'emploi.

Le troisième problème se pose parce que les changements de rémunération ne sont observés que pour les travailleurs qui avaient un emploi salarié en 2003 ainsi qu'en 2004. Le biais d'autosélection pourrait poser problème si les déterminants non observés de la rémunération ont également une incidence sur la décision des individus de participer au marché du travail. L'effet d'une éventuelle autosélection dans l'échantillon sera examiné au moyen d'une nouvelle méthode introduite par Semikyna et Wooldridge (2006).

3.2 Modèles de régression aux différences premières et à variables instrumentales

Nous adoptons une stratégie de modélisation en deux volets. Pour commencer, nous cherchons à éliminer l'hétérogénéité non observée et à tenir compte de l'endogénéité de la mobilité de l'emploi dans l'estimation du rendement de cette dernière, tout en ignorant la sélection de l'échantillon. Ensuite, nous évaluons l'effet d'un éventuel biais d'autosélection de l'échantillon sur les estimations. Considérons le modèle des effets non observés :

$$Y_{it} = X_{it}'\beta + (mobemploi)_{it}\gamma + \delta_i + u_{it}, \quad t=1,2$$

où Y_{it} représente le salaire horaire réel[2] du travailleur i au temps t, X_{it} est un vecteur des caractéristiques du travailleur i au temps t; $(mobemploi)_{it}$ est une variable binaire égale à 1 si un travailleur change d'employeur la deuxième année et à 0 autrement, β et γ sont des paramètres de régression qui doivent être estimés, δ_i traduit les effets non observés, et u_{it} est le terme d'erreur particulier. Le coefficient γ est le rendement de la mobilité de l'emploi. Les variables explicatives comprennent l'âge, l'ancienneté, la position sur la courbe de distribution des salaires avant le changement d'employeur, le niveau de scolarité, l'emploi permanent par opposition à temporaire, la profession, la situation syndicale, le revenu des autres membres du ménage, la formation théorique et les avantages sociaux. Considérons maintenant l'équation aux différences premières :

$$\Delta Y_{it} = \Delta X_{it}'\beta + (mobemploi)_{it}\gamma + \Delta u_{it}$$

où Δ indique la différence entre les valeurs de la deuxième et de la première année. En prenant les différences, on élimine les effets particuliers à la personne, ainsi que les caractéristiques temporellement invariantes des travailleurs et des emplois. Les covariables temporellement variantes retenues dans nos spécifications estimées sont les variations entre les périodes t et $t+1$.

Afin de résoudre le problème de l'endogénéité, nous avons de nouveau estimé le modèle aux différences premières en utilisant comme variables instrumentales le nombre de changements d'emploi au cours des cinq dernières années moins sa moyenne, le nombre d'emplois vacants moins sa moyenne, ainsi que la satisfaction à l'égard de l'emploi. L'utilisation du nombre de changements d'emploi antérieurs repose sur la notion intuitive selon laquelle le nombre de changements d'emploi au cours des

[2] La rémunération horaire nominale est déflatée au moyen de l'Indice des prix à la consommation pour chaque province, afin d'obtenir la rémunération horaire réelle en utilisant 2002 comme année de base.

cinq dernières années représente une approximation de l'investissement antérieur dans la recherche d'emplois et, par conséquent, qu'il tient compte des chocs antérieurs ayant une incidence sur la situation de mobilité courante. Parallèlement, le nombre de changements d'emploi ne devrait pas avoir d'effets sur la rémunération autrement que par la voie de la situation de mobilité courante. Nous supposons que les travailleurs changent d'employeurs plus fréquemment quand les offres d'emploi sont plus nombreuses, ce qui influe positivement sur la probabilité d'un changement volontaire d'emploi. L'utilisation de l'écart par rapport au nombre moyen de changements d'emploi, ainsi que de l'écart par rapport au nombre moyen d'emplois vacants donne lieu à une orthogonalité. Pour ce qui est de la variable instrumentale de satisfaction à l'égard de l'emploi, si on offre un niveau plus élevé de satisfaction, les individus semblent plus disposés à accepter une baisse de rémunération que s'ils étaient plus heureux dans leur emploi. On pourrait peut-être soutenir que les travailleurs satisfaits demandent moins d'augmentations de salaire pour compenser une situation d'emploi médiocre que les travailleurs qui sont très insatisfaits de leur emploi.

Nous supposons que les travailleurs sont plus susceptibles de quitter involontairement leur emploi si l'effectif d'un employeur est réduit significativement à cause de mises à pied, de congédiements et de compressions volontaires de l'effectif. D'où, l'utilisation comme variable instrumentale pour les changements d'emploi involontaires du nombre total de séparations involontaires (nombre de mises à pied plus le nombre de congédiements) moins sa moyenne, plutôt que le nombre d'offres d'emploi moins sa moyenne. Ces variables instrumentales sont fortement corrélées à la probabilité que les travailleurs changent d'employeur, ce qui en fait des instruments puissants. En outre, il paraît plausible de soutenir qu'elles ne sont pas corrélées aux caractéristiques personnelles non observables qui ont une incidence sur la rémunération.

Afin de tenir compte du biais de sélection de l'échantillon, nous avons utilisé une équation de sélection probit de Mundlak (1978) et avons calculé les inverses des ratios de Mills. Les équations de sélection de l'échantillon reflètent le fait qu'une partie de la population ne reçoit aucun salaire parce qu'elle est soit au chômage, soit inactive. Nous avons étoffé les modèles aux différences premières au moyen des inverses des ratios de Mills en utilisant la régression par les doubles moindres carrés à effets fixes (DMC-EF) sur les inverses des ratios de Mills, qui ont été calculés en partant de trois équations probit de participation distinctes (tous travailleurs confondus, travailleurs masculins et travailleurs féminins). Dans chaque cas, le test n'a pas donné lieu au rejet de l'hypothèse d'absence de sélection. Par conséquent, l'approche des DMC-EF a fourni des estimations cohérentes du rendement de la mobilité de l'emploi.

4. Résultats de l'estimation

4.1 Résultats empiriques sur la mobilité de l'emploi

La colonne 1 du tableau 4.1 fournit une liste partielle des variables explicatives pour deux régressions distinctes. La régression 1 comprend comme principale variable explicative les travailleurs qui changent d'employeurs par opposition à ceux qui ne changent pas d'employeurs, et la régression 2 utilise les travailleurs qui changent volontairement d'employeurs par opposition à ceux qui ne changent pas d'employeurs. Les équations de la croissance de la rémunération sont estimées pour l'ensemble des travailleurs, ainsi que pour les hommes et les femmes séparément. Les colonnes 2, 4 et 5 du tableau 4.1 contiennent les résultats de la régression des différences premières de l'analyse de la croissance de la rémunération pour le panel d'employés de 2003-2004. La mobilité de l'emploi entraîne un rendement salarial positif dans l'année après que le changement a eu lieu, comparativement au groupe de travailleurs non mobiles ayant des caractéristiques semblables. Comparativement aux travailleurs non mobiles, ceux qui changent d'employeurs gagnent une prime de 7,0 %. Ce rendement diffère de zéro de manière significative pour tous les travailleurs confondus, ainsi que pour les hommes et pour les femmes. Deuxièmement, le rendement est plus élevé pour les femmes que pour les hommes dans les modèles aux différences premières (DP). La mobilité volontaire est récompensée par un rendement plus élevé (9,6 %). Comme les travailleuses mobiles gagnent une prime salariale appréciable, la mobilité de l'emploi pourrait contribuer à atténuer, au cours du temps, les écarts salariaux entre les hommes et les femmes.

Les colonnes 3, 6 et 7 du tableau 4.1 contiennent les résultats des régressions avec variables instrumentales. Avant d'analyser les résultats, nous commençons par examiner la validité des variables instrumentales. Dans tous les modèles avec VI, le test d'Anderson-Rubin Wald mène au rejet de l'hypothèse nulle, ce qui indique que les trois variables instrumentales sont hautement significatives (1,0 %) et fortement corrélées à la mobilité de l'emploi. La statistique *J* de Hansen ne donne pas lieu au rejet de l'hypothèse nulle selon laquelle toutes les variables instrumentales sont sans corrélation avec le processus de perturbation dans tous les cas. C'est pourquoi la validité des variables instrumentales n'est pas mise en doute. L'hypothèse nulle d'endogénéité de la variable de mobilité de l'emploi a également été rejetée en utilisant le test de Hausman, ce qui confirme la nature bidirectionnelle de la relation entre la rémunération et la mobilité de l'emploi avancée dans diverses théories de la mobilité de l'emploi.

Tableau 4.1.1
Résultats des modèles de régression aux différences premières et à variables instrumentales

Certaines variables explicatives (1)	Tous travailleurs confondus		DP		VI	
	DP (2)	VI (3)	Hommes (4)	Femmes	Hommes (6)	Femmes (7)
1) Tous les travailleurs qui ont changé d'emploi par rapport à ceux n'ayant pas changé	0,0700***	0,0847***	0,0519***	0,0841***	0.0664***	0.1059***
Position sur la courbe de distribution des salaires						
Travailleurs faiblement rémunérés						
Travailleurs moyennement rémunérés	0,0321***	0,0315***	0,0366***	0,0312***	0.0386***	0.0310***
Travailleurs fortement rémunérés	0,0562***	0,0550***	0,0638***	0,0541***	0.0667***	0.0544***
Khi-deux-J de Hansen (2) Valeur P Khi-deux-test de Wald (3) Valeur P		3,595 (0,1657) 26,82 (0,0000)			3.971 (0.1373) 11.80 (0.0081)	3.601 (0.1652) 19.41 (0.0002)
2) Travailleurs ayant changé volontairement d'emploi comparativement à ceux qui n'ont pas changé	0,0956***	0,1135**	0,0920***	0,0993***	.0995***	0.1257***
Position sur la courbe de distribution des salaires						
Travailleurs faiblement rémunérés						
Travailleurs moyennement rémunérés	0,0314***	0,0311***	0,0335***	0,0316***	0.0316***	0.0315***
Travailleurs fortement rémunérés	0,0491***	0,0490***	0,0523***	0,0549***	0.0502***	0.0561***
Khi-deux-J de Hansen (2) Valeur P Khi-deux-test de Wald (3) Valeur P		2,242 0,3260 32,71 (0,0000)			3.686 0.1583 18.37 (0.0004)	2.746 0.2533 18.50 (0.0003)

Nota : Coefficients de régression significatifs au seuil de : ***= 1 %; **=5 %; *= 10 %.

Si nous passons maintenant à l'analyse des estimations par variables instrumentales du rendement de la mobilité de l'emploi, nous constatons que ces estimations sont plus élevées que celles obtenues par la méthode des différences premières. Comme dans le cas des estimations par DF, les femmes semblent bénéficier d'une plus forte augmentation salariale que les hommes. Si l'on examine les résultats selon la position du travailleur sur la courbe de distribution des salaires, deux constatations se dégagent : premièrement, les travailleurs fortement rémunérés obtiennent un rendement plus élevé que les travailleurs moyennement rémunérés. Ces derniers ont, à leur tour, un rendement plus élevé que les travailleurs faiblement rémunérés. Le rendement salarial de la mobilité de l'emploi augmente avec la position sur la courbe de distribution des salaires. Dans l'ensemble, si l'on considère la position sur cette courbe, le rendement de la mobilité de l'emploi est plus élevé pour les hommes que pour les femmes. Toutefois, il n'en est pas systématiquement ainsi dans le cas de la mobilité volontaire.

5. Conclusion

Les données de l'EMTE fournissent de nombreuses preuves qu'une prime salariale est associée à la mobilité de l'emploi. Il en est clairement ainsi de la mobilité globale ainsi que de la mobilité volontaire de l'emploi, ce qui confirme les résultats d'études antérieures. La mobilité volontaire est récompensée par un plus haut rendement. Royalty (1998), ainsi que Gladden et Taber (2000) ont constaté qu'aux États-Unis, les changements volontaires d'employeurs se traduisent par un accroissement de la rémunération; en Europe, Perez et Sanz (2005) ont tiré des conclusions semblables. Il existe également des preuves d'une forte mobilité salariale sans mobilité de l'emploi. Par contre, les données ne fournissent pas suffisamment de preuves d'une pénalité salariale liée à la mobilité involontaire de l'emploi. Le présent article a pour but d'examiner la relation entre la rémunération et la mobilité de l'emploi en se servant de données d'enquête longitudinales et ne tient compte que des effets fixes associés à l'employé. Les travaux de recherche à venir tiendront compte dans l'analyse de la nature liée des données d'enquête afin de refléter à la fois les effets des caractéristiques de l'employeur et de l'employé en utilisant des modèles multiniveaux.

Remerciements

L'auteur remercie Cynthia Bocci, Jean-François Beaumont, Harry François et Fritz Pierre des commentaires constructifs qu'ils ont fait durant l'élaboration de la présente étude.

Bibliographie

Altonji, J.G. et Williams, N. (2005). Do Wages Rise with Job Seniority? A Reassessment, *Industrial & Labor Relations Review*, 58, pp. 370-397.

Angrist, J.A. et Krueger, A.B. (2001). Instrumental Variables and the Search for Identification: from Supply and Demand to Natural Experiments. *Journal of Economic Perspectives,* 15, pp. 69-85.

Antel, J.J. (1991). The Wage Effects of Voluntary Labor Mobility with and without Intervening Unemployment, *Industrial and Labor Relations Review*, 44, pp. 299-306.

Arulampalam, W. (2001). Is Unemployment Really Scarring? Effects of Unemployment Experiences on Wages, *The Economic Journal*, 111, pp. F585-F606.

Becker, G. (1962). Investment in Human Capital: A Theoretical Analysis, *Journal of Political Economy*, 70, pp. S9-S49.

Blumen, I., Kogen, M. et McCarthy, P. (1955). The Industrial Mobility of Labor as a Probability Process, *Cornell Studies in Industrial and Labor Relations* 6, Ithaca: New York.

