TECHNIQUES D'ENQUÊTE

N° 12-001-XPB au catalogue

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

JUIN 2001

•

VOLUME 27

•

NUMÉRO 1

Statistique Canada Statistics Canada

Canada

Ca 005

STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA
1010330705

TECHNIQUES D'ENQUÊTE

JUIN 2001 • VOLUME 27 • NUMÉRO 1

Publication autorisée par le ministre
responsable de Statistique Canada

© Ministre de l'Industrie, 2001

Tous droits réservés. Il est interdit de reproduire ou de transmettre
le contenu de la présente publication, sous quelque forme ou
par quelque moyen que ce soit, enregistrement sur support
magnétique, reproduction électronique, mécanique, photographique,
ou autre, ou de l'emmagasiner dans un système de recouvrement,
sans l'autorisation écrite préalable des Services de concession
des droits de licence, Division du marketing,
Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Août 2001

N° 12-001-XPB au catalogue

Périodicité: semestrielle

ISSN 0714-0045

Ottawa

Statistique Canada Statistics Canada

Canada

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président G.J. Brackstone

Membres D.A. Binder
G.J.C. Hole
E. Rancourt (Gestionnaire de la production)
C. Patrick
R. Platek (Ancien président)
D. Roy
M.P. Singh

COMITÉ DE RÉDACTION

Rédacteur en chef M.P. Singh, *Statistique Canada*

Rédacteurs associés

D.R. Bellhouse, *University of Western Ontario*
D.A. Binder, *Statistique Canada*
C. Clark, *U.S. Bureau of the Census*
J.-C. Deville, *INSEE*
J. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
M.A. Hidiroglou, *Statistique Canada*
D. Holt, *University of Southampton, U.K.*
G. Kalton, *Westat, Inc.*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *University of Nebraska-Lincoln*
S. Linacre, *Australian Bureau of Statistics*
G. Nathan, *Hebrew University, Israel*
D. Norris, *Statistique Canada*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
F.J. Scheuren, *The Urban Institute*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
E. Stasny, *Ohio State University*
R. Valliant, *Westat, Inc.*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

Rédacteurs adjoints J.-F. Beaumont, P. Dick, H. Mantel et W. Yung, *Statistique Canada*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de Techniques d'enquête (n° 12-001-XPB au catalogue) est de 47 $ CA par année. Le prix n'inclus pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada: États-Unis 12 $ CA (6 $ × 2 exemplaires); autres pays, 20 $ CA (10 $ × 2 exemplaires). Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division de la diffusion, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au 1 800 700-1033, par télécopieur au 1 800 889-9734 ou par Courriel: order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et staticiens du Québec.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Volume 27, numéro 1, juin 2001

TABLE DES MATIÈRES

Dans ce numéro 1

Article Sollicité Waskbserg

G. NATHAN
Méthodes de téléenquêtes applicables aux enquêtes-ménages – Revue et réflexions sur l'avenir 7

Section spéciale sur l'estimation composite

A.C. SINGH, B. KENNEDY et S. WU
Estimation composite par régression pour l'Enquête sur la population active du Canada avec plan de sondage à renouvellement de panel 35

W.A. FULLER et J.N.K. RAO
Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada 49

P. BELL
Comparaison d'autres estimateurs pour l'Enquête sur la population active 57

J. GAMBINO, B. KENNEDY et M.P. SINGH
Estimation composite par régression pour l'Enquête sur la population active du Canada : Évaluation et application 69

Articles Réguliers

J.-K. KIM
Estimation de la variance après imputation 81

T.E. RAGHUNATHAN, J.M. LEPKOWSKI, J. VAN HOEWYK et P. SOLENBERGER
Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression 91

J. DUFOUR, F. GAGNON, Y. MORIN, M. RENAUD et C.-E. SÄRNDAL
Mieux comprendre la transformation des poids à l'aide d'une mesure de changement 105

P. ARDILLY et D. LE BLANC
Échantillonnage et pondération d'une enquête auprès de personnes sans domicile : un exemple français 117

Techniques d'enquête, juin 2001
Vol. 27, n° 1, pp. 1-2
Statistique Canada

Dans ce numéro

Ce numéro de *Techniques d'enquête* contient le premier d'une série d'articles annuels sollicités publiés en l'honneur de Joseph Waksberg. Une brève description de la nouvelle série et une biographie de Joseph Waksberg précèdent l'article proprement dit. J'aimerais remercier Danny Levine d'avoir écrit la biographie de Joseph Waksberg. J'aimerais également remercier David Binder, Paul Biemer, Graham Kalton et Chris Skinner, les membres du comité de sélection, pour leur choix d'un chercheur de renom dans le domain des enquêtes. Il est le premier auteur de la série Waskberg d'articles sollicités. Je dois des remerciements spéciaux à Graham Kalton qui, en tant que président fondateur du comité, a pris l'initiative, a négocié les arrangements nécessaires avec Westat, l'*American Statistical Association* et *Techniques d'enquêtes* pour mettre le projet en branle, et a travaillé pour rencontrer les délais prescrits par la revue pour la publication du numéro de juin.

L'auteur de l'article de 2001 de la série Waksberg est Gad Nathan. Son article, intitulé « Méthodes de téléenquêtes applicables aux enquêtes auprès des ménages – Revue et réflexions au sujet de l'avenir », fait l'historique de la méthodologie des enquêtes téléphoniques des années 1930 jusqu'à nos jours. Il évoque notamment les questions relatives au plan d'échantillonnage, aux bases de sondage, à la couverture, à la non-réponse et à la pondération. L'article se termine par l'examen de certains défis que posent les progrès technologiques les plus récents, comme le courrier électronique, Internet, le téléphone cellulaire et d'autres transformations technologiques et sociales, ainsi que des perspectives qu'ils offrent.

Ce numéro de *Techniques d'enquête* comprend aussi une section spéciale de quatre articles sur l'estimation composite. Le premier de ces articles, publié par A. C. Singh, Kennedy et Wu, décrit la méthode d'estimation composite par régression mise au point par Singh et ses collègues ces dernières années. Les auteurs comparent la nouvelle approche aux méthodes antérieures d'estimation composite, notamment les estimateurs composites *K* et *AK*. Ils donnent aussi une description heuristique de leur nouvelle méthode et les raisons pour lesquelles ils l'ont mise au point. Les avantages de la nouvelle méthode tiennent au fait qu'elle produit un ensemble unique de poids d'estimation qui assure la cohérence interne des estimations obtenues, tout en améliorant l'efficacité des estimateurs classiques de régression.

Fuller et Rao évaluent analytiquement les propriétés de l'estimation composite par régression. Ils commencent par décrire deux des premières variantes de l'estimation composite par régression, appelées estimateurs de régression modifiés (*MR*1 et *MR*2), puis analysent l'efficacité et le comportement des estimations dans le temps en se servant d'un modèle simple de série chronologique pour les estimations calculées d'après les données d'enquête par panel. Ils concluent qu'un estimateur modifié, qui serait un compromis entre *MR*1 et *MR*2, offrirait dans l'ensemble les propriétés les meilleures.

Dans son article, Bell compare l'application de divers estimateurs aux données de l'Australian Labour Force Survey. Il étudie notamment l'estimateur composite *AK*, la première variante de l'estimateur composite de régression appelée *MR*2, la variante de l'estimateur composite de régression de Fuller et Rao, ainsi qu'un meilleur estimateur linéaire non biaisé (MELNB) choisi comme étant une combinaison linéaire « optimale » des estimations fondées sur des données de panel. Il propose aussi un MELNB amélioré, obtenu en étalonnant le MELNB par rapport à des données de référence sur la population. La comparaison des estimateurs susmentionnés aux estimateurs de régression classiques porte sur l'écart entre les estimations qu'ils produisent et celles obtenues au moyen des estimateurs classiques, leurs écarts-types et leur utilité pour la désaisonnalisation et l'estimation de la tendance.

Le dernier article de la section spéciale, publié par Gambino, Kennedy et M.P. Singh, décrit l'estimateur composite de régression utilisé au Canada pour l'Enquête sur la population active. Cet estimateur se fonde sur les travaux de A. C. Singh et de ses collègues et sur le compromis proposé par Fuller et Rao. Les auteurs comparent les nouveaux estimateurs aux estimateurs de régression classiques utilisés antérieurement en les appliquant à plusieurs séries de données. Ils constatent que les nouveaux estimateurs sont ordinairement plus efficaces et plus stables et que leur utilisation pour la désaisonnalisation de la série d'estimations donne plus fréquemment de bons résultats.

Kim propose une nouvelle méthode d'estimation de la variance basée sur un modèle linéaire d'imputation par régression qui tient compte de l'imputation aléatoire. La méthode consiste à créer un ensemble de pseudovaleurs pour y, de sorte que l'estimateur classique de la variance calculé d'après ces pseudovaleurs tienne aussi compte de l'imputation. Le calcul des pseudovaleurs est décrit d'abord pour l'échantillonnage aléatoire simple, puis pour des plans de sondage complexes. L'auteur montre que la méthode est asymptotiquement équivalente à la méthode corrigée du jackknife de Rao et Sitter et étudie ses propriétés au moyen d'une étude en simulation.

Dans leur article intitulé « Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression », Raghunathan, Lepkowski, Van Hoewyk et Solenberger abordent la question importante de l'imputation dans le cas d'une structure de données complexe où il est difficile de formuler des modèles multidimensionnels complets explicites. Leur stratégie consiste à procéder à l'imputation, une variable à la fois, en la subordonnant toutefois à toutes les variables observées. Autrement dit, les imputations sont réalisées suivant une série de régressions multiples dont le modèle varie selon le type de variable visée par l'imputation.

Dans leur article, Dufour, Gagnon, Morin, Renaud et Särndal proposent une mesure de distance qui permet de mesurer l'incidence relative de l'ajustement de non-réponse, de la calibration et de l'interaction entre ces deux procédures. Cette mesure leur permet d'étudier et de mesurer le changement (du poids initial au poids final) qui est produit par la procédure de modification des poids. Ils utilisent cette mesure comme outil pour comparer l'efficacité de diverses méthodes d'ajustement pour la non-réponse par l'entremise d'une étude de simulation auprès des données de l'Enquête sur la dynamique du travail et du revenu. La mesure est également appliquée aux données de l'Enquête nationale longitudinale sur les enfants et les jeunes.

Ces dernières années, les efforts se sont multipliés en vue d'étudier les populations de sans-abri des grandes villes. La difficulté qu'il y a à élaborer une base de sondage et une méthode d'échantillonnage fiable et efficace, ainsi que la fluidité de la population au fil du temps rendent ce genre d'études particulièrement complexes. Le dernier article du présent numéro, publié par Ardilly et Le Blanc, décrit les méthodes d'échantillonnage et d'estimation adoptées pour une enquête menée à l'heure actuelle en France auprès des sans-abri. Y sont également décrits les problèmes et les défis que pose ce genre d'enquête. L'échantillonnage des personnes sans abri se fera indirectement, par échantillonnage des services qu'elles sont susceptibles d'utiliser, comme les refuges et les services de repas. Les auteurs montrent que la méthode de partage des poids est un moyen efficace d'obtenir des poids non biaisés pour diverses périodes de référence, comme une journée moyenne ou une semaine moyenne.

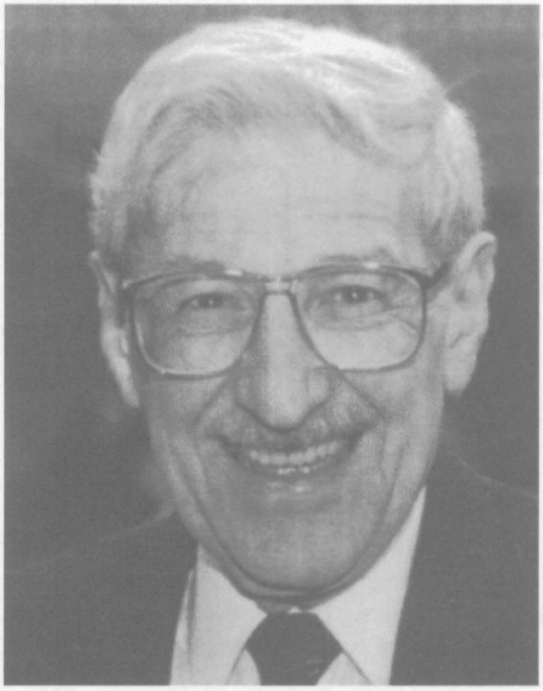
Finalement, j'aimerais profiter de l'occasion pour exprimer de sincères remerciements à Frank Mayda, Gestionnaire de la production de *Techniques d'enquête*, qui a récemment pris sa retraite. L'implication qu'il a eue dans *Techniques d'enquête* depuis 1987 est inestimable. J'aimerais aussi annoncer que Eric Rancourt remplace Frank Mayda en tant que Gestionnaire de la production.

M.P. Singh

Techniques d'enquête, juin 2001
Vol. 27, n° 1, pp. 3-5
Statistique Canada

Série Waksberg d'articles sollicités

Le comité de rédaction de *Techniques d'enquête* a décidé de publier une série d'articles annuels sollicités en l'honneur de Joseph Waksberg, pour souligner sa contribution importante à la méthodologie d'enquête. Chaque année nous inviterons un spécialiste renommé de la recherche en sondages à rédiger un article consacré à la rétrospective et à l'examen de la situation courante d'un domaine important de la méthodologie d'enquête. L'auteur reçoit un prix monétaire grâce à une subvention offerte par Westat en reconnaissance de la contribution de Joe Waksberg durant les nombreuses années où il a travaillé pour l'entreprise. L'*American Statistical Association* est chargée de la gestion financière et administrative de la subvention. L'auteur de l'article est choisi par un comité de quatre personnes désignées par *Techniques d'enquête* et l'*American Statistical Association*.



JOSEPH WAKSBERG

À l'heure actuelle, Joseph Waksberg (que tout le monde appelle « Joe ») est président du Conseil d'administration de Westat, une entreprise de recherche statistique établie à Rockville, Md. Au cours de sa carrière, qui s'étend maintenant sur plus de 60 ans, il a fait d'importantes contributions à la théorie de l'échantillonnage a développé, des applications innovatrices de la théorie et mené des travaux de recherche sur des problèmes très variés de méthodologie d'enquête. Il est l'auteur ou le coauteur de nombreux articles sur les méthodes d'échantillonnage, y compris la composition aléatoire, l'échantillonnage des populations rares, l'échantillonnage pour les enquêtes par panel et avec renouvellement de l'échantillon et le rôle de l'échantillonnage dans les recensements de population. Ses autres contributions vont de la recherche méthodologique sur l'évaluation de la population active à l'estimation régionale et à la création de modèles pour prédire les vainqueurs la nuit des élections, en passant par l'évaluation de la qualité des données de recensement aux États-Unis, les effets du télescopage et d'autres problèmes de mémorisation sur les résultats d'enquête et l'étude des effets des récompenses monétaires sur les taux de réponse et les coûts d'enquête. Son but a été d'améliorer et la théorie et la pratique. Enfin, mais non le moindre de ses accomplissements, il a été le maître et le mentor de nombreuses générations de statisticiens.

Né à Kielee, en Pologne, en septembre 1915, Joe a émigré aux États-Unis avec sa famille en 1921. En 1936, peu après avoir obtenu un diplôme en mathématiques à la City University of New York (CUNY), il s'est installé dans la région de Washington D.C. et, après un bref séjour au Navy Department, s'est joint au Census Bureau en 1940 à

titre de commis. Il y a travaillé pendant 33 ans jusqu'au moment de sa retraite, en 1973. À ce moment-là, il était directeur adjoint du service des méthodes statistiques, de la recherche et des normes. Au début des années 60, Waksberg et Neter ont entrepris une étude classique sur l'importance de divers problèmes de mémorisation. Cet effort marquant a permis de mettre au point des méthodes visant à réduire les effets de ces problèmes grâce à une stratégie innovatrice d'échantillonnage et de collecte de données. (Neter et Waksberg 1964; Neter et Waksberg 1965). Par la suite, Joe a continué de s'intéresser à ce domaine; par exemple, il a participé à la conception et à l'analyse des résultats d'une expérience réalisée pour déterminer le sens et la grandeur des biais éventuels dans le cas d'une enquête comportant une période de rappel d'un an pour le U.S. Fish and Wildlife Service (Chu, Eisenhower, Hay, Morganstein, Neter et Waksberg 1992). Cette expérience a eu un effet important sur le remaniement de l'enquête, mais a, avant tout et par dessus tout, élargit considérablement le champ des connaissances sur le biais de réponse quand on demande aux répondants de se remémorer la fréquence de certaines activités pour diverses périodes de rappel, et indiqué des méthodes pour minimiser l'erreur quadratique moyenne du plan de telles enquêtes.

La Current Population Survey (CPS) américaine, qui est considérée aujourd'hui comme un modèle d'efficacité statistique, reflète pleinement l'influence et la contribution de Joseph Waksberg à l'époque où il était responsable de l'échantillonnage, des normes statistiques et de la recherche pour le programme des enquêtes-ménages du Census Bureau. L'amélioration des méthodes d'échantillonnage et d'estimation, y compris l'utilisation d'échantillonnage à partir de bases-listes, les méthodes d'estimation de la variance par répétition, la détermination de la taille appropriée des grappes, le traitement des événements rares et l'estimation composite sont des changements qui portent tous son empreinte. Parallèlement, Joe a joué un rôle important dans la recherche expérimentale touchant à de nouveaux scénarios de renouvellement d'échantillon et d'estimation, à l'utilisation d'un seul répondant par ménage, et aux effets de périodes de rappel variables sur mesure de la population active.

Aucun exposé des travaux de Joe au Census Bureau ne serait complet sans la mention de ses nombreuses contributions au programme de recensement décennal. Le programme d'évaluation qu'il a développé, conçu et dirigé pour le Recensement de 1970 en est un bon exemple. Ce programme, qui regroupait 25 projets distincts, a été qualifié à l'époque de « radical »; aujourd'hui, il sert de modèle aux programmes courants de recherche sur le recensement décennal. Durant le Recensement de 1970, lorsque les premiers résultats des travaux sur le terrain ont indiqué une surestimation grave des unités « inoccupées », Waksberg a conçu, élaboré et mis en œuvre, en très peu de temps et avec des moyens financiers très limités, un programme innovateur d'enquête par sondage consistant à rendre de nouveau visite à un échantillon d'unités inoccupées qui a permis de montrer qu'une proportion importante de ces unités étaient, en fait, occupées. Il a alors mis au point et appliqué une méthode de rajustement à l'univers des unités inoccupées repérées durant le recensement, au niveau des petites régions (Waksberg 1998). Par la suite, après l'adoption de la loi sur le partage des revenus de 1972 qui exigeait que le Bureau produise des estimations annuelles de la population et du revenu par habitant pour les 39 000 unités gouvernementales des États-Unis, Waksberg a proposé d'utiliser des dossiers administratifs ainsi que des données d'enquêtes pour produire les estimations régionales requises de la population et du revenu par habitant. Il a entrepris des travaux de recherche sur l'appariement des dossiers de l'IRS couvrant deux années consécutives en vue d'obtenir des estimations régionales (au niveau du comté) des migrations brute et nette, et de la variation du niveau de revenu. Ces recherches ont mené au développement et à la mise en oeuvre d'un programme d'estimation pour petites régions qui est encore utilisé aujourd'hui.

Durant ses années à Westat depuis 1973, d'abord en qualité de statisticien principal et vice-président, et récemment comme consultant et président du conseil, Waksberg a continué de faire preuve de la même passion pour l'innovation, l'expérimentation et la qualité lorsqu'il s'efforçait de répondre aux besoins de ses clients, de mettre au point des échantillons ou d'exécuter des travaux de recherche sur les enquêtes. Lors de l'élaboration des plans de sondage de la National Health Interview Survey et de la National Health and Examination Survey du National Center for Health Statistics, il a participé à l'élaboration de méthodes innovatrices permettant de suréchantillonner plus efficacement les populations minoritaires (Waksberg 1973). Les travaux qu'il a réalisés en collaboration avec Judkins et Massey fournissent des renseignements importants sur la concentration résidentielle selon la race et l'origine ethnique, qui sont essentiels à l'évaluation de l'utilité dans certaines régions géographiques du suréchantillonnage des populations minoritaires des personnes pauvres, une autre sous-population qu'il est souvent nécessaire de suréchantillonner (Waksberg, Judkins et Massey 1997). En collaboration, il a developpé la méthode Mitofsky-Waksberg d'échantillonnage à deux phases des ménages abonnés au téléphone (Waksberg 1978). Cette méthode est devenue la méthode standard d'échantillonnage par composition aléatoire (CA) aux États-Unis. En vue de l'améliorer, Waksberg a étudié le biais que la méthode de composition aléatoire introduit comparativement à l'échantillonnage basé sur une liste (Waksberg 1983; Brick et Waksberg 1991). Cette étude, qui a permis de modifier la méthode et de la rendre plus efficace, a mené à une toute nouvelle méthode d'échantillonnage par CA (Brick, Waksberg, Kulp et Starer 1995). Plus récemment, il a participé a une étude sur des méthodes alternatives d'ajustement pour les ménages n'ayant pas de téléphone (Brick, Waskberg et Keeter 1996). Les travaux de Waksberg dans le domaine de

l'échantillonnage par CA illustrent fort bien le désir qu'il a eu toute sa vie de réexaminer constamment les méthodes statistiques et d'en découvrir de nouvelles, supérieures à celles considérées comme étant la norme, incluant celles qu'il avait contribué à établir.

M. Waksberg a partagé ses connaissances et son expérience avec d'autres dans de multiples endroits. Pendant de nombreuses années, il a enseigné à la Graduate School du U.S. Department of Agriculture et a été un professeur régulier du programme d'été sur les méthodes d'échantillonnage de l'Université du Michigan. Il a aussi rempli le rôle d'expert-conseil en échantillonnage et en techniques d'enquête auprès d'organismes officiels de la statistique de nombreux pays, sous le parrainage de la U.S. Agency for International Development et des Nations- Unies, ainsi qu'à la demande de pays particuliers. Il a fourni des conseils aux agences statistiques de la Chine, de l'Argentine, du Brésil, de Cuba, du Venezuela, de la Turquie et du Vietnam du Sud. Il a aussi représenté les États-Unis à des colloques internationaux sur la statistique, a joué le rôle d'expert technique sous les auspices des Nations-Unies et a fait partie d'une équipe envoyée dans divers pays d'Amérique du Sud par l'*American Statistical Association* pour coordonner les activités de leur société statistique nationale.

Il est membre de l'*American Statistical Association*, qui lui a conféré le titre de Fellow, de *l'Association internationale des statisticiens d'enquêtes* et de l'*Institut international de statistique*, et a participé à divers groupes de travail de la National Academy of Sciences chargés d'évaluer des programmes statistiques fédéraux particuliers. Il a été le premier lauréat du prix Roger Herriott décerné par la *Washington Statistical Society*, et les sections Government Statistics et Social Statistics de l'ASA pour « innovation dans la statistique fédérale », et est un récipiendiaire de la Gold Medal Award du U.S. Commerce Department. Enfin, son héritage le plus important sera peut-être le très grand nombre de collègues qui ont été inspirés par son exemple personnel, ses enseignements, son leadership, ainsi que sa gentillesse, sa bienveillance et sa compréhension.

BIBLIOGRAPHIE

BRICK, J. M., et WAKSBERG, J. (1991). Méthode pour éviter l'échantillonnage progressif dans une enquête téléphonique à composition aléatoire. *Techniques d'enquête*, 17, 31-46.

BRICK, J. M., WAKSBERG, J. et KEETER, S. (1996). Utilisation des données sur les interruptions du service téléphonique pour ajuster la couverture. *Techniques d'enquête*, 22, 187-199.

BRICK, J. M, WAKSBERG, J., KULP, D. et STARER, A. (1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly*, 59, 218-235.

CHU, A., EISENHOWER, D., HAY, M. MORGANSTEIN, D., NETER, J. et WAKSBERG, J. (1992). Measuring the recall error in self-reported fishing and hunting activities. *Journal of Official Statistics*, 8, 19-39.

NETER, J., et WAKSBERG, J. (1964). A study of response errors in expenditure data from household interviews. *Journal of the American Statistical Association*, 59, 18-55.

NETER, J., et WAKSBERG, J. (1965). Response Errors in Collection of Expenditure Data from Household Interviews: An Experimental Study. (Bureau of the Census Technical Paper No. 11). Washington, DC: U.S. Government Printing Office.

WAKSBERG, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. *Proceedings of the Social Statistics Section, American Statistical Association*, 1973, 429-434.

WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

WAKSBERG, J. (1983). A note on locating a special population using random digit dialing. *Public Opinion Quarterly*, 47, 576-579.

WAKSBERG, J. (1998). The Hansen Era: Statistical research and its implementation at the U.S. Census Bureau, 1940-1970. (avec discussion). *Journal of Official Statistics*, 14, 119-147.

WAKSBERG, J., JUDKINS, D. et MASSEY J. (1997). Suréchantillonnage géographique dans les enquêtes démographiques aux États-unis. *Techniques d'enquête*, 23, 69-80.

ARTICLE SOLLICITÉ WAKSBERG 2001

Auteur: Gad Nathan

Gad Nathan est professeur de statistique à la Hebrew University of Jerusalem et travaille de longue date pour le Central Bureau of Statistics d'Israel, le plus récemment à titre d'expert scientifique en chef. Il a obtenu son diplôme de doctorat au Case Institute of Technology, à Cleveland, en Ohio, et a publié de nombreux articles dans d'importantes revues statistiques, y compris *Journal of the American Statistical Association, Journal of the Royal Statistical Society, Techniques d'enquête, Journal of Official Statistics* et *Sankhya*. Ses principaux domaines de recherche sont la méthodologie d'échantillonnage, l'inférence à partir d'échantillons complexes, l'interview assistée par ordinateur et les téléenquêtes. Il a occupé divers postes de professeur invité et d'expert dans plusieurs établissements universitaires et bureaux de la statistique d'Amérique du Nord et d'Europe, et a été vice-président de l'*Institut international de statistique* et de l'*Association internationale des statisticiens d'enquête*, ainsi que président de l'Israel Statistical Association et de l'Israel Public Council of Statistics.

Techniques d'enquête, juin 2001
Vol. 27, n° 1, pp. 7-34
Statistique Canada

Méthodes de téléenquêtes applicables aux enquêtes-ménages – Revue et réflexions sur l'avenir

GAD NATHAN[1]

RÉSUMÉ

Nous appelons « téléenquête » les enquêtes pour lesquelles le mode principal ou unique de collecte des données repose sur un moyen électronique de télécommunications y compris le téléphone et d'autres dispositifs technologiques plus avancés, comme le courrier électronique, Internet, la vidéophonie ou la télécopie. Nous examinons, brièvement, les débuts des enquêtes par téléphone et, plus en détail, les progrès récents dans les domaines du plan de sondage et de l'estimation, de la couverture et de la non-réponse, ainsi que de l'évaluation de la qualité des données. Ces progrès méthodologiques ont fait de l'enquête téléphonique le mode principal de collecte des données dans le domaine des enquêtes par sondage au cours du dernier quart de siècle. D'autres moyens de télécommunication de pointe deviennent rapidement des compléments importants, voire même des concurrents, du service téléphonique fixe et sont déjà utilisés de diverses façons pour réaliser les enquêtes par sondage. Nous examinons leur potentiel pour les opérations d'enquête et l'effet que pourraient avoir les progrès technologiques actuels et futurs dans le secteur des télécommunications sur les pratiques d'enquête et leurs conséquences méthodologiques.

MOTS CLÉS : Enquêtes téléphoniques; enquêtes par Internet; plan d'échantillonnage; non-réponse; couverture.