Burdett, K (1978). A Theory of Employee Job Search and Quit Rates, *American Economic Review*, 68, pp. 212-220.

Flinn, C.J. (1986). Wages and Job Mobility of Young Workers, *Journal of Political Economy*, 94, pp: S88-S110.

Gladden, T. et Taber, C. (2000). Wage Progression among Less-skilled Workers, in David E. Card and Rebecca M. Blank (eds.) *Finding Jobs, Work and Welfare Reform*, New York: The Russell Sage Foundation, pp. 160-192.

Gregory, M. et Roberts, J. (2001). Unemployment and Subsequent Earnings: Estimating Scarring among British Men 1984-94, *The Economic Journal*, 111, pp: F607-F625.

Jovanovic, B. (1979). Job Matching and the Theory of Turnover, *Journal of Political Economy*, 87, pp. 972-990.

Light, A. et McGarry, K. (1998). Job Change Patterns and the Wages of Young Men, *The Review of Economics and Statistics*, 80, pp. 276-286.

Mortensen, D. (1988). Wage Separations and Job Tenure: On-the-job Specific Training or Matching, *Journal of Labor Economics*, 6, pp. 445-470.

Munasinghe, L. et Sigman, K. (2004). A Hobo Syndrome? Mobility, Wages, and Job turnover, *Labour Economics*, vol. 11, nº. 2, pp. 191-218.

Mundlak, Y. (1978). On the Pooling of Time Series and Cross-Section Data, *Econometrica*, 46, pp. 69-85.

Perez, J.H.G et Sanz, Y.R. (2005). Wage Changes through Job Mobility in Europe: A Multinomial Endogenous Switching Approach, *Labor Economics*, 12, pp. 531-555.

Pissarides, C. (1992). Loss of Skills during Unemployment and Persistence of Employment Shocks, *Quarterly Journal of Economics*, 107, pp. 1381-13911.

Stigler, G. (1962). Information in the Labour Market, *Journal of Political Economy,* 70, pp. 94-105.

Royalty, A. B. (1998). Job-to-job and Job-to-unemployment Turnover by Gender and Education Level, *Journal of Labor Economics*, 16, pp.441-479.

Semykina, A. et Wooldridge, J. M. (2006). Estimating Panel Data Models in Presence of Endogeneity and Selection: Theory and Application, discussion paper, Michigan State University, East Lansing, MI: Department of Economics.

Sicherman, N. et Galor, O. (1990). A Theory of Career Mobility, *Journal of Political Economy*, 98, pp. 169-192.

Topel, R. (1991). Specific Capital, Mobility and Wages: Wage Rise with Job Seniority, *Journal of Political Economy*, 99, pp. 145-176.

Vishwanath, T. (1989). Job Search, Stigma Effect and Escape Rate from Unemployment, *Journal of Labor Economics*, 7, pp. 487-502.

SESSION 21 – TECHNIQUES D'ANALYSE DE DONNÉES LONGITUDINALES

L'utilisation de techniques d'analyse exploratoire et confirmative de données longitudinales

Marcel de Toledo Vieira, Ronaldo Rocha Bastos, Henrique Steinherz Hippert et Augusto Carvalho Souza[1]

Résumé

Cette communication porte sur l'utilisation de diverses approches de l'analyse des données d'enquête longitudinale, y compris des techniques alternative d'analyse exploratoire des données et différentes stratégies de modélisation par régression aux fins d'analyses longitudinales des données de la British Household Panel Survey sur les attitudes à l'égard des rôles de l'homme et de la femme ainsi que leur relation avec les variables démographiques et économiques. La question générale posée dans la présente communication est la suivante : tirerait-on des conclusions et des inférences différentes, selon l'approche choisie? On procède à des analyses exploratoires ainsi que confirmatoires des données longitudinales au moyen de techniques d'analyse des correspondances et de modélisation par régression aux fins de l'analyse des relations adaptatives. Les résultats de l'analyse des correspondances sont confirmés de façon générale par les estimations des paramètres pour les modèles de régression, qui portent souvent le même signe que les relations qui s'affichent sur les cartes d'analyse des correspondances. Les preuves empiriques montrent que le choix de l'approche analytique et de la stratégie de modélisation est une question importante dans le contexte de l'analyse de données longitudinales. Par conséquent, nous recommandons de faire ce choix en fonction des objectifs de l'analyse longitudinale.

[1] Marcel de Toledo Vieira (marcel.vieira@ufjf.edu.br), Ronaldo Rocha Bastos (ronaldo.bastos@ufjf.edu.br) et Henrique Steinherz Hippert, Federal University of Juiz de Fora, Brésil; Augusto Carvalho Souza, Federal University of Minas Gerais, Brésil.

Mesures de la qualité de l'ajustement pour les modèles fondés sur une approche des équations d'estimation généralisées

Punam Pahwa[1]

Résumé

L'évaluation de la mesure dans laquelle le modèle est bien ajusté aux données (la qualité de l'ajustement) est une partie importante de tout processus de choix d'un modèle. Même si de nombreuses méthodes d'analyse de données longitudinales ont été élaborées au cours des vingt dernières années, il n'existe toujours pas de mesures standard raisonnables de la qualité de l'ajustement applicables à ces modèles. Pour les données longitudinales, nous avons besoin de statistiques d'évaluation de la qualité de l'ajustement pour choisir non seulement une fonction de réponse correcte mais une structure de corrélation/de covariance intra-sujet appropriée. Les statistiques de la qualité de l'ajustement fondées sur des méthodes de vraisemblance comme le test de rapport de vraisemblance et les critères d'information d'Akaike (i) requièrent des ajustements répétés des données à une famille de modèles emboîtés, (ii) exigent une spécification complète de la fonction de vraisemblance et (iii) ne peuvent être utilisées pour évaluer le caractère adéquat des modèles ajustés au moyen de l'approche fondée sur les équations d'estimation généralisées (EEG). Vonesh et coll. ont élaboré trois statistiques de la qualité de l'ajustement : (i) rc – coefficient de concordance servant à mesurer la concordance entre les réponses ajustées et les réponses observées; (ii) r($\omega\hat{}$) – une mesure de la concordance entre les structures de covariance hypothétique et réelle; et (iii) - pour tester l'égalité entre les structures de covariance hypothétique et réelle (indirectement, en testant l'égalité entre les structures de covariance « sandwich » et hypothétique). Ces trois mesures sont fondées exclusivement sur le modèle dont il est question. Nous proposons d'utiliser ces mesures pour évaluer la qualité de l'ajustement des modèles ajustés au moyen de l'approche EEG aux fins de l'analyse des données longitudinales recueillies (sans plan de sondage) sur la santé respiratoire des travailleurs dans les silos-élévateurs canadiens. Nous tâcherons de modifier ces mesures pour évaluer l'adéquation des modèles aux données d'enquête complexe longitudinales.

[1] Punam Pahwa, University of Saskatchewan, Canada (pup165@mail.usask.ca)

SESSION 22 – AJUSTEMENTS POUR LA NON-RÉPONSE ET L'ÉROSION

L'érosion de l'échantillon dans les études de cohorte : tendances, caractéristiques et ajustements

Ian Plewis, Lisa Calderwood et Sosthenes Ketende[1]

Résumé

La non-réponse étant omniprésente dans les études longitudinales, nous évaluons l'exactitude de modèles statistiques construits pour prédire divers types de non-réponse. Nous accordons une attention particulière aux mesures sommaires dérivées des courbes de fonction d'efficacité de l'observateur (ou courbe ROC, de l'anglais *receiver operating characteristic*) et des courbes logit des rangs (*logit ranks plots*). Nous appliquons les idées aux données des quatre premières vagues de la Millennium Cohort Study du Royaume-Uni. Les résultats donnent à penser que notre capacité de distinguer et de prédire la non-réponse n'est pas grande. Nous tirons des conclusions quant aux méthodes d'ajustement de la non-réponse.

Mots-clés : non-réponse, courbes ROC, coefficient de Gini, pondérations, Millennium Cohort Study.

1. Introduction

Le présent article a pour thème général l'apprentissage que nous pouvons tirer de la modélisation des prédicteurs des divers types de non-réponse dans les études longitudinales. Il est fréquent de modéliser le logit (ou le probit) de la probabilité de ne pas répondre, puis d'utiliser les valeurs prévues d'après le modèle pour produire les poids (c.-à-d. l'inverse de la probabilité) de non-réponse. Nous pouvons également utiliser l'information sur les prédicteurs statistiquement importants tirée de la régression logistique estimée comme description du mécanisme de formation des données manquantes, puis intégrer ce mécanisme dans une procédure d'imputation multiple, par exemple celle décrite par Carpenter et Plewis (2010). Toutefois, nous ne disposons souvent d'aucune donnée sommaire utile sur l'exactitude du modèle que nous choisissons et nous ne sommes donc pas certains de ce que valent nos renseignements sur les données manquantes et de leur incidence en ce qui concerne tant l'ajustement statistique que l'utilité des interventions possibles en vue de prévenir la non-réponse.

Les idées exposées dans le présent article sont illustrées au moyen des données de la Millennium Cohort Study (MCS), la quatrième d'une série d'études de cohortes de naissances de réputation internationale réalisées au Royaume-Uni. À la section 2, nous décrivons brièvement la MCS ainsi que les profils de non-réponse au cours des quatre premières vagues de l'étude. À la section 3, nous examinons comment peut être mesurée l'exactitude des modèles prédicteurs de la non-réponse et présentons deux mesures sommaires, à savoir le coefficient de Gini dérivé de l'aire sous la courbe de la fonction d'efficacité de l'observateur ou courbe ROC (de l'anglais *receiver operating characteristic*) et la pente de la courbe logit des rangs (*logit rank plot*) introduite par Copas (1999). À la section 4, nous examinons diverses spécifications pour les modèles qui prédisent la non-réponse et à la section 5, nous concluons par une discussion de ce qu'impliquent nos constatations en ce qui concerne l'ajustement statistique.

2. La Millennium Cohort Study

L'échantillon de la première vague de la MCS comprenait 18 818 bébés appartenant à 18 552 familles nés au Royaume-Uni sur une période de 12 mois en 2000 et 2001, et qui, à l'âge de neuf mois, vivaient dans des circonscriptions électorales sélectionnées du Royaume-Uni. Étant donné qu'au Royaume-Uni, les mères de presque tous les nouveau-nés ont droit à la prestation pour enfant, le registre de la prestation pour enfant a été utilisé comme base de sondage et le taux de réponse initial était de 72 %. Les secteurs où les proportions de familles noires et asiatiques sont élevées, les secteurs défavorisés et les trois plus petits pays du Royaume-Uni sont tous surreprésentés dans l'échantillon qui est structuré en grappes et stratifié de façon disproportionnée, comme l'a décrit Plewis (2007a). Les quatre premières vagues ont eu lieu lorsque les membres de la

[1] Ian Plewis, Social Statistics, University of Manchester, UK, M13 9PL (ian.plewis@manchester.ac.uk); Lisa Calderwood, Centre for Longitudinal Studies, Institute of Education, London, UK, WC1H 0AL; Sosthenes Ketende, Centre for Longitudinal Studies, Institute of Education, London, UK, WC1H 0AL.

cohorte avaient (environ) neuf mois, trois ans, cinq ans et sept ans. Les partenaires ont été interviewés dans la mesure du possible, et des données ont également été recueillies auprès des enfants membres de la cohorte en personne, ainsi qu'auprès de leurs frères et sœurs plus âgés.

2.1 Érosion de l'échantillon de la Millennium Cohort Study

Le tableau 2.1-1 montre comment l'échantillon de la MCS s'est érodé au fil du temps après la première vague. L'érosion de l'échantillon correspond aux non-répondants à une vague, c'est-à-dire les cas qui manquent à la vague t mais non à une ou à plusieurs vagues subséquentes, ainsi que les cas d'attrition qui, après avoir manqué une vague, ne réapparaissent pas. Il est impossible d'affecter définitivement les cas à l'une de ces deux catégories de non-réponse avant la fin de l'étude, mais le tableau 2.1-1 indique que l'érosion de l'échantillon de la MCS, comme celui de la plupart des études longitudinales dans le domaine des sciences sociales, correspond à un mélange de cas de non-réponse à une vague et d'érosion et est donc non monotone. Le tableau 2.1-1 révèle aussi qu'à la vague 2, les non-répondants se répartissent approximativement à parts égales entre la catégorie des refus et celle des autres cas non productifs (non localisés, non contactés, etc.), mais que les cas de refus deviennent plus dominants par après. Il convient de noter que la taille de l'échantillon admissible, qui n'inclut pas les enfants décédés et émigrés, augmente de la vague 2 à la vague 3; certains cas omis à la vague 1 en Angleterre ont été recrutés pour la première fois à la vague 2.

Tableau 2.1-1
Érosion de l'échantillon de la MCS selon le type de non-réponse

	Vague 2, 3 ans	Vague 3, 5 ans	Vague 4, 7 ans
Non-réponse à une vague	8,3%	3,3%	n,a,
Attrition	9,9%	16%	n,a,
Total	18%	20%	26%
Refus	9,1%	12%	19%
Autres cas non productifs	9,2%	7,3%	7,4%
Taille de l'échantillon admissible	18 385	18 944	18 756

s.o. sans objet car la non-réponse à une vague n'est pas définie pour la vague la plus récente.

2.2 Prédicteurs de la non-réponse à la vague 2 de la MCS

Les résultats des recherches présentées par Plewis (2007b) et Plewis et coll. (2008) sur les prédicteurs des divers types de non-réponse à la MCS sont résumés au tableau 2.2-1.