1. INTRODUCTION

À l'aube du nouveau millénaire, les télécommunications électroniques sont devenues un élément prédominant de presque tous les aspects de la vie contemporaine. Nos enquêtes par sondage ne font pas exception et le recours généralisé au téléphone comme mode principal de communication durant au moins le dernier quart de siècle, a fortement influencé les pratiques d'enquête. En fait, dans le domaine des enquêtes par sondage, l'enquête téléphonique est devenue l'instrument dominant de collecte des données, surtout en Amérique du Nord et en Europe de l'Ouest, qu'il s'agisse d'enquêtes menées auprès de ménages, de particuliers ou d'établissements. D'autres modes de télécommunication de pointe, comme le courrier électronique, Internet, la vidéophonie, la télécopie et la téléphonie mobile, deviennent rapidement des compléments importants, voire même des concurrents, du service téléphonique fixe. Ils sont déjà appliqués de diverses façons aux enquêtes par sondage et nous avons l'intention d'examiner ici les possibilités qu'ils offrent pour les opérations d'enquête et les conséquences méthodologiques de leur utilisation. Par conséquent, nous qualifions de « téléenquête » toute enquête pour laquelle le mode prédominant ou unique de collecte des données repose sur un moyen électronique de télécommunications, y compris le téléphone et d'autres dispositifs technologiques plus avancés. Les enquêtes classiques, fondées sur l'interview sur place, ou les enquêtes par la poste ne sont pas incluses, à moins qu'une de leurs composantes importantes ne s'appuie sur un instrument de télécommunication. Bien que le présent article se concentre sur les enquêtes auprès des particuliers et des ménages, une grande partie s'applique aussi aux enquêtes auprès des établissements. Nous parlons des « méthodes » de téléenquête, car il semble évident qu'aucune ne puisse convenir à elle seule à l'application de la pléthore de dispositifs de communication dont on pourrait disposer dans l'avenir et de leur combinaison.

Le présent article est rédigé en reconnaissance de la contribution unique de Joe Waksberg à l'élaboration de la méthodologie d'enquête en général et de celle des enquêtes téléphoniques en particulier. Il est généralement admis aujourd'hui que son article innovateur (Waksberg 1978) a frayé la voie à l'application généralisée et efficace de la composition aléatoire aux enquêtes téléphoniques et représente un jalon dans le développement de la méthodologie des téléenquêtes. Conjugué aux nombreux autres articles qu'il a publiés par la suite, ce document a eu une profonde influence théorique et pratique sur la méthodologie des enquêtes téléphoniques que nous nous proposons de décrire en partie ici.

Nous nous attacherons principalement aux aspects statistiques de la méthodologie des téléenquêtes, tout en reconnaissant qu'ils ne sont pas indépendants des aspects non statistiques, comme les caractéristiques cognitives de la téléinterview, l'administration des enquêtes et les considérations d'ordre éthique. Dans la section qui suit, nous passons brièvement en revue les débuts des enquêtes téléphoniques, jusqu'à 1978. À la troisième section, nous examinons de façon assez approfondie les faits plus récents dans le domaine du plan d'échantillonnage et de l'estimation, de la couverture et de la non-réponse, ainsi que de l'évaluation de la qualité des données. Enfin, à la quatrième section, nous étudions les effets que pourraient avoir les

[1] Gad Nathan, Departement of Statistics, Hebrew University, 91905 Jerusalem, Israel.

progrès technologiques courants et futurs du secteur des communications sur les pratiques d'enquête et leurs conséquences méthodologiques.

2. LES DÉBUTS DES ENQUÊTES TÉLÉPHONIQUES

Dans la section qui suit, nous examinons, brièvement et globalement, les débuts de l'utilisation du téléphone pour réaliser les enquêtes, afin de placer dans son contexte le développement des méthodes de téléenquête que nous décrirons plus tard. Le sujet est abordé de façon plus détaillée et plus complète dans plusieurs traités et articles, dont Blankenship (1977a), Groves, Biemer, Lyberg, Massey, Nicholls et Waksberg (1988), Frey (1989), Lavrakas (1993), Casady et Lepkowski (1998, 1999) et Dillman (1978, 2000).

Le téléphone a commencé à être utilisé pour réaliser des sondages dans les années 1930, en général comme moyen supplémentaire de collecte. Certains attribuent au moins partiellement, mais à tort, au sous-dénombrement téléphonique l'échec désastreux du sondage du *Literary Digest* qui avait prédit la victoire écrasante de Landon sur Roosevelt en 1936 (Katz et Cantril 1937; Payne 1956; Perry 1968). En fait, le sondage se fondait sur un questionnaire envoyé par la poste et, bien que l'on ait utilisé comme base de sondage des listes téléphoniques (conjuguées à des listes d'enregistrements de véhicules à moteur), il semble que l'échec soit attribuable davantage à la non-réponse qu'à la sous-représentation de la base de sondage (Bryson 1976; Squire 1988; Cahalan 1989).

Les premiers rapports sur la réalisation d'enquêtes par téléphone concernaient, pour la plupart, le domaine de la santé publique ou les applications d'étude de marché. Nombre de ces enquêtes s'appuyaient sur la combinaison d'interviews téléphoniques à d'autres modes de collecte et, dans certains cas, comportaient des comparaisons empiriques de taux de réponse ou de résultats, en vue d'évaluer les effets du mode de collecte des données. Par exemple, Cunningham, Westerman et Fischoff (1956) et Bennet (1961) décrivent l'utilisation d'enquêtes téléphoniques pour des études de suivi du traitement des patients, tandis que Fry et McNaire (1958) la décrivent dans le cas du suivi national d'une enquête par la poste en vue de recueillir les opinions du personnel hospitalier – utilisation qui a produit dans l'un et l'autre cas un taux de réponse élevé. Mitchell et Rogers (1958) ont recouru à l'interview téléphonique pour réaliser auprès des ménages ayant le téléphone une enquête sur la consommation de produits laitiers et ont comparé les résultats à ceux obtenus auprès d'un échantillon témoin de ménages ne possédant pas le téléphone. Cahalan (1960), quant à lui, compare les résultats d'interviews téléphoniques à ceux d'interviews sur place réalisées pour évaluer le lectorat des journaux et obtient des résultats favorables. Grâce à une étude téléphonique comparative de la remémorisation de la publicité et de l'utilisation des produits, Eastlack (1964) montre que le protocole rigoureux de rappels produit des résultats plus exacts qu'une méthode sans rappel. Coombs et Freedman (1964) mentionnent un taux de réponse élevé (92 %) dans le cas d'une enquête longitudinale sur la fécondité complétée par des interviews sur place. Sudman (1966) décrit plusieurs autres applications du téléphone pour réaliser les enquêtes, y compris la prise préalable d'un rendez-vous et le dépistage des populations rares, qui augmentent les taux de coopération et permettent de réduire les coûts.

C'est à la fin des années 1960 que les enquêtes téléphoniques ont vraiment pris leur essor, grâce à plusieurs faits nouveaux. En premier lieu, la croissance rapide de la pénétration du téléphone en Europe de l'Ouest et en Amérique du Nord a permis d'adopter l'interview téléphonique comme mode principal de collecte des données. Aux États-Unis, la proportion de ménages équipés du téléphone atteignait 88 % en 1970 (Massey et Botman 1988) et ce niveau de pénétration a été atteint un peu plus tard dans les pays d'Europe de l'Ouest, en Australie et en Nouvelle-Zélande (Trewin et Lee 1988). Parallèlement à l'augmentation rapide de la pénétration du téléphone dans de nombreux pays, à la fin des années 1960, les enquêteurs ont constaté une baisse importante des taux de réponse et éprouvé des difficultés à communiquer sur place avec les répondants pour recueillir les données. Cette situation les a poussé à envisager sérieusement le passage aux enquêtes téléphoniques en vue de réduire les coûts et d'obtenir un taux plus élevé de coopération. Ce sont les organismes commerciaux et universitaires de sondage qui ont adopté le plus rapidement l'interview téléphonique; les bureaux gouvernementaux de la statistique ont mis plus de temps à le faire. Par exemple, selon Federal Committee on Statistical Methodology (1984) en 1981, 11 % seulement des enquêtes de l'administration fédérale américaine comptaient une composante d'interview téléphonique, dans la plupart des cas combinée à d'autres modes de collecte.

Au début, l'interview téléphonique suscitait des appréhensions, même quand elle servait uniquement de mode complémentaire de collecte, car on craignait que le taux de non-réponse soit élevé et qu'un biais de réponse, considéré comme inhérent aux interviews non réalisées sur place, entâche les données. Les résultats des premières enquêtes téléphoniques semblaient, dans certains cas, confirmer ces craintes. Par exemple, une étude de la réception d'un dépliant réalisée par Larson (1952) suscite de sérieux doutes, fondés sur des interviews sur place subséquentes, quant à la validité des réponses téléphoniques. Pareillement, lors d'une enquête sur l'amélioration d'un service à la clientèle, Oakes (1954) obtient un taux de réponse beaucoup plus faible dans le cas de l'interview téléphonique que dans celui de l'interview sur place. Dans le cadre d'une enquête attitudinale sur les finances des consommateurs, Schmiedeskamp (1962) constate que les personnes interviewées par téléphone sont davantage portées à ne pas

prendre catégoriquement position. Lors d'une comparaison de l'envoi d'un questionnaire par la poste, de l'interview téléphonique et de l'interview sur place, Wiseman (1972) observe un effet de mode d'interview dans le cas de questions délicates (sur l'avortement et la contraception). Toutefois, c'est pour l'envoi d'un questionnaire par la poste et l'interview personnelle (par téléphone ou sur place) que la différence est la plus importante.

Les résultats de plusieurs études empiriques plus rigoureuses ont apaisé rapidement nombre des craintes susmentionnées. Ainsi, lors d'une expérience contrôlée bien conçue, Hochstim (1967) compare la collecte des données quand le mode principal est l'envoi d'un questionnaire par la poste, l'interview téléphonique ou l'interview sur place. L'étude donne des preuves convaincantes que les trois méthodes de collecte des données sont presque interchangeables si l'on s'en tient au taux de réponse, à la complétude de la réponse, à la comparabilité des résultats et à la validité des réponses. La différence importante entre les divers modes de collecte a trait au coût et joue manifestement en faveur de l'enquête par la poste ou par téléphone. Pareillement, un petit essai réalisé par Colombotos (1965) sur des échantillons d'une population de médecins ne révèle aucun écart significatif entre les réponses obtenues par téléphone et en personne. Janofsky (1971) constate que les répondants sont aussi disposés à exprimer leurs sentiments concernant des questions de santé par téléphone qu'en personne. L'étude de validation bien conçue menée par Locander, Sudman et Bradburn (1976) sur les effets du mode de collecte en cas de questions délicates ne révèle aucune différence significative entre les biais de réponse associés à l'interview téléphonique et à l'interview sur place. Enfin, au moyen d'une petite expérience sur le terrain minutieusement contrôlée, Rogers (1976) a testé les effets de diverses stratégies d'interview sur la qualité des réponses et sur les résultats sur le terrain, lors d'une enquête comportant une gamme de questions complexes sur les attitudes, les connaissances et les caractéristiques personnelles. De nouveau, l'étude montre que les données obtenues par téléphone sont de même qualité que celles obtenues lors d'interviews sur place. Groves et Kahn (1979) ont réalisé une étude nationale importante en vue de comparer l'interview téléphonique et l'interview sur place. Leur analyse approfondie des résultats de grandes enquêtes omnibus réalisées en appliquant les deux méthodes de collecte par University of Michigan Survey Research Center a fourni des données importantes sur la qualité des données qui ne témoignent d'aucun effet significatif du mode de collecte. Cette étude et d'autres qui ont été les précurseurs des études systématiques des effets du mode de collecte réalisées durant les années 1990 (décrites plus loin) ont contribué à la légitimisation des enquêtes téléphoniques en tant que méthode type de collecte des données.

Au départ, l'utilisation du téléphone pour réaliser les enquêtes par sondage se fondait surtout sur des échantillons sélectionnés à partir de bases de sondage générales, comme les annuaires téléphoniques, ou à partir de bases de sondage particulières pour les petites sous-populations. Vers la fin des années 1960, on avait pris conscience du taux croissant de numéros de téléphone non publiés et des différences importantes entre les ménages dont le numéro était publié et ceux dont le numéro ne l'était pas (pour plus de précisions, consulter la section 3.1.1). La méthode d'échantillonnage par composition aléatoire (CA), introduite pour la première fois par Cooper (1964) et améliorée par la suite par Eastlack et Assael (1966) et par Glasser et Metzger (1972), représente un progrès important qui a permis de surmonter le problème. Une des inefficacités inhérentes aux trois méthodes fondamentales de CA tenait au grand nombre de numéros composés qui ne produisaient pas d'interview (numéros non en service ou non résidentiels). Mitofsky (1970) a été le premier à proposer une méthode d'échantillonnage par CA à deux degrés pour contourner ce problème; par la suite, Waksberg (1978) l'a élaborée et lui a donné un support théorique solide. L'introduction de cette méthode, qui allait devenir la méthode de Mitofsky-Waksberg, a beaucoup contribué à la généralisation des enquêtes téléphoniques durant les années 1980 et 1990.

Enfin, les progrès réalisés durant les années 1960 et 1970 dans le domaine des télécommunications et de l'automation ont aussi donné un avantage à l'enquête téléphonique. L'accès universel à la composition automatique des numéros interurbains a permis de réaliser plus facilement des enquêtes nationales à partir d'un seul centre téléphonique ou à partir d'un petit nombre de centres d'interview avec tous les avantages d'un contrôle et d'une administration centralisés. Cependant, l'élément qui a stimulé le plus l'expansion des enquêtes téléphoniques est, sans aucun doute, l'introduction de l'interview téléphonique assistée par ordinateur (ITAO) durant les années 1970. Cet effet important tient à la fois à la simplicité de l'ITAO pour réaliser des enquêtes téléphoniques et aux possibilités qu'elle offre en ce qui concerne l'automatisation de nombreuses tâches importantes connexes à l'interview (comme la composition, l'établissement des calendriers de rappel, *etc.*).

L'une des premières applications de l'ordinateur à l'interview téléphonique a pris la forme d'une expérience de laboratoire axée sur l'utilisation d'un ordinateur muni de plusieurs postes de travail conçus pour recueillir des renseignements subjectifs (Shure et Meeker 1970). Le numéro spécial de *Sociological Methods and Research* (Freeman et Shanks 1983), publié après la conférence de Berkeley sur les techniques d'enquête assistée par ordinateur qui s'est tenue au printemps 1981 décrit bien les débuts de l'ITAO. Les organismes spécialisés dans les études de marché ont été les premiers à adopter les systèmes d'ITAO pour leurs opérations courantes. En 1972, Chilton Research Services avait déjà mis au point et utilisait régulièrement le Survey Response Processor (Fink 1983). D'autres organismes commerciaux d'enquête qui utilisaient des systèmes différents se sont vite rendu compte des avantages de

l'ITAO – par exemple, le système A&S/CATI™ (Dutka et Frankel 1980). Les organismes universitaires d'enquête par sondage leur ont emboîté le pas rapidement grâce au premier système mis au point à l'UCLA et à Berkeley pour la grande enquête sur l'incapacité en Californie basée sur l'ITAO (Shanks, Nicholls et Freeman 1981; et Shanks 1983). Un autre projet précoce de mise au point d'un système d'ITAO par un organisme universitaire d'enquête, selon une approche différente fondée sur des micro-ordinateurs, a été celui de l'Université du Wisconsin (Palit 1980; Palit et Sharp 1983). En Europe, les premiers organismes de recherche par sondage qui ont utilisé l'ITAO ont été le Social and Community Planning Research (SCPR, appelé aujourd'hui National Centre for Social Research) au Royaume-Uni (Sykes et Collins 1987) et l'Université d'état d'Utrecht, aux Pays Bas (Dekker et Dorn 1984). L'adoption des systèmes d'ITAO par les bureaux officiels de la statistique a été plus lente. Aux États-Unis, elle a débuté en 1982 au Census Bureau (Nicholls 1983) et au National Agricultural Statistics Service (Tortora 1985); parallèlement, elle a débuté la même année au bureau de la statistique des Pays Bas (1987). En 1987, selon un sondage mené auprès d'un échantillon (non probabiliste) de 27 organismes d'enquête (18 aux États-Unis et 9 ailleurs), presque tous utilisaient l'ITAO pour certaines ou toutes leurs enquêtes téléphoniques (Berry et O'Rourke 1988). Le rapport du Federal Committee on Statistical Methodology (1990) indique que le nombre estimatif d'installations d'ITAO dans le monde à la fin des années 1980 était supérieur à 1 000, et en 1988, le gouvernement américain collaborait avec 51 centres d'ITAO. Il convient de souligner que le développement de l'ITAO a fait rapidement part d'un mouvement plus général vers la collecte de données d'enquête assistée par ordinateur (CDEAO), qui englobe aussi l'interview sur place assistée par ordinateur (IPAO) et l'auto-interview assistée par ordinateur (AIAO) (Nicholls 1988). Pour un historique plus complet de la mise au point de l'ITAO et de la CDEAO en général, consulter Couper et Nicholls (1998).

3. PROGRÈS RÉCENTS DANS LE DOMAINE DES ENQUÊTES TÉLÉPHONIQUES

Durant le dernier quart de siècle, l'enquête téléphonique a définitivement fait ses preuves. Lyberg et Kasprzyk (1991) soutiennent qu'elle est devenue le mode principal de collecte de données dans les pays où la pénétration du téléphone est grande.

Des centaines d'articles scientifiques ayant trait à un large éventail d'aspects des enquêtes téléphoniques ont été publiés durant cette période. Plusieurs ouvrages généraux sur le sujet ont paru – Blankenship (1977a), Groves et Kahn (1979), Frey (1989) et Lavrakas (1993). Un certain nombre de conférences ont été consacrées à la méthodologie des enquêtes téléphoniques ou ont traité d'aspects particuliers de ces enquêtes. Les résultats ont été publiés dans des monographies ou dans des numéros spéciaux de revues scientifiques. En novembre 1987 s'est tenue, à Charlotte, en Caroline du Nord, une conférence importante sur la méthodologie des enquêtes téléphoniques qui a été suivie par la publication du compte rendu sous la direction de Groves, Biemer, Lyberg, Massey, Nicholls et Waksberg (1988) et d'un numéro spécial de la revue *Journal of Official Statistics*, sous la direction de Groves et Lyberg (1988b). La conférence sur la technologie des enquêtes assistées par ordinateur tenue à Berkeley au printemps 1981 (Freeman et Shanks 1983) portait principalement sur les enquêtes téléphoniques. L'ITAO a été le sujet principal de la conférence internationale sur la collecte de données d'enquête assistée par ordinateur InterCASIC 96 tenue à San Antonio, Texas, en décembre 1996 (Couper, Bethlehem, Baker, Clark, Martin, Nicholls et O'Reilly 1998) et de la troisième conférence internationale de l'ASC qui s'est tenue à Edinburgh en septembre 1999 (Banks, Christie, Currall, Francis, Harris, Lee, Martin, Payne et Westlake 1999).

Les auteurs susmentionnés fournissent des bibliographies détaillées comptant plusieurs centaines d'articles, de même que Khurshid et Sahai (1995), qui couvrent la période jusqu'à 1991, et *Survey Research Center* (2000), qui donne une mise à jour des bibliographies antérieures en ce qui concerne les plans d'échantillonnage pour les enquêtes téléphoniques auprès des ménages jusqu'à 2000.

La partie qui suit est consacrée à l'examen du développement des méthodes d'enquêtes téléphoniques auprès des ménages durant les 25 dernières années en se concentrant sur le plan d'échantillonnage, l'estimation, la couverture, la non-réponse et l'évaluation de la qualité des données.

3.1 Plan d'échantillonnage et estimation

La méthode d'échantillonnage appliquée aux enquêtes téléphoniques se fonde sur les principes généraux de l'échantillonnage. Son adaptation au contexte particulier des enquêtes téléphoniques a trait principalement à la base de sondage utilisée. Donc, nous adoptons la classification proposée par Lepkowski (1988) pour les méthodes d'échantillonnage téléphonique, selon la base de sondage utilisée – annuaire téléphonique et listes commerciales, composition aléatoire de numéros de téléphone (CA) et méthodes combinées (assistées par liste et base de sondage double).

3.1.1 Méthode d'échantillonnage axée sur une liste

Comme nous l'avons mentionné plus haut, les premières enquêtes téléphoniques étaient toutes réalisées auprès d'échantillons sélectionnés d'après des listes. Il s'agissait souvent d'enquêtes à mode mixte où l'on recourait à l'interview téléphonique pour compenser la non-réponse lors des interviews sur place ou pour réaliser un suivi. Ces scénarios dits d'« interview téléphonique à chaud » ont été appliqués à la Current Population Survey aux États-Unis et à

l'Enquête sur la population active au Canada (Drew, Choudhry et Hunter 1988). Dans ces cas, l'échantillonnage a pour base de sondage une liste générale à laquelle sont ajoutés des renseignements sur les numéros téléphoniques, et l'élaboration du plan de sondage ne tient compte d'aucune caractéristique particulière de l'usage du téléphone. Il en est de même pour les enquêtes téléphoniques « pures » réalisées auprès de populations particulières, comme les médecins, pour lesquelles on possède la liste complète des membres de la population avec les numéros de téléphone qui peut alors servir de base de sondage (voir, par exemple, Gunn et Rhodes 1981). Un autre exemple est celui où l'on recourt à l'interview téléphonique lors des cycles de suivi d'une enquête par panel pour laquelle le premier contact prend la forme d'une interview sur place. Par exemple, dans le cas de l'Enquête sur la population active d'Israël, la première prise de contact se fait lors d'une visite sur place et les deuxième et troisième cycles sont réalisés par téléphone auprès des ménages disposés à participer à l'enquête de cette façon (Nathan et Eliav 1988). Une méthode apparentée, utilisée récemment lors d'une étude pilote réalisée pour la US National Study of Health and Activity (Maffeo, Frey et Kalton 2000), consiste à tirer un échantillon régional, à obtenir dans la mesure du possible les numéros de téléphone en vue d'une interview téléphonique et de procéder à l'interview sur place pour ceux dont on n'a pas obtenu le numéro et pour ceux qui n'ont pas répondu à l'enquête téléphonique.

L'annuaire le plus facile à obtenir et le moins coûteux à utiliser comme base de sondage pour les enquêtes téléphoniques est, naturellement, l'annuaire téléphonique proprement dit, ou une version modifiée de cet annuaire. Au départ, on se servait de la version imprimée de l'annuaire, tandis qu'aujourd'hui, on se sert de la version électronique. Les principaux inconvénients de l'utilisation de l'annuaire téléphonique comme base de sondage, c'est-à-dire le sous-dénombrement, le surdénombrement, l'interview en double et le manque de données auxiliaires, ont été bien décrits. Le sous-dénombrement, qui est de loin le défaut le plus grave, inclut les ménages qui ne possèdent pas le téléphone, ainsi que ceux qui choisissent d'avoir un numéro de téléphone non publié ou ceux dont le numéro n'a pas encore été inséré dans l'annuaire. Nous traiterons à la section 3.3 du biais dû aux ménages qui ne possèdent pas le téléphone, lequel ne dépend naturellement pas de la base de sondage utilisée.

La proportion de numéros de téléphone non publiés varie considérablement selon le pays et le genre d'emplacement, ainsi que d'autres variables du ménage. Sykes et Collins (1987) publient un taux de numéros non publiés de 4 % pour les Pays-Bas et de 12 % pour le Royaume-Uni. Fréjean, Panzani et Tassi (1990) estiment à 14 % le taux de numéros non publiés en France et, aux État-Unis, durant les années 1970, on estimait que ce taux excédait 17 % à 19 % (Blankenship 1977b; Glasser et Metzger 1975). Selon Rich (1977), dans la région de la Californie desservie par Pacific Telephone, le taux de numéros de téléphone non publiés (excluant les numéros involontairement non publiés) est passé de 9 % en 1964 à 28 % en 1977. En outre, en Californie, environ 5 % de numéros de téléphone résidentiels seraient involontairement non publiés (attribués après la publication de l'annuaire). Selon des études plus récentes, le taux de numéros non publiés est considérablement plus élevé. Ainsi, d'après Genesys (1996), le taux national était de 40 % en 1993 et de 37 % en 1995, calculé d'après des échantillons nationaux de près de 100 000 interviews téléphoniques par CA, et selon Survey Sampling Inc. (1998), en 1997, le taux estimatif national de numéros non publiés aux États-Unis était de 30 %. Une étude à petite échelle réalisée dans la région de Jérusalem (Nathan et Aframian 1996) indique un taux de numéros non publiés de 27 %.

De nombreuses études révèlent des écarts importants entre les caractéristiques des ménages dont le numéro de téléphone est publié et de ceux dont le numéro ne l'est pas, différences qui pourraient être la source d'un biais de couverture en cas d'échantillonnage fondé sur l'annuaire téléphonique. Aux États-Unis, ces différences ont été mises en évidence, par exemple, par une étude de Brunner et Brunner (1971), qui ont observé des écarts hautement significatifs entre les ménages dont le numéro de téléphone est publié et ceux dont le numéro ne l'est pas, pour toute une série de variables démographiques et socioéconmiques. Leuthold et Scheele (1971) ont observé un taux plus élevé de non publication chez les noirs, les citadins, les jeunes, les personnes qui vivent en appartement, les personnes divorcées ou séparées et les travailleurs du secteur des services. Pareillement, Roslow et Roslow (1972) ont constaté des écarts significatifs entre les parts d'auditoire pour les ménages dont le numéro de téléphone est publié et ceux dont le numéro ne l'est pas. Glasser et Metzger (1975) ont montré que le taux de numéros non publiés était plus élevé dans l'Ouest, dans les grandes régions métropolitaines, chez les populations non blanches et chez les jeunes. Blankenship (1977b) et Rich (1977) ont noté des écarts hautement significatifs entre les ménages dont le numéro de téléphone est publié et ceux dont le numéro ne l'est pas en ce qui concerne le sexe et l'âge du chef de ménage, la profession, la taille du ménage et le revenu. Au Royaume-Uni, Sykes et Collins (1987) ont observé une plus forte proportion de numéros non publiés chez les jeunes, les personnes les plus pauvres et celles vivant à Londres. Pour la région de Jérusalem, Nathan et Aframian (1996) ont montré que les taux de ménages possédant un téléviseur et de ménages regardant la télévision (pour ceux qui possèdent un téléviseur) étaient plus faibles pour un échantillon sélectionné par CA que pour un échantillon sélectionné d'après un annuaire.

Outre le sous-dénombrement dû aux numéros non publiés, tels que décrits plus haut, les listes d'annuaire présentent les problèmes du surdénombrement, du compte multiple et du manque de renseignements auxiliaires à jour. Le surdénombrement a lieu lorsqu'une unité n'appartenant

pas à la population observée est incluse dans la base de sondage. Cette situation pourrait tenir au fait que, souvent, les numéros qui ne sont plus en service continuent de figurer dans l'annuaire, au fait que les numéros commerciaux ne sont pas toujours clairement désignés en tant que tels ou à d'autres cas non reconnus d'inadmissibilité. Le compte multiple a lieu lorsque la même unité est représentée plus d'une fois dans la base de sondage et que ce fait n'est pas reconnu. On peut habituellement repérer les comptes multiples durant l'échantillonnage si les entrées pour un même ménage sont énumérées consécutivement, mais on ne le peut pas si elles apparaissent séparément (par exemple, sous divers noms de famille). Si le compte multiple est confirmé durant l'interview (par exemple, en obtenant des renseignements sur le nombre de lignes raccordées dont dispose le ménage ou le nombre de numéros publiés dans l'annuaire) l'application de coefficients de pondération appropriés permet de résoudre le problème. Alors que le surdénombrement est surmontable, à un certain prix, le sous-dénombrement ne l'est pas, ce qui indique qu'il faut recourir à des bases de sondage plus représentatives que ne le sont les annuaires. Les listes produites par les entreprises commerciales, ordinairement pour des raisons de marketing, sont devenues une solution fréquente pour remplacer le traditionnel annuaire téléphonique (produit en général par la compagnie qui fournit le service téléphonique dans la région). Ces listes peuvent correspondre aux annuaires de la ville, être produites d'après les listes municipales d'adresses et complétées par les numéros provenant d'annuaires ou d'autres sources, les listes d'abonnés des compagnies de téléphone ou les listes nationales principales d'adresses, telles que celles produites par Donelley Marketing, Inc. aux États-Unis (Lepkowski 1988). Ces listes fournissent des données auxiliaires importantes, comme les données géographiques provenant du Recensement de la population et du logement et d'autres sources. En général, elles ne permettent pas d'éliminer le biais dû aux numéros non publiés et leur coût peut être élevé. Leur utilisation améliore parfois la variance d'échantillonnage, grâce aux données auxiliaires qui permettent d'élaborer un plan de sondage plus efficace. Éventuellement, les listes utilisées par les services d'urgence pour déterminer l'emplacement physique des personnes qui appellent pourraient servir de bases de sondage, mais les organismes d'enquêtes non gouvernementaux auraient de la difficulté à les obtenir.