Tableau 2.2-1
Prédicteurs de la non-réponse, vague 2 de la MCS

Prédicteur de la vague 1	Non-réponse à une vague	Attrition	Refus	Autres cas non productifs
Changement de résidence après la vague 1	✓	×	×	✓
Pays du Royaume-Uni	✓	✓	✓	✓
Revenu familial	×	✓	✓	×
Refus de répondre à la question sur le revenu	×	×	✓	×
Groupe ethnique	✓	✓	×	✓
Mode d'occupation du logement	✓	✓	×	✓
Type de logement	✓	✓	✓	✓
Âge de la mère	✓	✓	✓	✓
Niveau de scolarité	✓	✓	✓	✓
Fourniture d'une adresse stable	×	✓	✓	✓
Membre de la cohorte allaité	✓	✓	✓	✓
Maladie chronique	✓	✓	✓	✓
Partenaire présent		✓	✓	✓
Partenaire – sans interview	✓	✓	✓	✓

Le tableau 2.2-1, fondé sur une régression logistique multinomiale, montre que les variables mesurées à la vague 1 qui prédisent l'érosion ne prédisent pas forcément la non-réponse à une vague (et inversement). Il en est de même des prédicteurs du refus et des autres cas non productifs. Soulignons que certains prédicteurs — refus de répondre à la question sur le revenu, fourniture d'une adresse stable et pas d'interview du partenaire —sont des variables sans grand intérêt qui pourraient entrer dans la catégorie des « paradonnées » (Couper et Lyberg, 2005). Dans la conclusion, nous commentons brièvement la valeur de ce type de variables dans le processus de correction pour les données manquantes.

3. Résumé de l'exactitude des prévisions

Le tableau 2.2-1 montre clairement que les divers types de non-réponse à la vague 2 sont systématiquement liés à des variables mesurées à la vague 1. Ce qui est moins clair est la mesure dans laquelle les modèles arrivent à distinguer, ou à prédire, les répondants et les non-répondants. Nous pouvons considérer les fonctions estimées à partir des régressions logistiques comme des scores de risque, puis nous interroger quant à l'exactitude de ces scores. Une méthode très répandue pour évaluer l'adéquation des modèles pour des résultats binaires ou catégoriques consiste à utiliser l'une des diverses statistiques du pseudo-R^2 possibles. Hormis sa nature relativement arbitraire, qui rend les comparaisons entre ensembles de données difficiles, le pseudo-R^2 n'est pas utile dans le présent contexte, parce qu'il évalue l'adéquation globale du modèle et ne fait pas la distinction entre l'exactitude du modèle pour les répondants et les non-répondants pris séparément.

L'exactitude possède deux composantes connexes, la classification (ou discrimination) et la prédiction (Pepe, 2003). La classification s'entend des probabilités conditionnelles d'obtenir un score de risque (s) supérieur à un seuil choisi (t) sachant qu'une personne est ou n'est pas un répondant. La prédiction, par ailleurs, s'entend des probabilités conditionnelles d'être un non-répondant étant donné un score de risque supérieur ou inférieur au seuil.

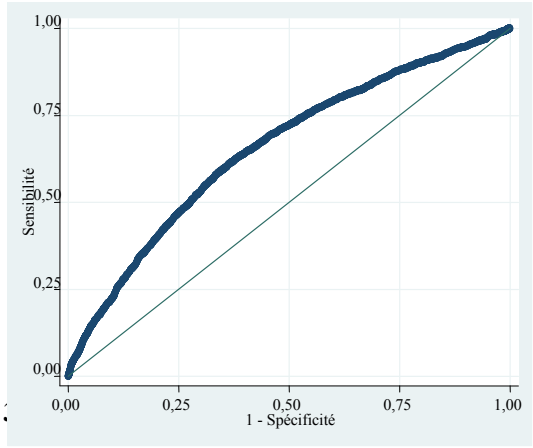
Plus formellement, soit D et $\bar{D}$, la présence et l'absence du mauvais résultat et + ($s > t$) et – ($s \leq t$), les tests positifs et négatifs dérivés du score de risque. Alors, pour la classification, nous nous intéressons à P(+|D), la fraction de vrais positifs (FVP) ou sensibilité du test, et à P(-|$\bar{D}$), sa spécificité, qui est égale à un moins la fraction de faux positifs (1 – FFP). Pour la prédiction, par contre, nous nous intéressons à P(D|+), la valeur prédictive positive (VPP) et à P($\bar{D}$|-), la valeur prédictive négative (VPN). Si la probabilité d'un test positif (P(+) = τ) est la même que la prévalence du mauvais résultat (P(D) = ρ), alors les inférences au sujet de la classification et de la prédiction sont essentiellement les mêmes. Dans ce cas, la sensibilité est égale à la VPP et la spécificité est égale à la VPN. Cependant, en général, (FVP, FFP, ρ) et (VPP, VPN, τ) communiquent des éléments d'information différents.

La mesure dans laquelle les scores de risque permettent de distinguer les répondants et les non-répondants peut être utilisée comme une indication de la mesure dans laquelle nos ajustements statistiques seront influents et, éventuellement, la mesure dans laquelle ils seront efficaces. Un manque de pouvoir de discrimination donne à penser soit que des prédicteurs importants manquent dans le score de risque, soit qu'une partie importante du mécanisme de formation des données manquante est essentiellement aléatoire. La mesure dans laquelle les scores de risque prédisent si un cas sera un non-répondant à la ou aux vagues subséquentes est un indice du succès qu'aura toute intervention destinée à réduire la non-réponse, aussi bien conçue et ciblée qu'elle soit.

3.1 Courbe ROC

Nous pouvons représenter graphiquement la fraction de vrais positifs (c.-à-d. la sensibilité) en fonction de la fraction de faux positifs (c.-à-d. 1 – la spécificité) pour tout seuil t. On obtient ainsi la courbe ROC, parfois appelée courbe de fonction d'efficacité de l'observateur (figure 3.1-1). La courbe ROC a toujours pour point d'ancrage les coordonnées (0,0) et (1,1) et, pour les grands échantillons et au moins certains prédicteurs mesurés de manière continue, elle est lisse et monotonement décroissante, mais possède toujours une pente positive. Krzanowski et Hand (2009) discutent en détail de la façon d'estimer les courbes ROC. La droite diagonale joignant le point [0, 0] (sensibilité = 0, spécificité = 1 : prédiction que personne n'aura un mauvais résultat de sorte que le seuil sur l'échelle de probabilité est égal à un) au point [1, 1] (sensibilité = 1, spécificité = 0 : prédiction que tout le monde aura un mauvais résultat de sorte que le seuil est égal à zéro) est la courbe ROC qui serait obtenue si les variables choisies pour construire le score de risque n'expliquaient aucune partie de la variation du résultat. Par conséquent, c'est l'aire sous la courbe (ASC), c'est-à-dire l'aire délimitée par la courbe ROC et la diagonale, qui nous intéresse, et celle-ci peut varier de 1 à 0,5. Le coefficient de Gini (= 2*ASC-1) est, toutefois, utilisé ici, car il s'agit d'une mesure plus naturelle, qui varie de 0 à 1.

Figure 3.1-1
Courbe ROC



Copas (1999) propose la courbe logit des rangs comme alternative à la courbe ROC pour évaluer le caractère prédictif d'un score de risque. Si le score de risque est calculé d'après une régression logistique, la courbe logit des rangs est la représentation graphique du prédicteur linéaire issu du modèle de régression logistique en fonction de la transformation logistique du rang proportionnel du score de risque. Plus généralement, il s'agit du tracé de (p_i) en fonction des logit des rangs proportionnels (r/n) où p_i est la probabilité estimée d'un mauvais résultat pour le cas i et r est le rang du cas i ($i = 1..n$) sur le score de risque. La pente de cette relation — qui peut varier de zéro à un — est une mesure du pouvoir prédicteur du score de risque et Copas soutient qu'elle est plus sensible aux changements de spécification du modèle qui sous-tend le score de risque que ne l'est le coefficient de Gini. La pente est indépendante de l'échelle et peut, par conséquent, être utilisée pour comparer les scores de risque pour le résultat d'intérêt. Une bonne estimation de la pente peut être obtenue en calculant les quantiles de la variable sur les axes des y et des x, puis en ajustant un simple modèle de régression.

3.3 Résultats

Le tableau 3.3-1 donne les coefficients de Gini et les pentes des courbes logit des rangs pour la non-réponse globale à la vague deux ainsi que pour les divers types de non-réponse. L'estimation de 0,39 pour le coefficient de Gini dans le cas de la non-réponse globale est relativement faible, ce qui indique que la distinction entre les non-répondants et les répondants d'après le score de risque n'est pas bonne, mais elle est légèrement meilleure pour les non-répondants à une vague et pour les autres cas non productifs que pour l'attrition et les refus. Un tableau semblable se dégage pour les pentes des courbes logit des rangs, quoique celles-ci fassent ressortir plus clairement les différences de caractère prédictif pour les divers types de non-réponse.

Tableau 3.3-1
Mesures de l'exactitude, vague 2 de la MCS

	ROC : Gini	Courbe logit des rangs : pente
Non-réponse globale	0,39	0,45
Non-réponse à une vague	0,43	0,52
Attrition	0,39	0,41
Refus	0,37	0,37
Autres cas non productifs	0,52	0,58

Limites de confiance à 95 % généralement de ± 0,02

Nous constatons qu'une variable de « consentement » — le répondant indique, à la vague 1, qu'il est d'accord que les enregistrements de l'enquête soient couplés aux données sur la santé provenant de sources administratives — est également associée à la non-réponse à la vague 2, quoique cette variable n'ait pas été utilisée dans les travaux antérieurs. Son inclusion dans le modèle n'a toutefois que très peu d'effet sur le coefficient de Gini, mais la valeur des pentes des courbes logit des rangs est plus élevée. Cette constatation fournit une certaine reconfirmation que les scores de risque pourraient être relativement robustes à certaines erreurs de spécification des régressions logistiques.

Il est possible d'utiliser les variables mesurées à la vague t+1 pour prédire la non-réponse à la vague t. Nous constatons que le changement de type de logement et de situation de partenariat entre les vagues t-1 et t+1, et le revenu familial à la vague t+1 prédisent tous la non-réponse à une vague, le coefficient de Gini passant de 0,43 à 0,46. Ce résultat a des incidences en ce qui concerne les méthodes d'ajustement, tel que nous en discutons plus loin.

4. Diverses stratégies pour prédire la non-réponse

Une des difficultés auxquelles se heurtent les analystes qui veulent utiliser les pondérations pour corriger la non-réponse est due au fait qu'idéalement, les poids doivent être réestimés à chaque vague. Cependant, la recherche de nouveaux prédicteurs de la non-réponse, puis le recalcul des poids et leur transmission aux analystes secondaires peut prendre du temps et représenter éventuellement un gaspillage de ressources techniques rares. Par conséquent, cela vaut la peine de voir si les changements de pondération après la vague 2 sont suffisamment importants pour justifier le recalcul des poids à chaque vague. Ici, nous nous concentrons sur la quatrième vague de la MCS et comparons les coefficients de Gini pour trois modèles pour la non-réponse globale fondés sur :

1. les variables de la vague 1, les valeurs de ces variables à la vague 1 et les coefficients à la vague 1;
2. les variables de la vague 1, les valeurs à la vague 1 et les coefficients à la vague 3;
3. les valeurs des variables à la vague 1, leurs valeurs à la vague 3 et les coefficients de la vague 3.

Nous constatons que quatre seulement des 15 variables de la vague 1 qui était prédictives de la non-réponse à la vague 2 ne sont pas prédictives de la non-réponse à la vague 4, et que les coefficients de Gini estimés pour les trois modèles valent 0,36 (n = 17 862), 0,37 (n = 17 862) et 0,36 (n = 12 729), valeurs à peine plus faibles que l'estimation de 0,39 à la vague 2. Il convient aussi de noter que la non-réponse partielle entraîne une réduction considérable de la taille de l'échantillon dans le cas de la troisième stratégie, ce qui constitue une difficulté générale quand on construit les poids de non-réponse.

D'autres travaux porteront sur le pouvoir de discrimination et la prédiction à la quatrième vague en incluant dans le modèle les autres variables de la troisième vague.

5. Conclusion

Trois remarques importantes se dégagent de la présente étude. La première est que l'utilisation d'un cadre construit en s'appuyant sur divers types de probabilités conditionnelles et de scores de risque produit des mesures sommaires de l'exactitude, telles que les coefficients de Gini et les pentes des courbes logit des rangs, qui nous permettent de comparer entre eux divers modèles de prédiction de la non-réponse. Ces comparaisons offrent un moyen d'évaluer l'utilité de l'ajout de prédicteurs supplémentaires dans les modèles et de comparer le pouvoir de prédiction et de discrimination pour divers types de non-réponse. Nous constatons, du moins dans le cas de la Millennium Cohort Study examiné ici, que nos meilleurs modèles de prédiction des données manquantes ne sont pas particulièrement exacts et ce manque d'exactitude laisse entendre que nous pourrions avoir de la difficulté à cibler efficacement les interventions susceptibles de prévenir la non-réponse. La question du ciblage soulève des questions quant au seuil d'exclusion optimal appliqué au score de risque qui, à son tour, requiert un examen des coûts et des avantages de l'intervention qui dépasse le cadre du présent article.

La deuxième remarque est que les modèles élaborés pour produire les poids de non-réponse à la vague 2 pourraient être utilisés de manière satisfaisante aux vagues ultérieures. Si cette hypothèse était corroborée par des études plus approfondies, en particulier l'étude de modèles comprenant des variables mesurées pour la première fois après la vague 1, par exemple les résultats des enfants à des tests éducatifs administrés à la maison, cela signifierait peut-être que les efforts en vue d'estimer de nouveaux modèles de non-réponse à chaque vague risque d'être peu judicieux et permettait aussi de contourner certains problèmes créés par la non-réponse partielle dans les modèles utilisés pour produire les poids correspondant aux probabilités inverses.

La dernière remarque a trait aux incidences de nos résultats en ce qui concerne l'ajustement statistique par d'autres moyens que l'utilisation des probabilités inverses comme pondérations. Nous constatons que nos modèles de prédiction des données manquantes sont, pour les non-répondants à une vague, améliorés lorsque l'on ajoute des variables provenant d'une vague ultérieure et que ces variables pourraient donc être incluses dans un processus d'imputation choisi pour un modèle d'intérêt particulier. En outre, le fait que certains prédicteurs importants de la non-réponse sont des variables dont l'inclusion dans un modèle d'intérêt est fort peu probable — par exemple, fournir une adresse stable — signifie que ces prédictions pourraient être utilisés comme variables instrumentales dans un modèle conjoint de type Heckman, dans lequel sont considérés simultanément les modèles d'intérêt et le mécanisme de formation des données manquantes, donc tenant compte de la non-réponse non ignorable. Carpenter et Plewis (2010) fournissent un exemple.

Bibliographie

Carpenter, J. et Plewis, I. (2010). Analysing Longitudinal Studies with Non-Response: Issues and Statistical Methods, in M. Williams and P. Vogt (eds.), *Handbook of Methodological Innovations*. Newbury Park, Ca.: Sage.

Copas, J. (1999). The Effectiveness of Risk Scores: The Logit Rank Plot, *Applied Statistics*, 48, pp. 165-183.

Couper, M. P. et Lyberg, L. E. (2005). The Use of Paradata in Survey Research, paper presented at the 54th Session of the International Statistical Institute, Sydney, Australia.

Krzanowski, W. J. et Hand, D. J. (2009). *ROC Curves for Continuous Data*, Boca Raton, Fl.: Chapman and Hall/CRC.

Pepe, M. S. (2003). **The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford: OUP.***

Plewis, I. (ed.) (2007a). *The Millennium Cohort Study: Technical Report on Sampling* (4th. ed.), London: Institute of Education, University of London.

Plewis, I. (2007b). Non-response in a Birth Cohort Study: The Case of the Millennium Cohort Study, *International Journal of Social Research Methodology*, 10, pp. 325-334.

Plewis, I., Ketende, S. C., Joshi, H. et Hughes, G. (2008). The Contribution of Residential Mobility to Sample Loss in a Birth Cohort Study: Evidence from the First Two Waves of the Millennium Cohort Study, *Journal of Official Statistics*, 24, pp. 365-385.

Étude de l'érosion dans l'Étude longitudinale sur le développement des enfants du Québec de 1998 à 2008

Catherine Fontaine et Robert Courtemanche[1]

Résumé

L'*Étude longitudinale sur le développement des enfants du Québec* (ÉLDEQ) a débuté en 1998. Après 10 années de collecte de données, on retrouve 1 402 répondants comparativement aux 2 120 répondants du premier volet. En vue d'une éventuelle prolongation de l'ÉLDEQ à partir de l'année 2011, il s'avère nécessaire d'évaluer les limites imposées par l'érosion de l'échantillon telles que le biais longitudinal des estimations notamment. L'analyse tient compte de la correction de non-réponse par le biais des pondérations. Enfin, quelques solutions sont présentées pour atténuer les effets de l'érosion.

Mots Clés : Érosion, non-réponse, biais.

1. Introduction

1.1 Présentation générale de l'enquête et de ses objectifs

L'Étude longitudinale sur le développement des enfants du Québec a débuté en 1998. Elle vise à représenter les enfants nés au Québec en 1997-1998 (naissances simples seulement). Un échantillon de 2 817 enfants a été sélectionné à partir du Fichier maître des naissances vivantes du Québec, selon un plan de sondage stratifié à trois degrés. Aucun ajout d'échantillon n'a été effectué depuis 1998. L'objectif principal de l'enquête est d'identifier les facteurs qui, mis en place pendant la petite enfance, contribuent à l'adaptation sociale et à la réussite scolaire des petits Québécois. Des collectes annuelles ont eu lieu de 1998 à 2006, puis bisannuelles depuis 2006. Plusieurs instruments de collecte sont utilisés à chaque collecte, dont le Questionnaire informatisé rempli par l'interviewer (QIRI) qui est administré à la personne qui connaît le mieux l'enfant (PCM). Des mesures sur l'enfant sont également prises, par le biais de tests cognitifs, d'évaluation de la condition physique ou par le Questionnaire papier administré à l'enfant (QPAE). Une pondération dite « transversale » est calculée à chaque volet pour représenter l'ensemble des enfants nés au Québec en 1997-1998 et qui ont complété au moins un instrument de collecte à ce volet[2]. D'autres pondérations s'ajoutent selon les volets, telles une pondération longitudinale et des pondérations pour des instruments de collecte secondaires. Jusqu'au volet 2006, la création des classes de pondération s'effectuait par segmentation (selon l'algorithme CHAID). Lorsque les classes étaient créées, la proportion pondérée de répondants était calculée à partir d'un volet de référence (antérieur au volet à pondérer). Les volets 1998, 2000 et 2002 ont été utilisés comme volets de référence pour les différentes pondérations.

1.2 Description de l'étude

L'étude de l'érosion s'inscrit dans un processus de réévaluation pour la poursuite de l'enquête. En effet, l'enquête avait été prévue initialement pour une seule phase (1998 à 2002). Une deuxième phase s'est ajoutée de 2003 à 2010, soit durant les années de fréquentation de l'enfant au niveau primaire (maternelle à la 6e année). Il y a présentement une volonté de poursuivre l'enquête pour une troisième phase, de 2011 à 2015, soit durant les années de fréquentation de l'enfant au niveau secondaire (secondaire 1 à 5). Cependant, comme l'échantillon a subi une érosion entre 1998 et 2008, il est justifié d'évaluer les limites qu'impose cette érosion sur les analyses produites, de même que les solutions à implanter lors de la troisième phase.

2. L'érosion dans l'ÉLDEQ

[1]Catherine Fontaine et Robert Courtemanche, Institut de la statistique du Québec, 200, chemin Sainte-Foy, Québec (Québec), G1R 5T4, Canada
(catherine.fontaine@stat.gouv.qc.ca et robert.courtemanche@stat.gouv.qc.ca).

[2] Sont exclus les enfants décédés ainsi que ceux dont les familles ont quitté le Québec définitivement.

L'érosion est définie généralement comme la perte prématurée des unités ayant été sélectionnées pour un suivi longitudinal, par rapport au sous-ensemble des unités ayant répondu au temps 1. Dans l'ÉLDEQ, l'érosion que l'on peut qualifier de « définitive » comprend les enfants dont la famille quitte le Québec définitivement, les décès et les refus chroniques de la part des parents à participer à l'enquête. Par contre, il peut ne pas être possible de déterminer certains cas d'érosion définitive avant la fin de l'enquête. En particulier, pour les cas où une famille ne répond pas à un volet d'enquête, mais est recontactée à chaque volet, et répond à un volet ultérieur. On peut considérer ces cas comme de l'érosion « temporaire ». Et c'est effectivement ce qui est préconisé par l'ÉLDEQ : la plupart des non-répondants d'une vague antérieure se voient contactés lors d'une vague ultérieure.

Le tableau suivant indique l'érosion à partir de l'échantillon de répondants en 1998.

Tableau 2.1-1
Érosion dans l'ÉLDEQ de 1998 à 2008

Volet	Âge des enfants	Nombre de répondants	Érosion (%)
1998	5 mois	2 120	0,0
1999	17 mois	2 045	3,5
2000	2 ½ ans	1 997	5,8
2001	3 ½ ans	1 950	8,0
2002	4 ans	1 944	8,3
2003	5 ans	1 759	17,0
2004	6 ans (maternelle)	1 492	29,6
2005	7 ans	1 528	27,9
2006	8 ans	1 526	28,0
2008	10 ans	1 402	33,9

Note : Ce tableau indique le nombre de répondants visés par la pondération transversale.

L'érosion qui fait l'objet d'une analyse est définie comme la non-réponse à un volet d'enquête, pour les volets de 1999 à 2008, par rapport aux enfants ayant participé à l'enquête en 1998. Cette définition inclut donc des cas d'érosion dite définitive et temporaire. Elle correspond aussi à la notion de non-réponse pour laquelle un traitement est effectué à chaque volet par le biais des pondérations. Ce traitement fait gonfler les poids des répondants d'enquête à un volet donné pour compenser la perte des non-répondants, par rapport à un volet antérieur. Il est à noter que les poids ne permettent de représenter que les enfants admissibles à l'enquête à un volet donné (les cas de déménagements hors Québec définitifs et les décès sont exclus).

L'estimation du biais longitudinal peut être définie comme étant l'écart entre une statistique mesurée pour une variable d'intérêt dans la population et l'estimation pondérée de cette statistique dans l'échantillon à un temps *t*. Par contre, cette définition implique que la statistique peut être calculée dans la population. Au moment où cette analyse a été réalisée, une telle mesure n'était pas disponible. C'est donc la notion de cohérence longitudinale qui s'est imposée pour remplacer la notion de biais longitudinal[3]. La cohérence longitudinale peut être estimée en calculant l'écart entre l'estimation pondérée d'une caractéristique au temps 1 (lorsque la taille de l'échantillon est la plus grande) et la même estimation à un temps ultérieur (temps *t*).

3. Méthodes d'analyse

3.1 Choix des caractéristiques à l'étude

Trois méthodes ont été utilisées pour effectuer cette analyse de l'érosion. Les deux premières méthodes concernent l'aspect de la cohérence longitudinale et la troisième concerne l'aspect relié à la précision. Elles seront décrites en détail aux sections ci-dessous. Pour ces méthodes, le même sous-ensemble de neuf variables, mesurées au volet 1998, a été utilisé. Ces variables ont été retenues principalement parce qu'elles sont liées aux mesures de réussite scolaire. Voici la liste de ces neuf caractéristiques :

1) revenus du ménage en 1998 (moins de 15 000 $);
2) santé du bébé en 1998 perçue par la PCM (bonne, mauvaise ou passable);

[3] On peut lire une définition de la notion de cohérence longitudinale dans Franklin et al. (2007).

3) niveau de suffisance du revenu du ménage en 1998 (insuffisants ou très insuffisants);
4) langue parlée à la maison par la mère en 1998 (ni français, ni anglais);
5) type de famille (monoparentale);
6) statut d'immigrant de la mère (immigrante européenne ou non européenne);
7) plus haut niveau de scolarité de la mère en 1998 (diplôme d'études secondaires ou moins);
8) région sociosanitaire de résidence au moment de l'échantillonnage (Montréal ou Laval);
9) âge de la mère en 1998 (moins de 25 ans).

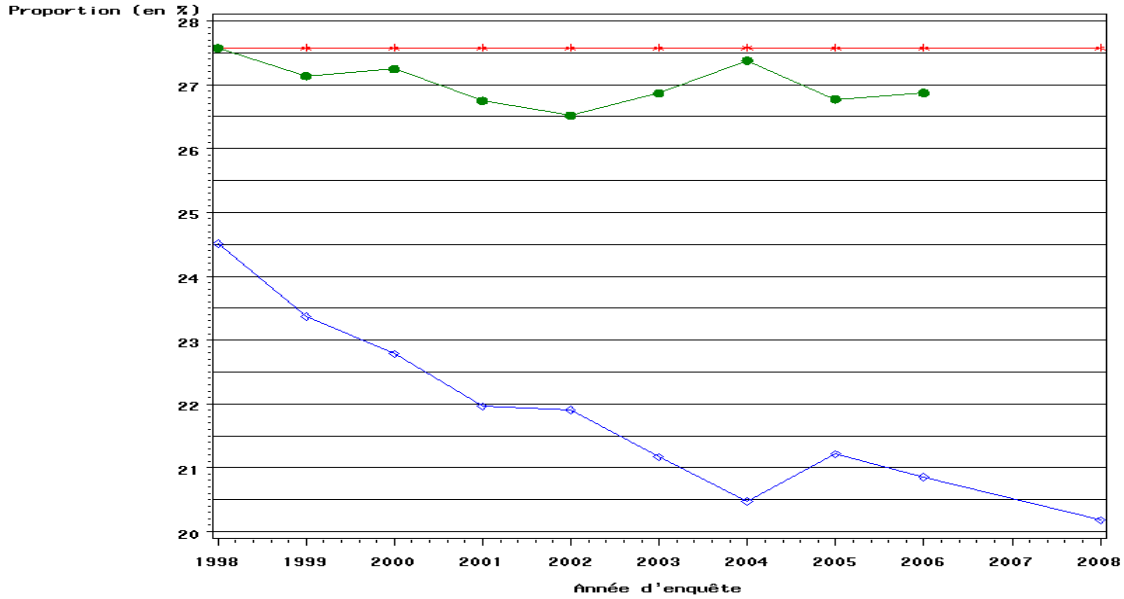
3.2 Analyse de la cohérence longitudinale

3.2.1 Graphiques

Pour illustrer visuellement la cohérence longitudinale, des graphiques ont été produits pour chacune des neuf caractéristiques à l'étude. Essentiellement, ces graphiques montrent l'évolution de l'estimation d'une caractéristique donnée, de 1998 à 2006. Deux courbes sont présentées, selon le différent poids de sondage qui est utilisé dans l'estimation (avec ou sans correction de la non-réponse). Ces courbes peuvent être comparées à une droite de référence, celle-ci étant l'estimation de la caractéristique au temps 1. Ces graphiques ont montré que les estimations avec une correction de la non-réponse permettent d'atténuer l'effet dû aux pertes d'échantillon. À l'inverse, les estimations sans correction présentent, généralement, une pente descendante lorsque les années avancent. L'analyse des graphiques révèle également que deux caractéristiques semblent plus problématiques quant au niveau de cohérence longitudinale (voir section 4 pour plus de détails).

Un seul graphique est présenté à titre d'exemple. C'est celui illustrant le niveau de suffisance du revenu du ménage. La courbe des estimations avec correction de la non-réponse[4] ne suit pas la même tendance que celle pour les estimations sans correction de la non-réponse[5]. Elle demeure plutôt près de l'estimation de référence[6] (volet 1998) même si la courbe des estimations sans correction s'en éloigne de plus en plus. Cela indique que la pondération a bien effectué son travail.