3.1.2 Composition aléatoire – La méthode de Mitofsky-Waksberg

L'application de la composition aléatoire (CA) pour réaliser les enquêtes téléphoniques est devenue un moyen très populaire, particulièrement aux États-Unis. Ces méthodes de CA se fondent sur une base de sondage qui englobe tous les numéros de téléphone possibles. L'idée a été proposée au départ par Cooper (1964), qui a ajouté des suffixes aléatoires à quatre chiffres à des préfixes connus pour réaliser une enquête locale. Par la suite, Eastlack et Assael (1966) et Glasser et Metzger (1972) ont amélioré cette méthode d'échantillonnage de base et lui ont donné une portée nationale grâce à la création de « banques valides » de numéros d'après les renseignements fournis par les compagnies de téléphone.

Jusqu'à récemment, la CA n'était utilisée, en grande partie, qu'aux États-Unis et au Canada. En effet, Sykes et Collins (1987) indiquent que les enquêtes téléphoniques étaient encore rares au Royaume-Uni à la fin des années 1980, principalement à cause de la faible pénétration du téléphone. Plus précisément, on recourait rarement à la CA, notamment à cause du manque d'uniformité de la longueur des numéros de téléphone à l'époque. Toutefois, récemment, étant donné la plus grande pénétration du téléphone au Royaume-Uni, de l'ordre de 96 % à la fin des années 1990, et la normalisation des numéros de téléphone qui comptent maintenant dix chiffres, les enquêtes par CA sont plus fréquentes – à cet égard, consulter, par exemple, Collins (1999) et Nicolaas, Lynn et Lound (2000). Pareillement, Gabler et Haeder (2000) indiquent qu'une méthode par CA, modifiée pour tenir compte de la longueur variable des numéros de téléphone (de 6 à 11 chiffres) est désormais la méthode normalisée d'enquête téléphonique en Allemagne.

Mitofsky (1970) a proposé le premier une méthode d'échantillonnage à deux degrés par CA pour parer à l'inefficacité des méthodes de CA élémentaires due au fait qu'il faut composer un grand nombre de numéros qui ne produisent pas d'interview (numéros non en service et numéros non résidentiels). Cette méthode, que Waksberg (1978) a élaborée par la suite et à laquelle il a donné un fondement théorique solide, a été appelée méthode de Mitofsky-Waksberg. Cette dernière ou ses variantes sont devenues les principales méthodes d'échantillonnage appliquées aux enquêtes téléphoniques, du moins aux États-Unis.

Elle se fonde sur le fait que les numéros de téléphone résidentiels sont, en général, regroupés en séries de numéros consécutifs ou dans des banques de numéros ayant les mêmes r premiers chiffres. Pour les États-Unis, la valeur de r est habituellement fixée à huit (pour les numéros de téléphone à 10 chiffres, y compris l'indicatif régional), de sorte que la taille de chaque banque ou grappe (UPE-unité primaire d'échantillonnage) soit $N = 100$. On suppose que la compagnie de téléphone peut fournir la liste de tous les préfixes opérationnels (indicatif régional plus les trois premiers chiffres du numéro), c'est-à-dire ceux auxquels a été attribué un numéro résidentiel. On ajoute à chaque numéro à six chiffres de cette liste toutes les combinaisons possibles de deux chiffres, afin d'obtenir une base de sondage de numéros à huit chiffres qui représente les M UPE qui forment la population. Les UPE de l'échantillon sont sélectionnées au hasard, de façon consécutive, à partir de cette base de sondage (avec remise) et, pour chaque UPE

sélectionnée, deux chiffres finals sont sélectionnés au hasard. Le numéro à 10 chiffres résultant est composé et, s'il n'est pas celui d'une résidence (conformément à la définition adoptée pour l'enquête), l'UPE est supprimée de l'échantillon. S'il s'agit d'une résidence, on sélectionne un échantillon aléatoire simple (sans remise) de k numéros résidentiels supplémentaires en composant des numéros sélectionnés au hasard (sans remise) à partir de l'UPE, jusqu'à ce que l'on ait obtenu k numéros résidentiels supplémentaires. La méthode de sélection des UPE se poursuit jusqu'à ce que l'on ait sélectionné un nombre établi, m, d'UPE. Il est facile de voir, si l'on suppose que le nombre de numéros résidentiels dans chaque UPE sélectionné, P_i, est au moins égal à k, que l'effectif total de l'échantillon de ménages ayant un numéro de téléphone résidentiel est $m\ (k+1)$ et que l'échantillon final est un échantillon de la population complète de ménages ayant un numéro de téléphone résidentiel sélectionnés avec probabilité égale.

Waksberg (1978) montre que, si nous représentons par $\pi = (\Sigma_{i=1}^{M} P_i)/(NM)$ la proportion de numéros résidentiels dans la population et par t la proportion d'UPE sans numéro résidentiel (c'est-à-dire pour lesquelles $P_i = 0$), alors le nombre prévu total d'appels est donné par $m[1 + (1-t)k]/\pi$, en supposant que $P_i \geq k+1$ pour toutes les UPE qui comptent au moins un numéro résidentiel. On peut abandonner la dernière hypothèse si l'on regroupe les UPE de sorte que la contrainte soit vérifiée pour chaque groupe ou si l'on applique des coefficients de pondération inégaux. On obtient les valeurs optimales des paramètres du plan de sondage dans le cas d'une fonction simple de coût et on étend la méthode au traitement d'enquêtes répétées. L'avantage principal de la méthode tient à la réduction du nombre prévu d'appels qu'il faut faire pour obtenir une taille donnée, efficace, d'échantillon, particulièrement si t, c'est-à-dire la proportion d'UPE sans numéro de téléphone résidentiel, est supérieur à 0,5. Groves (1977) publie, pour une étude nationale, des données indiquant que la valeur de t est d'environ 0,65. Il faut mettre en balance cet avantage et l'augmentation de la variance due à la mise en grappe. Néanmoins, si l'on tient compte des coûts, des calculs illustratifs pour des valeurs types des paramètres montrent que la réduction des coûts est de l'ordre de 20 % à 40 %.

Le principal inconvénient opérationnel de la méthode est dû à sa nature séquentielle qui rend difficile son application manuelle. Par contre, l'opération séquentielle ne pose aucun problème si le procédé de sélection est entièrement automatisé. Telle qu'elle est décrite plus haut, la méthode pose d'autres problèmes, que, dans la plupart des cas, de simples modifications permettent de surmonter. Si l'on suppose que l'on ne possède aucun renseignement a priori sur le nombre de ménages possédant un numéro de téléphone, on ne connaît pas les probabilités de sélection, mais on peut estimer la valeur de p à partir de l'échantillon. La nécessité pratique d'introduire une règle d'arrêt pour limiter le nombre d'appels sans répondant ou pour lesquels le répondant refuse de répondre, même s'il s'agit d'un numéro résidentiel, sous-entend que l'on ne peut appliquer la méthode strictement telle qu'elle est conçue, ce qui pourrait biaiser les résultats. Il est possible de contourner le problème des ménages qui possèdent plusieurs numéros de téléphone si l'on obtient des renseignements corrects sur le nombre de lignes distinctes, mais la pondération corrective qui est nécessaire réduit la simplicité de la pondération uniforme. Dans certains cas, il est possible d'obtenir les noms et les adresses par appariement des numéros obtenus par CA à des listes d'adresses, si bien que l'on peut envoyer un avis préalable à une partie des répondants éventuels. Cependant, la procédure est complexe et les difficultés que pose l'envoi d'un avis aux répondants (communes à toutes les méthodes de CA) rendent sa prise en considération difficile pour certains bureaux officiels de la statistique.

3.1.3 Modifications de la méthode de Mitofsky-Waksberg et d'autres méthodes de compositions aléatoires

Certains inconvénients de la méthode de base peuvent être surmontés grâce à la généralisation proposée par Potthoff (1987a, 1987b). La méthode se fonde sur la définition d'un ensemble de numéros de téléphone favorables. Il pourrait s'agir uniquement de numéros résidentiels, comme dans le cas de la méthode de Mitofsky-Waksberg, ou d'un ensemble plus général qui inclut tous les numéros résidentiels – par exemple, l'ensemble de numéros pour lesquels on obtient une tonalité (y compris le signal occupé, les messages enregistrés et les téléphonistes). La première étape correspond à un échantillonnage aléatoire simple d'un nombre déterminé, m, d'UPE. À partir de chaque UPE sélectionnée, on effectue un nombre déterminé d'appels, c, et, pour chaque appel, on détermine si le numéro est favorable ou non. Toute UPE pour laquelle les c numéros sélectionnés sont défavorables est rejetée. Les UPE retenues sont considérées comme étant de type I si elles ne comportent qu'un seul numéro favorable et de type II si elles comptent au moins deux numéros favorables. La deuxième étape consiste à sélectionner et à composer kc numéros provenant des UPE de type I et $k(c-1)$ numéros provenant des UPE de type II, où k est un nombre entier. Pour chaque numéro composé, on détermine si l'unité est résidentielle ou hors du champ d'observation, et l'on s'efforce de réaliser une interview dans le cas de toutes les unités résidentielles. Pour les UPE de type I, un segment séquentiel supplémentaire sélectionne des numéros de téléphone supplémentaires qui sont composés jusqu'à ce que l'on obtienne un total de k numéros favorables. Une tentative d'interview est faite pour chaque numéro favorable composé dans le cadre du segment séquentiel. Potthoff (1987a) montre que, dans certaines conditions, tous les numéros de téléphone résidentiels ont la même probabilité de sélection et développe des estimations non biaisées et des rapports estimatifs, ainsi que leurs variances.

Il présente aussi des comparaisons de coûts et certaines modifications qui permettent de surmonter des problèmes pratiques. Sa méthode réduit l'incertitude quant à la catégorie de numéro composé en cas de non-réponse, ainsi que le problème de l'épuisement des numéros résidentiels dans une UPE.

Un grand nombre de généralisations et de modifications supplémentaires de la méthode fondamentale de Mitofsky-Waksberg ont été proposées. Plusieurs visent à réduire le fardeau du filtrage des interviews et à améliorer le contrôle de la taille de l'échantillon auprès duquel a lieu le premier contact. Ainsi, Hogue et Chapman (1984) proposent de déterminer des nombres limites d'appels en se fondant sur une estimation de la probabilité qu'une UPE soit « peu peuplée », c'est-à-dire qu'elle ne contient qu'une faible proportion de numéros résidentiels, et d'établir une méthode optimale d'établissement du nombre limite en tenant compte du coût et de la variance. Alexander (1988) considère deux types de règles d'exclusion pour limiter le filtrage en vue d'une interview pour les préfixes pour lesquels la densité de numéros résidentiels est faible. Une « règle de croissance » met un terme au processus aussitôt qu'un nombre préétabli d'appels, c_i, ont été faits et que moins de i résidences ont été dépistées, où $\{c_i\}$ est une série croissante dans i. Une « règle de décroissance » met un terme au processus quand i résidences ont été dépistées et au moins c_i appels ont été faits, où $\{c_i\}$ est une série décroissante dans i. Les coûts de l'application de ces règles sont évalués dans le cas d'un modèle simple.

Lepkowski et Groves (1986a) proposent un plan d'échantillonnage à deux phases fondé sur l'appariement des préfixes sélectionnés à la première étape de la méthode de Mitofsky-Waksberg aux numéros d'un annuaire commercial pour obtenir le nombre de numéros de téléphones publiés pour chaque préfixe sélectionné. Puis, les préfixes sont répartis en deux strates – une strate à densité faible qui ne contient aucun numéro publié, ou uniquement un petit nombre de ces numéros, et une strate à densité élevée. Puis, on applique le plan d'échantillonnage de Mitofsky-Waksberg à la strate à faible densité et on sélectionne les numéros de téléphone avec probabilité proportionnelle au nombre de numéros de téléphone publiés dans la strate à forte densité.

Brick et Waksberg (1991) proposent d'utiliser un nombre fixé de numéros de téléphone à la deuxième étape, afin d'éviter tout bonnement l'échantillonnage séquentiel, ce qui simplifie l'opération. Le plan de sondage, proposé au départ par Waksberg (1984), n'est toutefois pas autopondéré et produit un léger biais et une augmentation de la variance. Brick et Waksberg (1991) énoncent des points à prendre en considération pour faire un choix entre les plans original et modifié de Mitofsky-Waksberg. Pour l'une des premières applications de la méthode modifiée de Mitofsky-Waksberg à la collecte de données sur les attitudes à l'égard de la santé, qui semble résulter d'un effort erroné d'application de la méthode originale, consulter Cummings (1979). Smith et Frazier (1993) comparent les méthodes originale et modifiée, au moyen de données recueillies dans le cadre du California Behavioral Risk Factor Surveillance System. Les résultas indiquent que la méthode modifiée accélère la collecte des données, ce qui produit un échantillon de plus grande taille au même prix. Cet avantage compense les effets plus importants du plan de sondage de cette méthode.

Le recours à la stratification et à la répartition non proportionnelle pour améliorer les « taux d'appels couronnés de succès», proposés par Palit (1983) est une autre variante de la méthode de base de Mitofsky-Waksberg. Palit et Blair (1986) procèdent à une évaluation des autres traitements des numéros de téléphone ne produisant pas de réponse applicables au plan de sondage de Mitofsky- Waksberg. Burke, Morganstein et Schwartz (1981) étudient la détermination optimale des paramètres pour la méthode de Mitofsky-Waksberg et Casady et Lepkowski (1991, 1993), ainsi que Tucker, Casady et Lepkowski (1992) s'intéressent à la répartition optimale pour la version stratifiée. D'autres problèmes ayant trait à la répartition produisant le coût minimal sont examinés par Palit (1983) et par Mason et Immerman (1988).