Figure 3.2.1-1
Graphique illustrant la cohérence longitudinale



3.2.2 Analyse des différences relatives

Bien qu'utiles, les graphiques ne permettent pas de saisir l'ampleur de la cohérence longitudinale, en raison de l'échelle qui diffère entre eux. C'est pourquoi une mesure de différence relative entre chacune des estimations pondérées d'un volet donné et

[4] Celle illustrée par le signe « ● ».
[5] Celle illustrée par le signe « ◊ ».
[6] Celle illustrée par le signe « + ».

l'estimation pondérée du volet 1998 a été calculée, pour les volets 1999 à 2006 (la pondération des données du volet 2008 n'avait pas encore été effectuée au moment de l'analyse). Les estimations pondérées incluent la correction de la non-réponse.

Formule 1 : Mesure de différence relative

$$Différence\ relative_t = \frac{\left(\hat{p}_{transversale\ pondérée\ volet\ t} - \hat{p}_{transversale\ pondérée\ volet\ 1998}\right)}{\hat{p}_{transversale\ pondérée\ volet\ 1998}}$$

pour les temps t=1999 à 2006.

Le tableau 3.2.2-1 présente, pour chacune des neuf caractéristiques, le maximum, en valeur absolue, parmi les différences relatives calculées de l'année 1999 à 2006.

Tableau 3.2.2-1
Maximum, en valeur absolue, de la différence relative par caractéristique

Caractéristique de 1998	Maximum (%)
Langue parlée à la maison par la mère	18,3
Statut d'immigrant de la mère	14,8
Type de famille	6,7
Région sociosanitaire	6,2
Revenus du ménage	5,9
Plus haut niveau de scolarité de la mère	5,8
Niveau de suffisance du revenu	3,8
Âge de la mère	3,1
Santé du bébé	2,5

3.3 Analyse de la précision

Le troisième aspect de l'analyse de l'érosion concerne la précision et la puissance Il sera traité simplement en présentant le nombre de répondants qui présentaient une caractéristique donnée en 1998, et qui étaient toujours répondants en 2006 ou 2008. Le tableau suivant présente les pertes (non pondérées) pour les caractéristiques à l'étude.

Tableau 3.3-1
Évolution du nombre de répondants entre 1998 et 2008

Caractéristique de 1998	Volet 1998	Volet 2006	Volet 2008	Perte entre 1998 et 2008 (non pondérée)
Langue parlée à la maison par la mère	128	55	53	59 %
Statut d'immigrant de la mère	253	128	120	53 %
Revenu du ménage	248	140	121	51 %
Type de famille	171	100	86	50 %
Niveau de suffisance du revenu	511	316	284	44 %
Plus haut niveau de scolarité de la mère	564	370	327	42 %
Région sociosanitaire	571	371	338	41 %
Âge de la mère	159	109	96	40 %
Santé du bébé	473	330	292	38 %
Nombre de répondants au volet	2 120	1 526	1 402	34 %

Note : Ce tableau indique le nombre de répondants visés par la pondération transversale.

4. Constats et solutions proposées

Les deux variables où le niveau de cohérence longitudinale est le moins bon, parmi les variables étudiées, sont habituellement des caractéristiques qui permettent de définir le modèle de non-réponse. Ce sont les variables du statut d'immigrant de la mère et de la langue parlée par la mère à la maison. C'est donc dire que la probabilité de répondre varie selon ces caractéristiques et que l'ajustement qui est fait permet de diminuer le biais de non-réponse pour ces caractéristiques (et celles qui leur sont reliées). Par contre, il semble que même ces corrections ne permettent pas de rétablir un bon niveau de cohérence longitudinale. Le tableau 3.3-1 permet de constater que ces deux caractéristiques ont subi de grandes pertes depuis 1998, en proportion, relativement aux autres variables. Il est ainsi plus difficile de minimiser les biais dans ce cas. En outre, la méthode de

segmentation de l'échantillon de répondants (qui permet de créer les groupes de pondération) peut avoir utilisé ces deux variables pour définir le modèle de non-réponse pour un sous-groupe d'enfants seulement.

Les deux variables qui présentent les pertes de répondants les plus faibles (environ 40 %) avaient également un très bon niveau de cohérence longitudinale. Ce sont les variables représentant les enfants dont la mère avait moins de 25 ans et représentant les enfants dont la santé perçue du bébé était bonne, passable ou mauvaise en 1998. L'une de ces variables est considérée lors de la modélisation de la non-réponse (âge de la mère), l'autre ne l'est pas.

D'autres constats sont également énoncés dans Fontaine et Courtemanche (2009). Ainsi, toutes ces analyses révèlent que les caractéristiques de 1998 qui sont davantage touchées par l'érosion dans l'ÉLDEQ sont le statut d'immigrant de la mère; les revenus du ménage; la langue parlée par la mère à la maison et le type de famille. Si elles sont analysées et qu'elles sont également liées à la réussite scolaire, il y a un risque accru de biais longitudinal. Et ce, en dépit du fait que ce sont des caractéristiques qui sont habituellement prises en compte lors de l'ajustement pour la non-réponse effectué pour les différents instruments/volets. Enfin, elles sont généralement associées à de faibles nombres de répondants en 2008, ce qui peut entraîner des restrictions au niveau de l'analyse. Quelles sont alors les solutions pour corriger cette situation dans le futur?

En ce qui concerne la solution proposée pour améliorer l'aspect de la cohérence longitudinale, elle consiste en des travaux méthodologiques, effectués afin de considérer une méthode de correction de la non-réponse alternative à celle utilisée actuellement. Ces deux méthodes fonctionnent en corrigeant le poids de référence d'un volet antérieur à l'aide de l'inverse du taux de réponse pondéré pour le volet actuel, à l'intérieur d'un sous-ensemble d'enfants : ce sont les groupes de pondération. La méthode actuelle utilise une segmentation de l'ensemble de données basée sur des tests d'association (khi-deux). La méthode alternative considérée, appelée la méthode du score, crée des groupes selon les valeurs d'un score, celui-ci étant la probabilité estimée de réponse issue d'un modèle de régression logistique. En effet, des études récentes laissent supposer les bienfaits de la méthode du score (Haziza et Beaumont, 2007). Le résultat de ces travaux méthodologiques indique que les quatre caractéristiques qui avaient été ciblées comme étant le plus associées à l'érosion voient leur niveau de cohérence longitudinale s'améliorer avec l'utilisation de la méthode du score. La méthode du score a donc été implantée lors de la pondération des données du volet 2008.

Quant à l'aspect portant sur la précision des estimations, les solutions proposées visent à restreindre la diminution du nombre de répondants, de plus en plus marquée au cours des volets les plus récents. La fidélisation des répondants, la réduction du fardeau de réponse (en limitant le nombre d'instruments de collecte) et l'utilisation de données administratives (pour l'ensemble de l'échantillon) sont autant de mesures susceptibles d'améliorer la puissance des analyses.

5. Conclusion

En conclusion, l'étude de l'érosion pour l'ÉLDEQ jusqu'à 2008 a permis de confirmer que le sous-ensemble des enfants touchés par cette perte proviennent de familles ayant l'une (ou plusieurs) des caractéristiques suivantes : revenus du ménage faibles; monoparentalité; mères immigrantes; mères qui ne parlent ni français, ni anglais à la maison. Ce sous-ensemble représente un domaine d'intérêt dans l'étude de la réussite scolaire.

Par contre, cette analyse est compromise par le risque de biais potentiel et le nombre d'enfants disponibles pour la réaliser. Ainsi, un changement dans la méthode de création des classes de pondération a été implanté lors de la pondération des données de l'année 2008 pour tenter de réduire les effets de l'érosion sur le biais longitudinal. Il sera également possible de cibler ce sous-groupe d'enfants lors de la collecte de données pour minimiser la non-réponse future. Cela a déjà été effectué lors de collectes précédentes.

Cette étude de l'érosion a donc démontré l'importance de la pondération dans la minimisation des biais dus à l'érosion, puisque la pondération améliore le niveau de cohérence longitudinale de caractéristiques reliées à la réussite scolaire. Ce constat a d'ailleurs été confirmé par une analyse complémentaire à partir de données populationnelles, provenant du ministère de l'Éducation, des Sports et des Loisirs du Québec (MELS).

En terminant, les utilisateurs de données de l'ÉLDEQ ont été informés des conclusions émanant de l'étude de l'érosion. Cela leur permettra d'évaluer dans quelle mesure celles-ci peuvent limiter les conclusions de leurs analyses futures.

Bibliographie

Fontaine, C. et R. Courtemanche, R. (2009). Étude de l'érosion pour l'Étude longitudinale sur le développement des enfants du Québec de 1998 à 2008, rapport non publié, Québec, Canada : Institut de la statistique du Québec.

Franklin, S. et al. (2007). The National Longitudinal Survey of Children: 10-Year Review of the Methodology, rapport non publié, Ottawa, Canada: Statistique Canada.

Haziza, D. et J.-F. Beaumont, J.-F. (2007). On the construction of imputation classes in surveys, *International Statistical Review*, 75, p. 25-43.

Modélisation de la non-réponse pour une enquête longitudinale à l'aide de paradonnées : Application à l'enquête sur la dynamique du travail et du revenu

Beatrice Baribeau, Wisner Jocelyn, et Junior Chuang[1]

Résumé

L'Enquête sur la dynamique du travail et du revenu (EDTR) est une enquête longitudinale conçue pour recueillir des données sur le travail, le revenu et les circonstances familiales en vue de produire des estimations longitudinales et transversales. À l'heure actuelle, des renseignements figurant dans la base de sondage, provenant de l'Enquête sur la population (EPA), sont utilisés pour produire un ajustement pour la non-réponse en se servant de l'algorithme CHAID (Chi-Square Automatic Interaction Detection). Dans le présent article, nous examinons les ajustements pour la non-réponse en se servant de paradonnées par comparaison à la méthode courante. Selon les résultats de l'étude, bien que la méthode courante d'ajustement pour la non-réponse produise des erreurs-types plus faibles pour certaines estimations clés, le modèle comprenant à la fois des paradonnées et des données tirées de la base de sondage offrent des avantages sous forme d'une corrélation plus forte avec la réponse et des variables autres que le revenu, ainsi qu'une réduction du speudo-biais relatif.

Mots clés : paradonnées, ajustement pour la non-réponse, pseudo-biais relatif.

1. Introduction

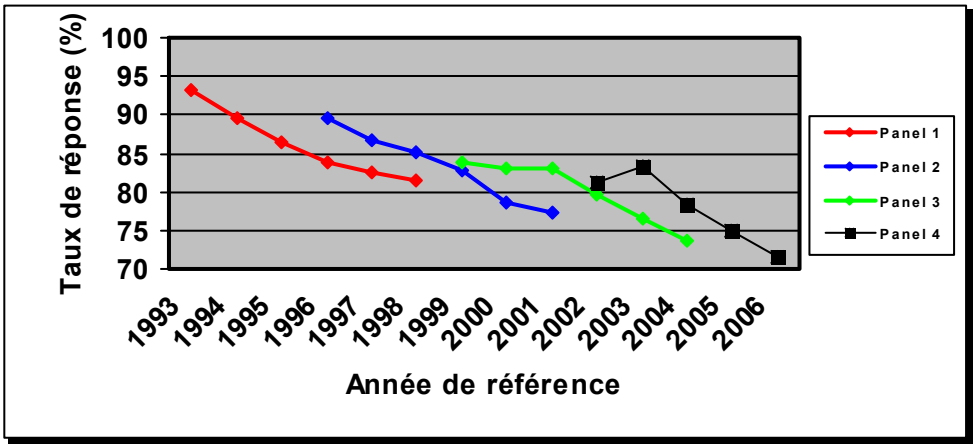
1.1 Introduction à l'Enquête sur la dynamique du travail et du revenu

L'Enquête sur la dynamique du travail et du revenu (EDTR) est une enquête téléphonique, longitudinale, annuelle conçue pour recueillir des données au sujet du travail, du revenu et des circonstances familiales en vue de produire des estimations longitudinales et transversales. Elle consiste en un suivi longitudinal d'un sous-échantillon de personnes provenant d'environ 17 000 ménages répondant à l'Enquête sur la population active (EPA). Chaque panel de l'EDTR est suivi pendant six ans, un nouvel panel étant ajouté tous les trois ans. Puisque l'échantillon est tiré des ménages répondant à l'EPA, les renseignements démographiques de base sur les personnes qui participent à l'EDTR sont directement disponibles. Les ménages pour lesquels est enregistré un refus catégorique ou deux années consécutives de non-réponse sont éliminés de la collecte. Cet aspect du processus de collecte des données influe sur l'utilisation des paradonnées pour la modélisation de la non-réponse, étant donné que ces paradonnées ne sont disponibles que pour les ménages ayant participé à la collecte.

La figure 1.1.-1 donne les taux de réponse longitudinaux selon la vague et le panel. L'année indiquée au bas de la figure représente l'année de référence de l'enquête. L'année de la collecte des données est supérieure d'une unité à l'année de référence. La figure révèle clairement une diminution des taux de réponse au cours de l'existence du panel, ainsi que d'un panel au suivant. Cette baisse est un facteur qui motive la recherche d'un ajustement optimal pour la non-réponse. L'adoption d'un plan de collecte adaptatif pour l'année de référence 2009 et une refonte de l'enquête comportant un panel à durée de vie indéfinie sont d'autres facteurs dont il faut tenir compte dans l'analyse de l'ajustement pour la non-réponse. Le plan de collecte adaptatif s'appuie sur les paradonnées pour concentrer les efforts de collecte de manière à accroître les taux de réponse et à s'assurer de la représentativité de l'échantillon. L'utilisation d'un ajustement pour la non-réponse en s'appuyant sur les mêmes variables complète bien cette méthode. La refonte de l'enquête qui pourrait inclure un panel de durée indéfinie signifie qu'au fur et à mesure du vieillissement du panel, l'information de la base de sondage deviendra de plus en plus dépassée. Le panel utilisé dans les présents travaux de recherche était le panel 4, qui est le dernier panel démontré. Nous l'avons choisi en raison des données provenant de plusieurs vagues disponibles pour l'analyse et de sa pertinence, puisqu'il s'agit d'un panel récent.

[1] Beatrice Baribeau, Statistique Canada, 100 Promenade Tunney's Pasture, Ottawa, Canada, K1A 0T6, Beatrice.Baribeau@statcan.gc.ca; Wisner Jocelyn, Statistique Canada, 100 promenade Tunney's Pasture, Ottawa, Canada, K1A 0T6, Wisner.Jocelyn@statcan.gc.ca; Junior Chuang, Simon Fraser University, Canada

Figure 1.1-1
Taux de réponse longitudinal selon la vague et le panel



1.2 Méthode courante d'ajustement pour la non-réponse

À l'heure actuelle, les données de la base de sondage, qui proviennent de l'Enquête sur la population active (EPA), sont utilisées pour produire des variables dichotomiques en vue de modéliser la non-réponse. L'ajustement pour la non-réponse est basé sur l'algorithme CHAID (Chi-Square Automatic Interaction Detection) qui retient les variables les plus significatives pour expliquer la réponse. En se basant sur ces variables, l'algorithme construit un arbre dont les feuilles correspondent aux groupes de réponses homogènes (GRH). Les GRH sont contraints de contenir au moins 30 unités, avec un taux de réponse pondéré minimal de 50 %. Les poids des répondants sont corrigés en les multipliant par l'inverse du taux de réponse pondéré, tandis qu'un poids nul est attribué aux non-répondants. L'un des avantages de l'utilisation de l'algorithme CHAID est que, dans la construction de l'arbre, il tient compte de la dissymétrie et est donc moins contraignant que les méthodes de mise en grappes (voir Beaumont, 2005) ou les méthodes basées sur la propension à répondre (Watson et Wooden, 2006).

1.3 Paradonnées disponibles

L'EDTR est une enquête téléphonique assistée par ordinateur. Les paradonnées disponibles [voir Scheuren, 2005 pour une discussion intéressante des concepts des paradonnées] proviennent des fichiers de l'historique des transactions Blaire (BTH pour *Blaise Transaction History*). Les fichiers BTH contiennent des renseignements tels qu'un enregistrement pour chaque appel fait à un ménage, l'heure et le jour de l'appel et le résultat de l'appel (c'est-à-dire signal occupé, interview obtenue, etc.). L'information est consignée au niveau du ménage, c'est-à-dire que les appels représentent les appels faits au ménage, non à des membres particuliers de celui-ci. Les ménages peuvent compter plusieurs enregistrements, chacun représentant un appel téléphonique individuel. Puisque les caractéristiques des appels, telles que le jour de l'appel, peuvent varier, seul l'appel le plus récent a été retenu pour chaque ménage. Le dernier appel est celui qui correspond le plus probablement aux résultats d'une réponse ou d'une non-réponse. Comme les renseignements tirés de la base de sondage, les renseignements provenant des paradonnées peuvent être convertis en variables dichotomiques telles que le « nombre d'appels faits au ménage » qui seront utilisées par l'algorithme CHAID.

2. Hypothèses de l'étude

2.1 Données du quatrième panel

Les données de cinq vagues du quatrième panel ont été utilisées pour la présente étude. Environ 1 860 (sur 34 000) unités ont été exclues de l'étude. La majorité des exclusions correspondaient à des personnes qui étaient des enfants à la première vague, mais étaient devenus des adultes à la cinquième vague (1 700 unités) et à des non-répondants pour lesquels n'existait aucune donnée BTH indiquant qu'il s'agissait de non-répondants (1 150 unités). Le chevauchement entre les deux groupes est important. Le premier groupe a été exclu à cause d'un manque de variables d'intérêt à la première vague, ce qui aurait produit une composante imprévisible du pseudo-biais relatif.

2.2 Imputation longitudinale des paradonnées

Pour les cas dont la non-réponse antérieure avait entraîné l'élimination de l'échantillon utilisé pour la collecte, aucune paradonnées n'existait. Dans ces circonstances, nous avons procédé à une imputation longitudinale de données issues de la

vague la plus récente de paradonnées, à condition que les données proviennent aussi d'une unité non répondante. L'information tirée d'une vague antérieure de non-réponse ne reflétait peut-être pas la situation dans la vague la plus récente de non-réponse, mais elle était plus actuelle que l'information provenant de la base de sondage. Dans certains cas, dans la dernière vague de paradonnées disponibles, l'unité était un répondant; ces enregistrements ont été éliminés de l'étude. L'une des limitations de la présente étude est que les unités non répondantes exclues pourraient différer des unités non répondantes incluses.

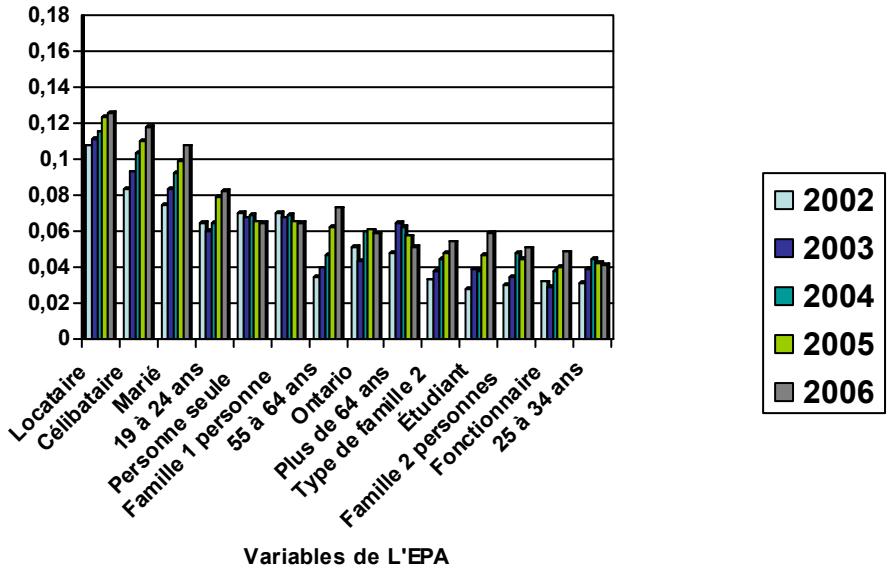
3. Analyse des ajustements pour la non-réponse

3.1 Corrélation entre les variables auxiliaires et la réponse

La formation des groupes de réponses homogènes (GRH) dépend de l'existence d'un lien entre les variables auxiliaires et la réponse; donc, il est souhaitable qu'il existe une forte corrélation entre la réponse et la variable auxiliaire. La figure 3.1-1 donne la valeur absolue des corrélations entre les variables auxiliaires tirées de l'EPA et la réponse pour les années de référence 2002 à 2006. Nous utilisons la valeur absolue de la corrélation pour montrer la grandeur de cette dernière, laquelle, pour la comparaison entre les variables auxiliaires, était plus intéressante que le sens de la corrélation.

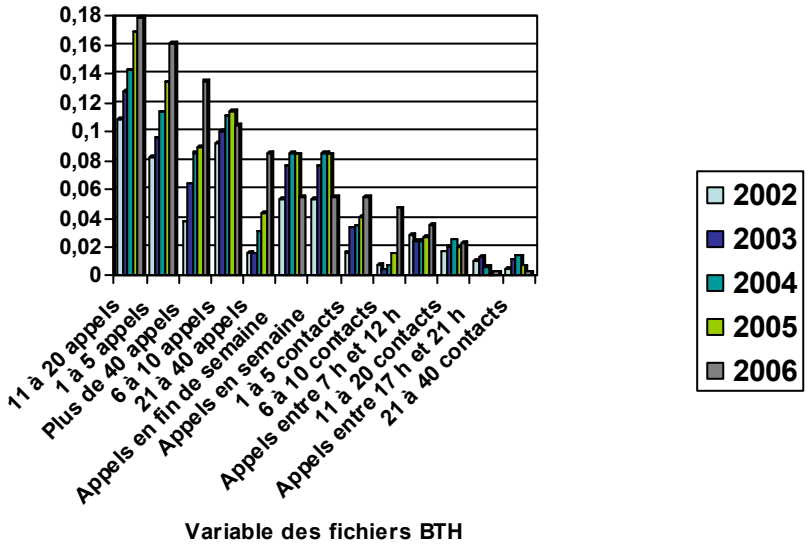
Comme le montre la figure 3.1-1, la corrélation augmente au cours de la vie du panel pour la plupart des variables. Les variables les plus corrélées à la réponse étaient les états de locataire, de célibataire et de personne mariée; par conséquent, elles représentaient les premières branches de l'arbre utilisées pour créer les GRH.

Figure 3.1-1
Valeur absolue des corrélations entre les variables de l'EPA et la réponse au cours de la vie du panel



La figure 3.1-2 donne la valeur absolue de la corrélation pour les variables auxiliaires tirées des paradonnées. Comme les variables de l'EPA, les variables des fichiers BTH sont devenues plus corrélées à la réponse au cours du temps. Cependant, trois variables BTH dépassent la corrélation la plus élevée des variables de l'EPA. La variable BTH la plus corrélée, « de 11 à 20 appels », avait une valeur absolue de corrélation de 0,18. Quand nous avons inclus les variables de l'EPA ainsi que les variables des fichiers BTH dans l'algorithme CHAID pour exécuter un ajustement de la non-réponse, les variables BTH étaient systématiquement sélectionnées pour former les premières branches de l'arbre.

Figure 3.1-2
Valeur absolue de la corrélation entre les variables des fichiers BTH et la réponse au cours de la vie du panel



3.2 Corrélation entre les variables auxiliaires clés et les variables d'intérêt

La deuxième mesure concernant les variables auxiliaires est la corrélation entre les variables auxiliaires clés et les variables d'intérêt. Il est souhaitable que cette corrélation soit forte, puisque cela donne lieu à une plus faible erreur-type associée aux estimations.

Dans l'ensemble, les variables auxiliaires de ni l'une ni l'autre source n'étaient fortement corrélées aux variables d'intérêt. Le tableau 3.2-1 montre que les variables de la base de sondage étaient plus corrélées au revenu que les variables des fichiers de paradonnées. Il en était de même pour les sources individuelles de revenu. Cependant les variables des fichiers de paradonnées avaient tendance à être plus corrélés aux autres variables que le revenu, comme l'illustre le tableau 3.2-2.

Tableau 3.2-1
Corrélation entre les variables auxiliaires et le revenu

Variable de l'EPA	Corrélation	Variable BTH	Corrélation
Célibataire	0,10	1 à 5 appels	0,05
Locataire	0,08	6 à 10 appels	0,01
Plus de 64 ans	0,06	11 à 20 appels	0,02
Ontario	0,04		

Tableau 3.2-2
Corrélation entre les variables auxiliaires et le nombre d'heures de travail

Variable de l'EPA	Corrélation	Variable BTH	Corrélation
Locataire	0,06	1 à 5 appels	0,14
Célibataire	0,02	6 à 10 appels	0,06
Ontario	0,02	11 à 20 appels	0,05
Plus de 64 ans	0,01		

3.3 Le pseudo-biais relatif

Sans les estimations de population réelles, il n'est pas possible de mesurer le biais. À l'instar de Singh et coll. (1995), nous avons utilisé une mesure de la cohérence longitudinale baptisée pseudo-biais relatif. Dans cette mesure, l'estimation à la première vague peut-être considérée comme l'approximation la meilleure possible de l'estimation de population réelle, une attrition ayant lieu aux vagues subséquentes. Dans le cas de l'attrition due à la non-réponse, il pourrait exister un biais de non-réponse si les non-répondants diffèrent des répondants, à moins qu'il soit tenu compte de cette différence dans l'ajustement pour la non-réponse. Si ce dernier produit une correction adéquate pour les non-répondants, l'utilisation des poids ajustés d'une vague ultérieure combinés à la valeur de la variable d'intérêt à la vague 1 devraient produire une estimation

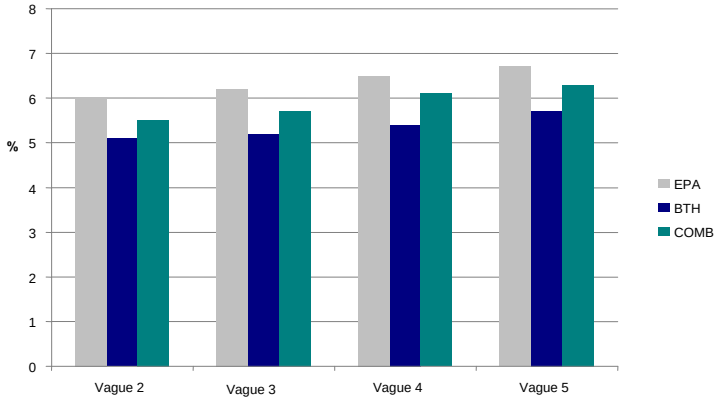
proche de celle de l'estimation à la vague 1. La mesure de la variation en pourcentage de la nouvelle estimation par rapport à l'estimation à la vague1 est égale au pseudo-biais relatif, ce qui s'exprime comme il suit :

$$Pseudo\text{-}biais\text{-}rel = \left(\frac{\hat{Y}_{1,S}^{(wi)} - \hat{Y}_{1,LFS}^{(w1)}}{\hat{Y}_{1,LFS}^{(w1)}} \right) * 100.$$

$\hat{Y}_1^{(wi)}$ est une estimation pour une variable d'intérêt particulière (revenu, salaire, loyer) et une source donnée S (EPA, BTH, EPA et BTH combinés) en utilisant le poids provenant de la vague *i (*wi*)* et les valeurs de *y* provenant de la première vague.