3.1.4 Méthodes assistées par liste

Bien que les méthodes de CA permettent d'éviter le sous-dénombrement inhérent aux annuaires à cause des numéros non publiés, elle continue de poser le problème fondamental du sous-dénombrement dû aux ménages non abonnés au téléphone (pour plus de détails, consulter la section 3.3). En outre, le manque de données auxiliaires (dont les données géographiques), que fournissent fréquemment les listes, cause des inefficacités, même dans le cas des modifications les plus perfectionnées de la méthode fondamentale exposées plus haut. Par conséquent, on a recherché d'autres méthodes en vue de combiner les échantillons obtenus par CA à des échantillons obtenus d'après des listes et des annuaires. L'un des premiers efforts dans cette direction, qui a été proposé par Stock (1962) et élaboré par Sudman (1973), se fonde sur le remplacement des deux derniers chiffres des numéros de téléphone sélectionnés à partir d'un annuaire par des chiffres sélectionnés au hasard. Hauck et Cox (1974) ont appliqué la méthode pour procéder à une étude méthodologique des effets du mode de sélection lors du filtrage d'une sous-population particulière. Une version simplifiée, appelée communément méthode de l'« unité en plus », consiste à remplacer chaque numéro de téléphone échantillonné d'après un annuaire par le numéro obtenu en ajoutant une unité (ou un autre nombre, ce qui donne alors la méthode du « nombre en plus»). En principe, cette mesure permet d'éviter le biais dû aux numéros non publiés. Étant donné sa simplicité, elle a eu beaucoup de succès auprès des spécialistes des études de marché. Cependant, plusieurs

études, dont celles de Landon et Banks (1977) et de Mullet (1982), montrent qu'en fait, cette méthode n'est pas dépourvue de biais et que son efficacité est faible.

Forsman et Danielsson (1997) proposent une méthode axée sur un modèle pour l'échantillonnage par la méthode du «nombre en plus», fondée sur l'hypothèse qu'un préfixe comprend un mélange aléatoire de numéros publiés et non publiés. Le modèle, qui a été testé empiriquement, fournit des estimations non biaisées. Ghosh (1984) a proposé une méthode améliorée qui consiste à continuer d'ajouter une unité au dernier numéro de téléphone composé aussi longtemps qu'on n'a pas rejoint un ménage et d'arrêter aussitôt que l'on en rejoint un. Quoi qu'il persiste un biais, celui-ci est plus faible que celui que produit la simple méthode de l'« unité en plus ». D'autres méthodes assistées par liste avec composantes CA sont examinées par Potter, McNeill, Williams et Waitman (1991) qui stratifient les préfixes d'après les dénombrements de numéros de téléphone publiés, tout en s'assurant que soient inclus des blocs ne contenant aucun numéro publié.

Brick, Waksberg, Kulp et Starer (1995) proposent une méthode assistée par liste qui permet de surmonter le problème gênant de la nature séquentielle de l'échantillonnage de deuxième phase du scénario de Mitofsky-Waksberg. La méthode consiste à répartir le fichier des préfixes de central (banques 100) en deux strates. La première comprend tous les préfixes de central qui comptent au moins un numéro de téléphone résidentiel publié et la deuxième, ceux qui n'en contiennent aucun. Limiter l'échantillonnage à la première strate réduit spectaculairement la proportion de numéros non résidentiels qu'il faut composer, mais cause un biais de couverture. Ils étudient le biais et concluent que ces méthodes d'échantillonnage tronqué sont efficaces et présentent des avantages opérationnels, tandis que le biais de couverture résultant (environ 4 %) n'est pas très important. La méthode a été utilisés à grande échelle pour remplacer la méthode classique de Mitofsky-Waksberg. Statistique Canada l'applique depuis 1991 à l'Enquête sociale générale pour la sélection de l'échantillon complet, avec échantillonnage aléatoire simple dans les banques de numéros qui contiennent au moins un numéro résidentiel (Norris et Paton 1991). La modification de ce plan de sondage inclut la stratification complète des banques de numéros d'après les renseignements fournis par des listes, l'utilisation de la CA simple pour les strates contenant une faible proportion de banques pour lesquelles aucun numéro résidentiel n'est publié et l'application de la méthode de Mitofsky-Waksberg aux strates restantes. Casady et Lepkowski (1993) comparent ce plan de sondage à d'autres plans stratifiés en se servant d'un modèle de coûts. Selon leur étude, pour des rapports de coûts faibles (de sélections productives aux sélections improductives), les plans de sondage par CA à deux et à trois strates sont aussi efficaces que la méthode de Mitofsky-Waksberg et, pour les rapports de coûts élevés, ils sont supérieurs.

3.1.5 Plans à plusieurs bases de sondage

En vue d'éliminer certains biais de sous-dénombrement inhérents aux enquêtes téléphoniques dont la base de sondage est un annuaire ou des numéros de téléphone, on a accordé de plus en plus d'attention aux enquêtes à base de sondage double et à mode mixte combinant l'interview par téléphone et l'interview sur place. Ces enquêtes allient l'échantillonnage classique pour les interviews sur place à l'échantillonnage par CA ou d'après un annuaire pour les interviews téléphoniques. Biemer (1983) a étudié la composition optimale pour ce genre de plans de sondage grâce à une étude en simulation, et McCarthy et Bateman (1988) proposent de recourir à la programmation mathématique pour atteindre une répartition optimale des unités d'échantillonnage pour un plan à base de sondage double qui permet l'analyse a postériori des effets des variations du plan de sondage et des paramètres de coût sur l'optimisation. Choudhry (1989) propose une optimisation variable selon le coût pour estimer les proportions et Brick (1990) propose pour cela de recourir à l'échantillonnage multiple. Dans une série d'articles, Groves et Lepkowski (1985, 1986), Lepkowski et Groves (1984, 1986b) et Traugott, Groves et Lepkowski (1987) mettent au point des modèles d'erreur pour ces plans d'enquête à base de sondage double. Ils présentent aussi les résultats d'expériences visant à comparer les taux de réponse et les biais éventuels pour l'échantillonnage par CA ou d'après une liste et pour plusieurs méthodes d'interview. Les résultats ont été appliqués aux États-Unis à la grande enquête National Crime Survey.

Whitmore, Mason et Hartwell (1985) décrivent l'application de méthodes à base de sondage et mode doubles lors d'une étude de l'exposition personnelle au monoxyde de carbone dans deux régions métropolitaines parrainée par la US Environment Protection Agency et d'une étude à l'échelle des États des besoins en services sociaux. Dans les deux cas, ils se sont servis de listes d'annuaire disponibles dans le commerce que l'on a combinées à l'échantillonnage régional des ménages. D'après l'analyse de leurs résultats, ils recommandent l'utilisation de plans à base de sondage double afin de profiter de l'efficacité relative de l'interview téléphonique et d'éliminer les biais inhérents à l'utilisation d'annuaires comme base de sondage. Waksberg, Brick, Shapiro, Flores-Cervantes et Bell (1997) décrivent l'utilisation combinée de l'échantillonnage par CA et de l'échantillonnage aréolaire pour la US National Survey of America's families durant laquelle on s'est concentré, en particulier, sur la population à faible revenu. On a remis aux ménages ne possédant pas le téléphone repérés durant le dépistage aréolaire des téléphones cellulaires pour leur permettre de répondre aux interviews téléphoniques, donc de ne pas devoir former les intervieweurs chargés des questions de filtrage dans le cas d'un questionnaire non téléphonique (Cunningham, Berlin, Meader, Molloy, Moore et Pajunen 1997).

3.2 Autres problèmes d'échantillonnage

3.2.1 Échantillonnage de populations spéciales

Le coût relativement faible de l'interview téléphonique a fait de celle-ci un moyen de premier choix de filtrer les grands échantillons en vue de repérer de petites populations particulières. Ainsi, Sudman (1978) discute des conditions dans lesquelles l'utilisation d'un échantillon téléphonique pour la sélection par tri d'un sous-groupe dont les membres seront, en bout de ligne, interviewés sur place est plus efficace que le filtrage sur place. L'analyse des fonctions de coût montre que le filtrage téléphonique est efficace, sauf si l'homogénéité intragrappe est faible, la densité d'interview est faible et(ou) les coûts de dépistage et de filtrage sont faibles comparativement aux coûts de l'interview. Blair et Czaja (1982) proposent une modification de la méthode de Mitofsky-Waksberg pour repérer les populations spéciales géographiquement regroupées et décrivent une application à la population noire. Toutefois, comme le fait remarquer Waksberg (1983), quand les grappes sont épuisées, cette méthode nécessite une repondération qui pourrait réduire son efficacité. Autrement dit, la méthode pourrait être efficace pour la population noire, mais ne pas l'être nécessairement pour d'autres minorités. Un autre plan de sondage téléphonique applicable à la population noire des États-Unis est proposée par Inglis, Groves et Heeringa (1987). Mohadjer (1988) propose un plan d'échantillonnage par CA avec stratification des zones de préfixe. Hartge, Brinton, Rosenthal, Cahill, Hoover et Waksberg (1984) décrivent la sélection de ménages par la méthode de Mitofsky-Waksberg conjuguée à l'échantillonnage stratifié de personnes dans les ménages pour sélectionner un groupe témoin représentatif de la population dans le cas de quatre études épidémiologiques. Une simulation de l'échantillonnage aléatoire simple en vue de déterminer l'efficacité de la méthode porte Perneger, Myers, Klag et Whelton (1993), à conclure que celle-ci est efficace.

Les enquêtes aréolaires sont un autre exemple de population spéciale que l'enquête téléphonique permet de traiter efficacement. Même si, en général, les centraux téléphoniques ne coïncident pas avec des régions géographiques exactes, le degré de correspondance est important et, si l'on procède à une sélection par tri des unités comprises dans la zone définie, l'interview téléphonique permet de réduire considérablement les coûts. Par exemple, Banks et Hagan (1984) décrivent la réduction du filtrage que doivent réaliser les intervieweurs grâce à la combinaison de l'échantillonnage à partir d'une liste et de l'échantillonnage par CA lors d'une enquête visant à évaluer l'efficacité des programmes de santé dans des zones de services particulières. Pareillement, Campbell et Palit (1988) testent une combinaison d'échantillonnage par liste et de composition totale (CT) en se servant comme base de sondage de tous les numéros couverts par les circonscriptions téléphoniques correspondant à une région de recensement donnée et constatent que cette méthode réduit considérablement les coûts de dénombrement comparativement à l'interview sur place.

3.2.2 Échantillonnage de personnes dans les ménages

Presque toutes les enquêtes-ménages comprennent des questions sur des personnes qui vivent dans le ménage. Dans certains cas, tous les membres du ménage sont inclus dans l'échantillon, mais souvent, pour diverses raisons, on sélectionne un ou plusieurs membres du ménage auxquels on demande de répondre personnellement à certaines questions. La méthode classique de Kish (Kish 1949), utilisée principalement lors des enquêtes par interview sur place, pose des problèmes particuliers dans le cas des enquêtes téléphoniques, car elle oblige à obtenir la liste complète des membres du ménage par téléphone. Ces renseignements sont plus difficiles à obtenir par téléphone que lors d'une interview sur place durant laquelle certaines personnes peuvent être présentes. Il convient toutefois de souligner que, dans de nombreux cas, il faut de toute façon recueillir les renseignements sur la composition du ménage. En outre, la manipulation des règles de sélection par l'intervieweur (par exemple, pour obtenir un taux de réponse élevé) soupçonnée depuis longtemps dans le cas des interviews sur place, est presque impossible dans le cas des enquêtes par ITAO (pour lesquelles l'intervieweur ne peut visualiser la sélection).

Troldahl et Carter (1964) proposent une méthode qui n'oblige à déterminer que le nombre de personnes de chaque sexe dans le ménage. Puis, des règles probabilistes (par exemple, l'« homme le plus âgé ») sont appliquées pour sélectionner la personne qui répondra aux questions, de sorte que l'on connaît les probabilités de sélection de chaque personne. Cependant, la probabilité de sélection n'est pas positive pour tous les membres du ménage (par exemple, dans les ménages comptant trois hommes, celui d'âge intermédiaire n'est jamais sélectionné). La méthode (appelée « méthode de Troldahl-Carter ») a été modifiée par Bryant (1975) afin de tenir compte du fait que des ménages peuvent compter plus de deux personnes de même sexe. Une variante de la méthode, proposée par Salmon et Nichols (1983) et par O'Rourke et Blair (1983), consiste à sélectionner la personne qui sera la prochaine (ou a été la dernière) à célébrer son anniversaire (méthode du « prochain anniversaire» ou du «dernier anniversaire »), pour s'assurer que la probabilité de sélection soit la même pour tous les membres du ménage, en supposant que la date de l'interview soit aléatoire. Évidemment, cette hypothèse est raisonnable dans le cas des enquêtes réalisées sur une période de 12 mois, mais ne l'est pas pour les enquêtes dont la période d'interview est plus courte. Ce facteur et d'autres pourraient donner lieu à une corrélation entre les probabilités de sélection et les caractéristiques individuelles. Une autre méthode de sélection proposée par Hagan et Meier (1983), ne nécessite aucune donnée préliminaire sur la composition du ménage et consiste à sélectionner une

personne prédéterminée (par exemple, « l'homme le plus âgé »). De nouveau, cette méthode ne permet pas d'assurer que la probabilité de sélection soit positive pour tous les membres du ménage.

Les méthodes susmentionnées ont fait l'objet de plusieurs comparaisons empiriques. Czaja, Blair et Sebestik (1982) ne constatent aucune différence significative entre les taux de réponse ni les profils démographiques des deux versions de la méthode de Troldahl-Carter et de la méthode de Kish. Hagan et Meier (1983) comparent leur méthode, décrite plus haut, à celle de Troldahl-Carter et constatent que la méthode qu'ils proposent produit un taux de refus nettement plus faible, sans différence significative entre les profils démographiques. Salmon et Nichols (1983) comparent quatre méthodes de sélection des répondants dans un ménage – méthodes de Troldahl-Carter, d'alternance homme/femme, du prochain anniversaire et de sélection nulle, dans le cadre d'une enquête téléphonique auprès d'un petit échantillon. Ils concluent que la méthode du prochain anniversaire est assez efficace pour sélectionner un échantillon représentatif des membres du ménage. Oldendick, Bishop, Sorenson et Tuchfarber (1988) ne constatent aucune différence significative entre la méthode de Kish et celle du dernier anniversaire. Lors d'une étude fondée sur la méthode du dernier anniversaire, Romuald et Haggard (1994) observent que le taux d'autosélection des personnes bien informées en vue de participer à l'enquête est plus élevé que prévu. Ils étudient l'effet qu'ont les indices utilisés pour raviver la mémoire sur l'autosélection du répondant et concluent que cet effet n'est pas significatif. Lavrakas, Bauman et Merkle (1993) évaluent l'effet de l'utilisation de la méthode du dernier anniversaire sur la couverture à l'intérieur d'une unité dans le cas d'une enquête nationale et présentent des données qui donnent à penser que la méthode aboutit à une sélection incorrecte dans de nombreux cas. Forsman (1993) passe en revue les expériences d'échantillonnage dans les ménages réalisées par 18 sociétés privées spécialisées dans les sondages d'opinion et décrivent un test permettant de comparer les méthodes de Kish, du prochain/dernier anniversaire et de Toldahl-Carter. Ils concluent que cette dernière méthode est un peu meilleure que la méthode de Kish et que toutes deux sont supérieures à la méthode des anniversaires. Pareillement, Binson, Canchola et Catania (2000) décrivent, pour une enquête téléphonique nationale, une comparaison triple entre les méthodes de Kish, du prochain anniversaire et du dernier anniversaire et notent des écarts significatifs entre les taux d'abandon enregistrés pour les trois méthodes aux premiers stades du processus de filtrage. La méthode de Kish produit le taux d'abandon le plus élevé et celle du « prochain anniversaire », le taux le plus faible. Ils supposent que les intervieweurs, plutôt que les répondants, sont la source principale du taux élevé de refus dans le cas de la méthode de Kish, puisque cette dernière nécessite la liste complète des membres du ménage.

3.3 Couverture et non-réponse

3.3.1 Couverture téléphonique

Jusqu'à récemment, le problème de la non-couverture téléphonique était un inconvénient important des enquêtes par téléphone. Même aux États-Unis, le sous- dénombrement des personnes (dans les ménages ne possédant pas le téléphone) était encore de 7,2 % à la fin de 1986 (Thornberry et Massey 1988). Au milieu des années 1980, le sous-dénombrement téléphonique des ménages était inférieur à 10 % dans la plupart des pays occidentaux, le taux de couverture le plus élevé (99 %) étant celui observé en Suède. Néanmoins, le taux de sous-dénombrement téléphonique demeurait élevé dans certains pays, comme le Royaume-Uni (25 %), l'Italie (29 %), l'Irlande (50 %) et Israël (30 %) (Trewin et Lee, 1988). À la fin du siècle, la situation avait changé spectaculairement, puisque, dans la plupart des pays occidentaux, la couverture téléphonique atteignait virtuellement la saturation. Elle était de 94,4 % aux États-Unis en 1999 (NTIA 2000), de 96,6 % en Australie en 1996 (St. Clair et Muir 1997), de 97,0 % au Royaume-Uni (OFTEL 1999), de 97,3 % en Israël (Bureau central de la statistique 2000) de 97,9 % en Finlande (Kuusela et Vikki 1999), de 98,2 % au Canada (Statistique Canada 1999) et de 99 % en Allemange (République fédérale d'Allemagne 1999).

De toute évidence, le principal problème que pose le sous-dénombrement téléphonique tient davantage au sous-dénombrement différentiel qu'au sous-dénombrement global, et au fait que le sous-dénombrement téléphonique est fortement corrélé au large éventail de variables démographiques, économiques et de santé. Cette situation a été confirmée par un grand nombre d'études empiriques menées aux États-Unis et ailleurs – à cet égard, consulter par exemple Groves et Kahn (1979), Collins (1983, 1999), Thornberry et Massey (1983, 1988), Trewin et Lee (1988), ainsi que Botman et Allen (1990). L'augmentation rapide de la couverture téléphonique globale au cours de la dernière décennie n'a pas modifié radicalement cette situation. Ainsi, en Finlande, où le sous-dénombrement téléphonique global était de 2,1 % en 1999, le sous-dénombrement des ménages à faible revenu (moins de 675 Euros par mois) était de 11,3 % (contre 0 % pour les ménages à revenu élevé) et celui des ménages vivant dans un logement loué, de 4,9 % (Kuusela et Vikki 1999). En Israël, le sous-dénombrement téléphonique était de 17,9 % pour le décile inférieur de revenu comparativement à 0,8 % pour les deux déciles supérieurs et de 24,9 % pour les ménages composés d'un seul adulte et d'au moins trois enfants comparativement à 2,4 % pour les ménages sans enfant comptant au moins trois adultes (Bureau central de la statistique 2000). Pareillement, aux États-Unis, les variations géographiques sont importantes et le sous-dénombrement téléphonique est corrélé au problème de logement, à la race, au niveau de scolarité, au revenu et à la mobilité (Shapiro, Battaglia, Hoaglin, Buckley et Massey 1996; Giesbrecht, Kulp et

Starer 1996; Fox et Riley 1996; NTIA 2000). Selon Anderson, Nelson et Wilson (1998), qui se sont fondés sur les données de la National Health Interview Survey, et selon Ford (1998) qui s'est fondé sur les données de la National Health and Nutrition Examination Survey, les caractéristiques liées à la santé des personnes qui vivent dans un ménage ayant le téléphone diffèrent quelque peu de celles des personnes qui vivent dans un ménage n'ayant pas le téléphone. Cependant, les deux études mènent à la conclusion que les effets de la couverture téléphonique sont faibles.

Par contre, le problème principal que pourrait poser la couverture téléphonique dans un avenir proche a trait au lancement et à la prolifération rapide de la téléphonie mobile. À la fin des années 1990, la proportion de ménages ayant accès à au moins un téléphone mobile atteignait 76 % en Finlande, 59 % au Danemark, 35 % en Italie (Rouquette 2000) et 52 % en Israël (Bureau central de la statistique 2000). La situation ne poserait aucun problème si ces téléphones mobiles étaient simplement complémentaires du service fixe existant. Toutefois, on possède déjà des preuves convaincantes d'une tendance, dans plusieurs pays, à considérer le téléphone mobile comme un remplacement, plutôt qu'un complément, du service téléphonique fixe. Selon Kuusela et Vikki (1999), en Finlande, 20 % des ménages possèdent maintenant uniquement un ou plusieurs téléphones mobiles, mais ne possèdent aucun téléphone fixe et dans un an, le nombre de téléphones mobiles surpassera le nombre de lignes téléphoniques fixes. Comparativement, les chiffres sont de 3 % pour le Royaume-Uni (OFTEL 2000) et de 2,9 % pour Israël (Bureau central de la statistique 2000). Autrement dit, la couverture des services téléphonique fixes a diminué pour s'établir à 77 % en Finlande et à 94 % au Royaume-Uni et en Israël. En Allemagne, on estime que la proportion de ménages abonnés à un service téléphonique fixe diminuera pour s'établir à 92 % d'ici à 2004 (Gabler et Haeder 2000). De surcroît, les caractéristiques des personnes qui possèdent uniquement un téléphone mobile sont assez différentes de celles qui s'abonnent à un service téléphonique fixe. D'après Kussela et Vikki (1999), en Finlande, les premières sont généralement jeunes, vivent souvent seules, en appartement, en région urbaine. Il convient de souligner que le passage de la téléphonie fixe à la téléphonie mobile ne semble pas se produire dans les mêmes proportions en Amérique du Nord, en raison de stratégies différentes d'établissement des prix.

En théorie, l'échantillonnage par CA pourrait être étendu à la téléphonie mobile. Toutefois, en pratique, l'exercice pourrait être assez difficile, parce que le téléphone mobile est, de par sa nature, un appareil personnel plutôt qu'un appareil du ménage. Échantillonner des personnes dans un ménage en communiquant avec l'un des membres par téléphone mobile est virtuellement impossible. Interviewer par téléphone mobile des personnes qui peuvent se trouver n'importe où est également une tâche fort difficile. Même déterminer le nombre total de numéros de téléphone (services mobile et fixe) que possède un ménage (information nécessaire pour la pondération) peut être une tâche fastidieuse. À la section 4, nous considérons certains moyens éventuels d'aborder cette question et d'autres problèmes que pose le passage à la téléphonie mobile.

Le sous-dénombrement des personnes dans les ménages couverts a trait principalement à la méthode de sélection des personnes dans le ménage (voir la section 3.2.2) et au sous-dénombrement dû au fait que l'on n'obtient pas toujours la liste complète des personnes qui composent le ménage. Maklan et Waksberg (1988) ont étudié l'effet du second facteur en comparant les données sur les particuliers provenant d'une enquête par CA à celles provenant de la US Current Population Survey et du Recensement de la population. Selon ces auteurs, bien que la taille moyenne des ménages soit comparable, les résultats de l'enquête par CA sont biaisés en faveur des ménages de deux personnes et en défaveur des ménages d'une seule personne. Les écarts pourraient être dus en partie à l'application de règles de résidence différentes, mais les résultats n'indiquent pas que les personnes sont sous-dénombrées dans le cas de l'enquête par CA. Ils décrivent aussi une expérience qui a consisté à poser des questions plus détaillées sur la composition du ménage et ne constatent presque aucune amélioration de l'exactitude de la déclaration. Lors d'une expérience comparable, réalisée dans le cadre d'une enquête sur l'usage du tabac, Bercini et Massey (1979) ont testé les effets de l'utilisation de noms dans la liste de composition du ménage et de la position de la question sur la composition du ménage (avant ou après la première interview). Selon eux, l'utilisation des noms et la position de la question sur la composition du ménage ont toutes deux un effet sur la réponse et l'obtention de la composition du ménage après l'interview est la méthode qui donne les meilleurs résultats.

3.3.2 Non-réponse

La non-réponse et le biais qui y est associé est un problème fondamental de toutes les études par sondage, mais l'interview téléphonique pose des problèmes particuliers de non-réponse. L'un des problèmes principaux tient à l'ambiguïté des résultats de nombreux essais de composition – par exemple, ligne continuellement occupée ou pas de réponse, numéro raccordé à un télécopieur, un modem d'ordinateur ou un répondeur. Récemment, des dispositifs automatisés de filtrage ont été mis au point pour repérer les numéros de téléphone raccordés à des enregistrements indiquant s'ils sont ou non en service (Casady et Lepkowski 1999). Ainsi, du matériel et des logiciels personnalisés ont été développés pour déceler les enregistrements « à trois tonalités » qui indiquent que le numéro « n'est pas en service » de sorte que, s'ils sont composés, ces numéros puissent être éliminés de l'échantillon. L'appariement des listes de numéros aux fichiers des

« pages jaunes » permet d'éliminer a priori un grand nombre de numéros commerciaux. Cette méthode et d'autres réduisent le coût du filtrage et l'ambiguïté ayant trait aux appels qui restent sans cesse sans réponse.

Les progrès technologiques, tels que le « renvoi automatique d'appels » et l'identification de l'appelant, favorisent la non-réponse. En outre, il est plus facile de refuser ou de mettre fin à une interview téléphonique qu'à une interview en personne. Groves et Lyberg (1988a) ont étudié en profondeur ce problème de non-réponse et d'autres que posent l'interview téléphonique « à froid » et les moyens mis en œuvre aux États-Unis pour les résoudre. Plus précisément, comme l'ont fait CASRO (1982) et White (1983), ils recommandent que la définition du taux de non-réponse inclue, au dénominateur, une estimation du nombre de numéros restés sans réponse qui sont des numéros en service, en plus du nombre d'interviews complètes et incomplètes, du nombre de numéros qui ont produit un refus et du nombre d'autres unités non interviewées. L'estimation du nombre de numéros admissibles qui n'ont produit aucune réponse est exprimée sous forme de proportion par rapport au nombre de numéros admissibles qui ont produit une réponse. Cependant, cette estimateur pourrait être biaisé. Par exemple, l'usage intensif de répondeurs par les entreprises sous-entend que presque toutes les entreprises répondront et pourront être identifiées comme étant des entreprises. En outre, comme le fait remarquer Massey (1995), dans le cas du filtrage, ou sélection par tri, il faut modifier cette mesure en définissant, pour le filtrage des ménages, un taux de réponse qui correspond à la proportion estimative de ménages admissibles identifiés comme tels lors du tri de sélection, plutôt que sous forme de proportion de l'ensemble des ménages filtrés pour déterminer leur admissibilité. Cunningham, Brick et Meader (2000) présentent plusieurs mesures détaillées du taux de réponse et du taux d'admissibilité pour chaque étape d'une enquête avec filtrage, ainsi que du taux de réponse global, lors de la description de la méthodologie de la National Survey of America's Families.

En général, les taux de non-réponse sont plus élevés dans le cas des enquêtes téléphoniques que dans celui des enquêtes sur place, pour les raisons susmentionnées – consulter Hochstim (1967), Groves et Kahn (1979), Fitti (1979), Groves et Lyberg (1988a) pour la situation aux États-Unis, Wilson, Blackshaw et Norris (1988) ; et Collins, Sykes, Wilson et Blackshaw (1988) pour la situation au Royaume-Uni, ainsi que Drew, Choudry et Hunter (1988) en ce qui concerne les enquêtes gouvernementales au Canada. Ces derniers auteurs procèdent aussi à une comparaison des interviews téléphoniques « à froid » et « à chaud » qui ne révèle que de faibles écarts entre les taux de non-réponse. Plus récemment, une analyse de la situation pour 39 enquêtes téléphoniques américaines réalisée durant les années 1990 (Massey, O'Connor et Krotki 1997) a montré une légère réduction supplémentaire du taux de réponse, la moyenne étant de 62 % et la fourchette variant de 42 % à 79 % (toutefois, il semble qu'au Canada, les taux de réponse n'aient pas diminué ces dernières années). L'utilisation croissante de dispositifs technologiques (répondeurs, renvoi automatique des appels, lignes téléphoniques multifonctionnelles) et la fréquence croissante de la sollicitation par téléphone, déjà reconnue par Biel (1967) comme pouvant poser des problèmes pour les enquêtes téléphoniques, comptent parmi les facteurs auxquels on peut imputer cette augmentation de la non-réponse. L'American Statistical Assocation (1999) estime que la diminution du taux de participation aux enquêtes causée par la quasi saturation provoquée par les appels des entreprises de télémarketing est un problème grave auquel les spécialistes des études par sondage ne se sont pas attaqués pleinement. L'Association conclut que, si la tendance ne se renverse pas, les enquêtes téléphoniques telles que nous les connaissons pourraient disparaître dans les cinq ans à venir. Kalton (2000) est du même avis.