En utilisant le revenu pour mesurer le pseudo-biais relatif, la figure 3.3-1 montre le pseudo-biais relatif pour trois ajustements pour la non-réponse. Le premier ajustement est basé sur les variables de l'EPA, le deuxième uniquement sur les variables BTH, tandis que le troisième (dénommé COMB) est fondé sur une combinaison des deux sources. Comme l'attrition longitudinale augmente à chaque vague, il en est de même du pseudo-biais relatif. La figure 3.3-1 montre que plus l'ajustement dépend des paradonnées, plus le pseudo-biais relatif observé est faible. Cette tendance s'observe également pour plusieurs autres variables de revenu et des variables autres que le revenu.

Figure 3.3-1
Pseudo-biais relatif du revenu selon l'ajustement



3.4 Le coefficient interquartile

Nous avons examiné l'intervalle interquartile pour évaluer l'homogénéité des poids après l'ajustement pour la non-réponse. L'homogénéité des poids dans un GRH peut donner lieu à une plus grande stabilité de l'estimation des erreurs-types. Le coefficient interquartile est défini comme étant (Q3-Q1)/médiane. La figure 3.4-1 montre le coefficient interquartile moyen calculé sur l'ensemble des GRH pour les trois ajustements. Un faible coefficient interquartile est souhaitable. L'ajustement au moyen des variables de l'EPA seulement produit le coefficient interquartile le plus faible. Ce résultat est celui prévu, étant donné que les variables de l'EPA comprennent la province et la corrélation entre la province et la grandeur des poids.

Figure 3.4-1
Moyennes du coefficient interquartile

Ajustement	2002	2006
EPA	0,596	0,929
BTH	1,201	1,283
EPA et BTH combinés	1,083	0,736

3.5 Erreur-type pour le revenu total

Les analyses effectuées aux sections 3.3 et 3.4 s'appuient sur les poids obtenus après la correction pour la non-réponse. Nous avons examiné les erreurs-types des estimations du revenu après avoir effectué un ajustement pour tenir compte des valeurs influentes et du calage. Laroche (2007) discute en détails des divers types de poids utilisés pour l'EDTR. La plus forte corrélation des variables de l'EPA avec les variables d'intérêt, ainsi que le coefficient interquartile plus faible indiquent que les estimations pourraient être plus stables que si l'on utilise les poids basés sur les variables du fichier BTH seulement. Comme le montre le tableau 3.5-1, il en est effectivement ainsi. Ce tableau donne le coefficient de variation (CV) et la variation en pourcentage des erreurs-types (e.-t.) pour l'ajustement au moyen des variables BTH seulement et au moyen des variables BTH (paradonnées) et des variables de l'EPA combinées, selon le groupe d'âge. La seule variable pour laquelle la différence était

statistiquement significative est celle pour le groupe des 55 à 64 ans en 2001. L'examen des estimations selon la province a produit des résultats semblables. Alors que, dans l'ensemble, les erreurs-types des estimations augmentaient, les CV demeuraient faibles.

Tableau 3.5-1
Coefficient de variation et variation en pourcentage des erreurs-types pour le revenu selon l'âge

	BTH		EPA et BTH combinés	
Âge en 2001	CV	Variation en pourcentage de l'e.-t.	CV	Variation en pourcentage de l'e.-t.
Moins de 20 ans	3,1%	-1,6%	3,0%	-4,4%
20 à 24 ans	2,9%	-7,3%	2,9%	-7,5%
25 à 34 ans	2,0%	1,8%	2,1%	2,8%
35 à 44 ans	1,8%	0,7%	1,8%	0,8%
45 à 54 ans	2,5%	4,9%	2,5%	3,7%
55 à 64 ans	2,5%	2,3%	2,5%	-0,2%
65 ans et plus	2,0%	-2,4%	2,3%	12,3%
Tous âges confondus	3,1%	2,8%	3,0%	3,1%

4. Conclusion et futurs travaux

Les variables auxiliaires tirées des paradonnées étaient plus corrélées à la réponse et à certaines variables autres que le revenu que les variables tirées de la base de sondage. Les variables de la base de sondage étaient plus corrélées aux variables de revenu et produisaient un coefficient interquartile plus faible que les variables des paradonnées. La corrélation plus forte avec le revenu et le coefficient interquartile plus faible indiquaient une plus grande stabilité de la variabilité des estimations pour le revenu, ce qui a été confirmé par une plus faible erreur-type pour le revenu. Cependant, étant donné que les CV des estimations du revenu demeuraient faibles quand la correction pour la non-réponse s'appuyait uniquement sur les paradonnées et que cet ajustement réduisait le pseudo-biais relatif, le compromis entre la variabilité et le biais était acceptable. Si l'on pense à d'autres facteurs, tels que l'adoption d'un plan de collecte adaptatif et une refonte planifiée de l'enquête qui comprendra l'adoption d'un panel d'une durée indéfinie, un ajustement idéal pour la non-réponse pourrait intégrer à la fois les variables de la base de sondage et celles des paradonnées.

Les futurs travaux comprendront la production des estimations et des CV pour chaque année du quatrième panel. Avant de pouvoir intégrer un nouvel ajustement pour la non-réponse utilisant des paradonnées, il faudra proposer et tester un traitement pour les unités non répondantes exclues à l'heure actuelle. Il faudrait produire les estimations et les corrélations pour un plus grand nombre de variables autres que le revenu et calculer les pseudo-biais relatifs après le calage.

Bibliographie

Beaumont, J-F. (2005). L'utilisation de renseignements sur le processus de collecte des donnéespour traiter la non-réponse totale au moyen de l'ajustement de poids, *Techniques d'enquête*, 31, 249-254.

Laroche, S. (2007). Pondérations longitudinale et transversale de l'Enquête sur la dynamique du travail et du revenu, Catalogue no. 75F0002MIF — No. 007, Statistique Canada.

Scheuren, F. (2005). Les paradonnées de la conception à la réalisation, *Recueil du Symposium international 2005 de Statistique Canada.*

Singh, A.C., Wu, S. et Boyer, R. (1995). Longitudinal survey nonresponse adjustment by weight calibration for estimation of gross flows. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 396-401.

Watson, N. et Wooden, M. (2006), Modelling Longitudinal Survey Response: The Experience of the HILDA Survey, HILDA PROJECT DISCUSSION PAPER SERIES No.2/06, Australia.

SYMPOSIUM DE STATISTIQUE CANADA – BANQUET DU 25ÈME ANNIVERSAIRE

Symposium de Statistique Canada – Banquet du 25ème anniversaire

Gordon Brackstone

1. Introduction

Lorsqu'on m'a d'abord demandé de prendre la parole à l'occasion du symposium, je nous ai imaginés réunis dans une quelconque salle de bal sans fenêtres d'un hôtel, quelque part, mais non dans cette espèce de forêt pluviale de la côte Ouest avec mâts totémiques en arrière-plan et vue sur la majestueuse rivière des Outaouais. J'espère être à la hauteur de la situation et du décor.

On m'a demandé de parler du symposium du point de vue de quelqu'un qui a assisté au tout premier, puis à la vingtaine qui ont suivi. Je tiens à souligner que j'étais là au tout début, mais que je n'étais pas l'initiateur. La brève note de David Binder dans le programme explique comment le symposium a vu le jour grâce aux efforts déployés par le Comité de recherche en méthodologie. En outre, vous avez sûrement vu l'exposition et la vidéo présentées par les organisateurs au Palais des congrès, avec de nombreux souvenirs et fragments de symposiums tenus dans le passé et de nombreux visages familiers lorsqu'ils étaient beaucoup plus jeunes. Mon rôle à cette étape consistait essentiellement à donner mon consentement bienveillant. J'avais mis sur pied un comité de la recherche et du développement de la méthodologie et il s'agissait d'éviter de contrecarrer une bonne initiative et de trop m'immiscer. Grâce au Comité (initialement sous la direction de M.P. Singh et, plus tard, de David Binder), mon intervention a rarement été nécessaire et le plus souvent, de toute façon, on n'en tenait pas compte. Cependant, j'ai participé chaque année au choix du thème pour l'année suivante, pour garantir que le symposium ne soit pas axé sur l'analyse des données provenant d'enquêtes complexes année après année, comme certains le souhaiteraient peut-être, et que ses assises budgétaires demeurent solides.

Voyons d'abord ce que nous célébrons cette année. Les 25 ans du premier symposium qui s'est tenu en 1984. Les plus alertes parmi vous voudront savoir pourquoi on en célèbre les 25 ans en 2009 s'il s'en est tenu un chaque année depuis 1984. Ne vous en faites pas trop : 25 ou 26, c'est suffisamment proche pour les statistiques publiques. On a cessé de compter, semble-t-il, après que j'ai eu pris ma retraite. Ou bien, les règles de comptabilité ont été élaborées par un statisticien du troisième type. Ce qui est sûr et certain, c'est que cette année marque le 25e anniversaire du premier symposium.

J'aimerais d'abord dire quelques mots sur l'année 1984 et en particulier sur les circonstances et l'environnement ayant trait au lancement de la série de symposiums, c'est-à-dire la situation à Statistique Canada et dans le domaine des techniques d'enquête. Je ferai ensuite plusieurs commentaires et observations sur la série de symposiums au fil des ans et je terminerai par quelques réflexions sur l'avenir; et tout cela, avant qu'on vous serve le dessert.

2. 1984

Le livre sur l'année 1984 a déjà été écrit. Pour ceux d'entre nous qui avons grandi dans les années 1950 et 1960, 1984 a toujours été une année spéciale, longtemps attendue grâce au livre d'Orwell. Elle représentait un avenir éloigné, mais pas trop éloigné, dans lequel Big Brother surveillait les citoyens, et le ministère de la Vérité déterminait la réalité et réécrivait l'histoire en conséquence. Des concepts tels que novlangue, crimepensée et doublepensée faisaient partie du scénario. Orwell a écrit son livre en 1948, durant l'après-guerre où l'État a créé bien des programmes sociaux et où il a été appelé à jouer un rôle beaucoup plus grand dans la vie des citoyens qu'auparavant. En outre, Orwell avait été témoin des rôles joués par les gouvernements en Union Soviétique et en Allemagne avant la guerre.

Quand l'année 1984 est enfin arrivée, certaines des prédictions d'Orwell s'étaient réalisées et d'autres non. Par exemple, la novlangue était bien établie. StatCan est un exemple de novlangue (tout comme les noms astucieux et irrévérencieux se terminant par Can inventés pour les autres ministères). *FedProv* est un bon exemple de novlangue et de doublepensée en un seul concept. Son utilité tient à ce qu'il peut représenter, selon le cas, des approches communes harmonisées d'importantes questions nationales ou d'interminables chamailleries entre secteurs de compétence sans espoir de résolution. Nous n'avions pas un ministère de la Vérité, minivrai, mais nous avions Statistique Canada qui définissait la vérité, plus ou moins deux

erreurs-types, 19 fois sur 20. Et à Statistique Canada, le Système de comptabilité nationale réécrivait l'histoire sans aucun problème, publiant ouvertement des révisions historiques. Ce n'est que plus tard que l'assurance-chômage est devenue l'assurance-emploi.

Orwell avait long à dire sur la scène internationale et le rôle de la guerre comme condition nécessaire pour stimuler l'économie (il n'avait pas entendu parler de dépenses d'infrastructure); toutefois, il n'a pas su prédire la tenue d'un symposium sur des questions de méthodologie.

3. La situation à Statistique Canada

Ici dans le Pré, l'année 1984 a marqué une étape à Statistique Canada pour des raisons tout à fait autres. Une grande réorganisation cette année-là a réparti les structures de l'organisation en six secteurs qui demeurent pratiquement inchangés aujourd'hui. Ce changement reflétait la pensée transformatrice du statisticien en chef, Martin Wilk, qui avait résolu d'ouvrir l'organisme vers l'extérieur dans ses relations avec les clients et d'accroître à l'interne la collaboration et l'efficacité des activités. C'est ce qu'il a fait durant ses quatre brèves années à la barre. Martin Wilk croyait fermement en l'importance de la recherche-développement, de l'analyse et de la collaboration avec d'autres organismes pour atteindre les objectifs.

L'un des changements issus de la restructuration en 1984 a été le rétablissement d'une seule Direction de la méthodologie. De 1979 à 1984, les divisions de la méthodologie étaient éparpillées dans les directions spécialisées dans différents secteurs. Durant cette période de séparation (si j'ose utiliser ce mot ici au Québec), les directeurs de la méthodologie ont constaté le besoin pour leurs divisions d'entretenir des relations étroites et de mener des activités conjointes, comme avant 1979. En fait, le besoin de ces contacts et échanges est ressorti davantage durant cette période de séparation. L'une des initiatives prises dans ce contexte a été la création d'un comité de recherche en méthodologie chargé de coordonner à l'échelle des divisions les travaux de recherche-développement qui n'étaient par directement associés à certains programmes spécialisés. Comme David Binder l'a mentionné dans sa note, c'est ce comité qui a proposé la tenue du premier symposium sur des questions de méthodologie. Le comité a été maintenu lorsque la Direction de la méthodologie a été créée en 1984 et il existe encore aujourd'hui, son nom et ses fonctions ayant été quelque peu modifiés.

Il importe de signaler plusieurs autres institutions dans le domaine de la méthodologie qui datent de cette période. Le CCMS, le comité consultatif externe de Statistique Canada sur les questions de méthodologie, qui se réunit deux fois l'an, tiendra sa 50^{e} réunion l'année prochaine. Le Comité consultatif des méthodes et des nombres, qui est un comité interne (le comité de gestion chargé des enjeux organisationnels dans ce domaine) a tenu sa première réunion en 1983. La revue *Techniques d'enquête* a célébré son 25^{e} anniversaire il y a quelques années, en 2000, mais elle a été adoptée comme publication officielle de Statistique Canada en 1984 (chose intéressante, le dernier numéro de *Techniques d'enquête* publié dans l'ancien format a été celui consacré aux articles du premier symposium sur des questions de méthodologie qui a eu lieu en 1984). Des échanges périodiques sur les questions de méthodologie avec l'U.S. Cernsus Bureau ont également été institués au cours de cette période, puis élargis de manière à inclure la participation du BLS.

Il s'agit donc d'une période où Statistique Canada a reconnu l'importance de la méthodologie pour ses programmes, a reconnu la nécessité de renforcer ses liens et consultations externes et, à l'interne, s'est efforcé d'accroître la collaboration et l'efficacité. La série de symposiums a été conçue dans cet environnement.

Étant donné les priorités et l'influence de Martin Wilk, il était juste que le premier symposium soit ouvert par lui, qu'il s'intéresse à l'analyse des données d'enquête et qu'il soit coparrainé par le Laboratoire de recherche en statistique et probabilités des universités Carleton et d'Ottawa, un organisme qui a continué de coparrainer bon nombre des symposiums subséquents et qui a joué un rôle de premier plan en aidant à asseoir les premiers symposiums sur de solides assises financières.

4. Perfectionnement professionnel

L'un des défis permanents que présente la gestion d'un effectif composé d'un grand nombre de méthodologistes consiste à fournir suffisamment de possibilités de perfectionnement professionnel. Comme l'effectif regroupe au moins 200 membres (ils sont maintenant près de 300, je crois), on ne peut offrir à tous la possibilité d'assister régulièrement à des conférences. La participation des fonctionnaires à des conférences, surtout à l'étranger, entraîne depuis longtemps des formalités administratives et exige des approbations spéciales. En plus de peser la valeur intrinsèque de la participation pour l'employé et pour l'organisme, il a toujours été nécessaire de considérer également la perception du public en ce qui a trait à de nombreux déplacements à l'étranger dans le climat budgétaire actuel.

En outre, au début des années 1980, il y avait assez peu de conférences en lien direct avec les travaux de méthodologie. Nous étions aux prises avec bien des problèmes qui ressemblaient à ceux d'autres organismes statistiques nationaux, de sorte qu'il fallait trouver des expériences comparables à l'extérieur du pays. Les réunions de l'American Statistical Association offraient toujours une gamme de séances pertinentes pour notre travail, même si elles ne représentaient qu'une petite proportion de l'ensemble des séances. Mais il nous était rarement possible d'y envoyer plus de 20 personnes. Et avant que la Société statistique du Canada se dote d'une section des méthodes d'enquête, elle avait peu de lien étroit avec notre travail. Il est vrai que l'Institut international de statistique, et surtout l'Association internationale des statisticiens d'enquêtes, organisait tous les deux ans plusieurs séances d'intérêt pratique, mais peu d'employés pouvaient y aller, car elles avaient lieu le plus souvent ailleurs qu'en Amérique du Nord. C'est pourquoi il est apparu intéressant de tenir chaque année un symposium en lien avec notre travail et auquel un grand nombre de nos employés pourraient assister. Si nous ne pouvions pas envoyer nos gens à l'étranger autant que nous l'aurions voulu, nous pouvions au moins inviter des conférenciers reconnus à venir ici. Pour moi, cela a toujours été l'une des raisons d'être du symposium.

Comme toute conférence, le symposium offre d'autres mesures incitatives et procure d'autres avantages aux auteurs des communications et les fait mieux connaître. Les échéances permettent souvent d'accomplir de véritables progrès. La visibilité des produits peut rehausser les réputations professionnelles et, il faut l'espérer, celles de Statistique Canada également. Le symposium peut offrir aux jeunes employés l'occasion de s'habituer aux communications dans un milieu familier avant de s'aventurer plus loin.

Certes, depuis les années 1980, il s'est tenu bien des réunions sur des aspects de la méthodologie, y compris des conférences thématiques ponctuelles, souvent sous l'égide de l'American Statistical Association, ainsi que des rencontres intergouvernementales sous les auspices d'organismes des Nations Unies ou de l'UE. Il n'en demeure pas moins que le symposium demeure une importante occasion de perfectionnement professionnel pour nos méthodologistes et, il est à espérer, pour ceux de bien d'autres pays.

5. Thèmes

Il suffit de parcourir la liste des thèmes couverts par les 25 (ou 26) symposiums au fil des années pour se faire une idée des défis et des priorités en matière de méthodologie sur 25 ans. Les thèmes se déclinent au goût du jour.
Comme je l'ai dit, il convenait que le premier symposium, ouvert par Martin Wilk, traite de l'analyse des données d'enquête. À Statistique Canada, les méthodologistes s'intéressent depuis toujours à la conception d'enquêtes, de sorte que cela témoignait bien que la préoccupation ultime des utilisateurs était l'analyse des données, que les méthodologistes devaient s'y attacher davantage et qu'il faudrait en tenir compte à l'étape de la conception. L'analyse revient périodiquement dans les thèmes du symposium, jusqu'à cette année. Comme l'a dit François hier dans son discours d'ouverture, le thème des enquêtes longitudinales de cette année est repris de symposiums tenus depuis 1989 et 1992. Le symposium de 1989 sur l'analyse des données dans le temps faisait écho à une rencontre sur les enquêtes par panel tenue à Washington. En 1989, l'analyse des données dans le temps était encore limitée par l'absence d'enquêtes longitudinales.

Le deuxième symposium, sur les statistiques sur petits domaines, a eu lieu à un moment où Statistique Canada proposait une approche plus organisée et plus cohérente du sujet à l'échelle de l'organisme. Il s'est déroulé sous forme de conférence au Centre des congrès d'Ottawa, qui a été démoli récemment. Il combinait un certain nombre des grandes considérations organisationnelles avec l'intérêt méthodologique pour l'estimation sur petits domaines. Il a fait l'objet d'un livre publié chez les éditions Wiley.

En 1986, nous étions de nouveau au sous-sol de l'immeuble R.H.-Coats à discuter des données manquantes dans les enquêtes. En remontant dans le temps, j'ai cru d'abord, ironie du sort, qu'il nous manquait aussi les actes du symposium, mais ils figurent dans le volume 12 de *Techniques d'enquête.* L'année suivante, en 1987, le symposium a porté sur l'utilisation des données administratives. Il s'est tenu dans la vieille gare ferroviaire qui est devenue le Centre de conférences du gouvernement. Le thème était d'actualité, vu les efforts renouvelés pour élargir l'utilisation des données fiscales et la sensibilisation croissante aux problèmes de respect de la vie privée. Je crois qu'il est correct de dire que ce symposium a été le premier à produire un ensemble d'actes autonome. C'était le quatrième symposium et nous parlions déjà d'en faire une série.

Les symposiums de 1985 et de 1987 ont été couronnés de succès, mais ils ont entraîné beaucoup de travail et ils se sont déroulés à l'extérieur de Statistique Canada. Certains d'entre nous appréciaient l'intimité qu'offrait le sous-sol de Statistique Canada et surtout la facilité d'accès pour les employés. Mais le nombre de places était limité et le contrôle de l'accès des visiteurs à l'immeuble posait un problème. De 1988 à 1996, nous avons utilisé le Centre de conférences Simon A. Goldberg à Statistique Canada, mais c'était bondé et les commissaires des incendies étaient après nous. En fait, nous n'avions qu'une salle de séance plénière. Je me souviens très bien de la table ronde à la fin du symposium de 1994, un format que nous avons utilisé

dans divers symposiums pour tâcher de garder l'auditoire jusqu'à la fin. Les statisticiens en chef de l'Australie, du Canada, de la France et de la Nouvelle-Zélande se sont battus pendant quatre rondes pour voir qui sortirait vainqueur! Nous avons eu l'idée d'alterner l'étendue thématique des symposiums de manière que les thèmes de portée générale soient traités à l'extérieur de Statistique Canada et que ceux de portée restreinte soient débattus au sous-sol. Ainsi, en 1997, nous avons tenu le symposium pour la première fois au Palais des congrès (là où nous sommes réunis cette année) et en 1988 et 1999, nous revenions dans les locaux de Statistique Canada. Mais même les thèmes de portée restreinte attiraient plus de participants que ne pouvait accommoder le sous-sol en toute sécurité et nous avons dû abandonner l'idée d'alterner ainsi les symposiums. Nous étions, semble-t-il, incapables d'organiser un petit symposium, seulement un grand. Depuis 1999, il ne s'est plus tenu de symposium au Centre de conférences Simon A. Goldberg.

Le symposium de 1993 visait à remédier à une perception voulant que la méthodologie aurait négligé les enquêtes-entreprises. D'où la première conférence internationale sur les enquêtes-établissements (CIEE). C'était aussi la première fois que le symposium se tenait non pas dans la région d'Ottawa-Gatineau mais assez près, soit à Buffalo, New York, pour nous permettre d'y envoyer un bon nombre de participants de Statistique Canada. Nous avons loué un autobus et pris l'autoroute à péage. Le thème a été repris en 2000, à Buffalo, et de nouveau à Montréal, en 2007.

La question de la qualité des données est omniprésente dans nos symposiums, mais elle a retenu une attention toute particulière en 1990, en 1996 (erreur non due à l'échantillonnage) et en 2001. Nous avons généralement évité d'imposer des restrictions à l'application dans un secteur spécialisé (sauf pour celles implicites dans la CIEE), mais en 2006 nous avons jugé que le thème de la santé de la population était suffisamment vaste et présentait un défi suffisamment grand pour justifier un symposium sur les problèmes de mesure dans ce domaine.

6. Influence et imitation

L'imitation est un signe de succès. À mon avis, la série des symposiums sur les questions de méthodologie a été imitée sur deux plans. D'abord, à Statistique Canada, d'autres milieux professionnels ont reconnu la valeur d'une possibilité offerte régulièrement au personnel de présenter leurs travaux et d'en discuter avec leurs pairs à l'intérieur et à l'extérieur de l'organisme. Aujourd'hui, la collectivité des TI et les économistes et sociologues tiennent chaque année des conférences professionnelles.

En deuxième lieu, d'autres bureaux statistiques ont institué des conférences périodiques sur la méthodologie. Il est peut-être un peu présomptueux de soutenir que ce sont là des imitations du modèle de Statistique Canada, mais nous pouvons au moins dire que ces autres organismes sont parvenus à une conclusion semblable à celle de Statistique Canada au sujet de la valeur de pareilles réunions.

7. Bilinguisme

Je ne saurais conclure mes remarques sans dire un mot sur une particularité de la série des symposiums, soit le bilinguisme. Les deux langues officielles du pays sont utilisées comme langues de travail, de sorte que les employés de Statistique Canada peuvent présenter leurs communications et participer dans la langue officielle de leur choix. La vaste majorité des participants de l'extérieur s'expriment en anglais, mais j'ai toujours été frappé par l'appréciation, et parfois l'étonnement, de ceux qui viennent d'autres pays francophones et qui constatent qu'ils peuvent s'exprimer en français s'ils le souhaitent. J'ai également observé un avantage caché du bilinguisme, à savoir l'amélioration de la version anglaise d'une communication lorsqu'un traducteur peine pour donner un sens à la version française – et probablement vice-versa. Les traducteurs nous obligent à nous exprimer plus clairement dans notre langue première.

8. L'avenir

Et maintenant, que nous réserve l'avenir? Je sais que vous pensez que l'avenir, c'est le dessert et que vous espérez qu'il sera servi bientôt. Je vous demande d'être patients avec moi. Un 25^{e} anniversaire, c'est très bien, mais cela ne veut pas dire nécessairement qu'il faut maintenant viser un 50^{e} symposium. La véritable question qui se pose est celle de savoir si le symposium continue de répondre aux besoins des méthodologistes à Statistique Canada ainsi que sur le plan international, et s'il représente toujours le meilleur moyen de satisfaire à ces besoins. Le monde a beaucoup changé depuis 1984, particulièrement les moyens de communication disponibles aux statisticiens. Comme nous l'avons vu, le symposium s'est adapté et a évolué de nombreuses façons depuis 1984; cependant, il demeure essentiellement une réunion de personnes pendant une brève période chaque année. Il entraîne des préparatifs importants, ce dont peuvent attester des générations de comités

organisateurs. Il y a donc lieu de nous demander si, aujourd'hui, il n'existe pas d'autres mécanismes qui peuvent compléter ou même remplacer un symposium annuel. Je ne me rappelle pas avoir entendu beaucoup de critiques au sujet des symposiums au fil des ans. Les gens sont très polis, surtout au Canada, et ils semblent généralement contents d'avoir l'occasion d'assister au symposium. Toutefois, j'ai entendu un commentaire négatif formulé de temps à autre, à savoir le regret de manquer de temps pour discuter de bon nombre des éléments traités. Cette critique a été formulée à propos non seulement de la série de symposiums, mais de bien des réunions et conférences qui attirent les participants en grand nombre et où la discussion est limitée par le format, les modalités et la nécessité de respecter le programme. Si on envisageait de modifier le cadre du symposium, ce serait là où apporter des améliorations ou des solutions de rechange. N'oubliez pas de remplir vos formulaires d'évaluation – plus de discours aux banquets, peut-être.

9. Conclusion

Lorsque nous avons tenu le premier symposium, en 1984, je ne m'attendais pas à en célébrer ici le 25e anniversaire. Nous avons fait tout ce chemin grâce au dévouement, à la persévérance et au soutien sans relâche des organisateurs, contributeurs et participants. Je tiens à signaler tout particulièrement les efforts déployés par les comités des programmes et les comités chargés des arrangements locaux, grâce auxquels les symposiums ont pu se dérouler. Ce sont leurs membres qui travaillent surtout dans les coulisses pour s'assurer que tout ce dont nous avons besoin est en place. Au cours de ces 25 années, il s'agit probablement de 400 personnes qui ont rendu possible la tenue de cette série de symposiums. Bon nombre d'entre elles sont présentes ici ce soir. Pour terminer, j'aimerais remercier les organisateurs chargés du programme et des arrangements locaux qui nous ont permis de faire tout ce chemin. Votre travail a été tout aussi important pour les progrès réalisés dans le domaine de la méthodologie que les travaux des auteurs dont les communications ont été en vedette au cours des 25 dernières années. Merci.