Comme cela est le cas pour la non-couverture téléphonique, le biais dû à la non-réponse introduit dans les estimations d'enquête est encore accentué par la corrélation entre la non-réponse et nombre de caractéristiques socio-économiques. Groves et Lyberg (1988a) se sont fondés sur l'examen de travaux antérieurs pour cerner les principaux corrélats de la non-réponse téléphonique. Il s'agit de l'âge (le taux de refus est plus élevé chez les personnes âgées – consulter aussi Collins et coll. 1988) et le niveau de scolarité (le taux de réponse est plus élevé pour les groupes dont le niveau de scolarité est faible – consulter, par exemple, Cannel, Groves, Magilavy, Mathiowetz, Miller et Thornberry 1987). Par contre, il existe des preuves que l'écart entre les taux de non-réponse observés pour les régions urbaines et rurales est plus faible dans le cas des enquêtes téléphoniques que dans celui des enquêtes sur place (Groves et Kahn 1979). Des articles plus récents sur les effets de la non-réponse se concentrent sur des problèmes particuliers. Ainsi, Diehr, Collins, Sykes, Wilson et Blackshaw (1992) étudient le lien entre le taux de réponse et d'autres variables sommaires au niveau du préfixe et de la personne et notent que la non-réponse est corrélée à l'âge, à la race, à la taille de la famille et au type de cette dernière. Dans une étude de l'effet des appels de suivi sur la qualité des estimations d'enquête, Merkle, Bauman et Lavrakas (1993) montrent que l'âge et la situation d'emploi sont les principales variables corrélées au nombre de rappels. Kalsbeek et Durham (1994) étudient l'effet de la non-réponse dans le cadre d'une enquête téléphonique de suivi sur l'allaitement maternel auprès de femmes à faible revenu et constatent que la non-réponse est corrélée principalement à l'âge et au degré d'urbanisation. Enfin, Hox, DeLeeuw et Kreft (1991) recourent à la modélisation multiniveaux pour réaliser une méta-analyse de grande portée des rapports sur les comparaisons de la non-réponse selon le mode d'enquête. Leur étude, fondée sur l'analyse par modélisation multiniveaux de 45 études (dont 35 comportant une composante téléphonique),

indique que le taux de réponse est nettement plus faible pour les enquêtes téléphoniques que pour les enquêtes sur place si l'on applique des modèles à pente constante. Cependant, si l'on choisit des modèles à pente aléatoire, l'écart n'est plus significatif.

Certains chercheurs ont étudié l'effet des variables opérationnelles d'enquête sur la non-réponse pour essayer de réduire cette dernière dans le cas des enquêtes téléphoniques. Ainsi, lors d'une expérience en prévision de la US National Crime Survey, Sebold (1988) constate que le fait de doubler la période d'enquête (pour passer de deux à quatre semaines) augmente le taux de réponse de trois points de pourcentage. Brick et Collins (1997) étudient l'effet de l'envoi préalable de lettres et des questions de filtrage sur le taux de réponse à la US National Household Education Survey. Selon eux, l'adoption d'une méthode de rejet à la sélection augmente considérablement le taux de réponse, mais l'envoi d'une lettre préalable ne rehausse pas l'effet de la présélection. La durée de l'interview (Collins et coll. 1988) et les caractéristiques vocales de l'intervieweur (Oksenberg et Cannel 1988) sont d'autres variables qui influent sur le taux de réponse. L'effet de la méthode de sélection de personnes dans le ménage sur la non-réponse (en particulier la nécessité d'obtenir la liste des membres du ménage) a déjà été mentionnée à la section 3.2.2.

Enfin, ces dernières années, on a assisté à une augmentation significative de l'utilisation de répondeurs et de dispositifs d'identification des appelants en vue de filtrer les appels non désirés, situation qui augmente la probabilité de non-réponse. Par exemple, en France, la proportion de ménages équipés d'un répondeur est passée de 21 % en 1995 à 40 % en 1999 (Rouquette 2000); l'augmentation est la même en Allemagne (République fédérale d'Allemagne 1999), tandis qu'aux États-Unis, la proportion est passée d'environ 25 % en 1988 (Tuckel et Feinberg 1991) à plus de 73 % en 1997 (*Decision Analyst* 1997). Cependant, d'après une enquête téléphonique nationale, Tuckel et Feinberg (1991) concluent que, comparativement à d'autres groupes de non-réponse initiale (par exemple, « pas de réponse » ou « signal occupé »), les ménages dotés d'un répondeur sont plus susceptibles de répondre et moins susceptibles de refuser de participer à l'enquête, ce qui produit un taux de réponse pour ces ménages qui n'est définitivement pas plus faible que celui observé pour les autres formes de non-réponse. En fait, selon une étude réalisée par Oldendick et Link (1994), il semble que le taux d'utilisation d'un répondeur pour filtrer les appels des enquêteurs est de l'ordre de 2 % à 3 % seulement. Cependant, les personnes qui filtrent effectivement les appels ont tendance à appartenir aux catégories supérieures de revenu, à vivre en région urbaine et à avoir atteint un niveau élevé de scolarité. Pareillement, Piazza (1993) constate, après examen de la foule de données provenant de la California Disability Survey, enquête téléphonique pour laquelle le nombre de rappels est élevé, que, s'il est plus difficile au départ d'entrer en contact avec les propriétaires d'un répondeur, ces personnes sont au moins aussi susceptibles de participer à l'enquête que les personnes qui ne possèdent pas de répondeur. Il fait également remarquer que tomber sur un répondeur donne la certitude qu'un ménage a été rejoint et que ses membres ne veulent pas manquer les appels importants. Selon une étude conçue par Xu, Bates et Schweitzer (1993) pour étudier l'effet des messages laissés sur les répondeurs, les ménages qui possèdent un répondeur sont plus susceptibles d'être contactés et de participer à l'interview que ceux qui n'en possèdent pas. De surcroît, le fait de laisser un message sur le répondeur augmente de façon significative le taux de réponse et réduit de façon significative le taux de refus. Pareillement, Harlow, Crea, East, Oleson, Fraer et Cramer (1993) observent, en se fondant sur les résultats d'une expérience contrôlée, que laisser un message sur le répondeur augmente le taux de réponse de 15 %, après correction pour tenir compte de l'effet de l'âge, de l'intervieweur et de la ville de résidence. Selon un essai randomisé réalisé par Koepsell, McGuire, Longstreth, Nelson et van Belle (1996), laisser un message sur le répondeur produit une augmentation globale de 20 % du taux de réponse. Bien que, lors d'une étude similaire, Tuckel et Shukers (1997) n'aient constaté aucun effet significatif, les résultats globaux d'un éventail d'études indiquent que l'usage accru des répondeurs à un effet favorable sur le taux de réponse aux enquêtes, probablement parce qu'il donne la possibilité de laisser un message positif, donc de permettre aux personnes appelées de filtrer les appels des télévendeurs.

Tuckel et O'Neill (1996) estiment que la proportion de ménages américains possédant un dispositif d'identification des appelants est passé de 3 % en 1992 à 10 % en 1996. Selon une étude nationale, qui comprend l'analyse du profil des abonnés à l'option d'identification des appelants et des propriétaires de répondeurs, ces auteurs concluent que ces dispositifs technologiques ne constituent pas encore des obstacles importants à la recherche par sondage téléphonique, puisque leurs propriétaires ont tendance à utiliser les dispositifs de filtrage principalement pour rejeter les appels indésirables provenant de personnes dont ils connaissent le numéro plutôt que les appels provenant de numéros inconnus. Cependant, ils soulignent que la possibilité de filtrer les appels augmentera probablement le taux de réponse par répondeur aux appels de suivis répétés.

3.3.3 Pondération et rajustement des données

Une attention particulière doit souvent être accordée à la pondération et au rajustement des données recueillies par enquête téléphonique. Bien que les plans d'échantillonnage se fondent habituellement sur la sélection avec probabilités égales, en pratique ces conditions ne sont pas toujours atteintes. Par exemple, en théorie, les plans d'échantillonnage par CA sont autopondérés, mais, en réalité, la multiplicité des lignes téléphoniques (numéros) que possède

un ménage peut causer une sélection avec probabilités inégales. Le cas échéant, si l'on recueille des renseignements sur le nombre de lignes téléphoniques auxquelles le ménage est raccordé, la correction est simple. Pareillement, une repondération est nécessaire pour tenir compte des UPE pour lesquelles le nombre de numéros faisant partie du champ d'observation est inférieur à la taille requise de l'échantillon de grappes. Le fait qu'il soit souvent difficile de déterminer si un numéro de téléphone pour lequel on n'obtient aucune réponse après des tentatives répétées d'appel est effectivement un cas de non-réponse d'un numéro qui fait partie du champ d'observation ou, en fait, un numéro hors du champ d'observation pose un problème supplémentaire. La non-réponse, le sous-dénombrement inhérent aux ménages qui ne possèdent pas le téléphone et la nécessité manifeste d'utiliser une certaine forme d'estimateur de la multiplicité dans le cas des plans d'échantillonnage à bases de sondage multiples, en s'appuyant sur les données des bases de sondage dans lesquelles l'unité est représentée, sont d'autres problèmes qu'il faut résoudre.

Ces problèmes sont traités, dans le cas de l'échantillonnage national par CA réalisé par le US National Center for Health Statistics, dans une série d'articles publiés par Thornberry et Massey (1978), Botman, Massey et Shimizu (1982), ainsi que Massey et Botman (1988). Ces articles décrivent les rajustements de la pondération effectués pour la US National Health Interview Survey (NHIS) réalisée par CA et pour une enquête sur l'usage du tabac afin de tenir compte de la multiplicité des lignes ou des numéros de téléphone par ménage, de la couverture téléphonique et de la non-réponse. Ces rajustements se fondent sur des données externes pour la race et la région géographique, et sur des données d'enquête pour la non-réponse et les lignes téléphoniques multiples. Plusieurs méthodes de rajustement et de pondération sont comparées et évaluées. Chapman et Roman (1985) comparent la substitution à la correction pour la non-réponse lors d'une étude de faisabilité ayant trait à la NHIS par CA et constatent que le biais et la variance sont comparables. Drew et Groves (1989) comparent diverses méthodes de rajustement pour la non-réponse unitaire fondées sur l'utilisation de données administratives externes au moyen d'un modèle explicite de prédiction de la réponse et sur les probabilités de réponse estimées d'après les données concernant les appels de suivi. Casady et Sirken (1980) proposent un estimateur de multiplicité pour un plan d'échantillonnage à bases de sondage multiples applicable aux données de la US National Health Interview Survey. Brick (1990) compare l'estimateur de multiplicité à l'estimateur à bases multiples classique dans le cas d'une enquête par CA sur la scolarité.

Goksel, Judkins et Mosher (1991) décrivent des rajustements, fondés sur la modélisation de la propension à la non-réponse, dans le cas d'un suivi téléphonique après une interview sur place dans le cas de la US National Survey of Family Growth. Bull, Pederson et Ashley (1988) proposent une correction fondée sur la propension à répondre selon l'intensité des efforts de suivi et selon la catégorie d'usage du tabac pour une enquête canadienne sur les attitudes à l'égard des dispositions législatives limitant l'usage du tabac.

À la suite d'une comparaison des ménages qui n'ont pas le téléphone aux ménages « en transition » (c'est-à-dire ceux qui ont acquis ou perdu récemment le service téléphonique) faite par Keeter (1995), Brick, Waksberg et Keeter (1996) proposent d'utiliser les données sur l'interruption du service téléphonique pour faire la correction pour le sous- dénombrement dû aux ménages qui n'ont pas le téléphone. Les résultats montrent que ce genre de correction peut réduire l'erreur quadratique moyenne. Hoaglin et Battaglia (1996) comparent une méthode modifiée de stratification a posteriori ainsi qu'une estimation basée sur un modèle avec stratification a posteriori simple pour corriger les données de la non-couverture dans le cas d'une enquête par CA sur la couverture de la vaccination. La méthode modifiée de stratification a posteriori s'appuie sur des données nationales sur les taux de vaccination pour les enfants qui vivent dans des ménages ayant et n'ayant pas le téléphone, ainsi que sur des données démographiques et socioéconomiques utilisées pour la stratification a posteriori simple, tandis que, dans le cas de la méthode basée sur un modèle, on se fonde sur un modèle logit pour estimer la probabilité de vivre dans un ménage ayant le téléphone. L'étude montre que l'utilisation de la méthode modifiée de stratification a posteriori produit une amélioration, mais que les résultats de la méthode modifiée de stratification a posteriori et de la correction basée sur un modèle diffèrent peu. Frankel, Srinath, Battaglia, Hoaglin, Wright et Smith (1999) appliquent une correction similaire, fondée sur des données sur l'interruption du service téléphonique, aux données de la NHIS et montrent de façon concluante qu'elle réduit considérablement le biais.

3.4 Qualité des données – Erreurs de réponse et effets de mode

La qualité des renseignements recueillis par téléphone a toujours été une question controversée. Comme on l'a mentionné à la section 2, les craintes quant à la qualité prétendument inférieure des données recueillies par interview téléphonique ont été apaisées très tôt, en grande partie grâce à certaines évaluations empiriques à grande échelle réalisées durant les années 1960 et 1970. Cependant, certaines données contradictoires provenant de diverses études sur la qualité relative des interviews par téléphone et sur place persistaient néanmoins. Même si l'analyse approfondie des données de grandes enquêtes omnibus réalisées selon les deux modes d'interviews par le University of Michigan Survey Research Center (Groves et Kahn 1979) ont fourni des renseignements importants sur la qualité des données et d'autres questions, les comparaisons entre les modes d'interview et la comparaison à des données externes n'ont pas été concluantes. En vue de résoudre le

problème, de Leeuw et van der Zoowen (1988) ont procédé à une méta-analyse à grande échelle de 28 études empiriques importantes comportant une comparaison entre l'interview sur place et l'interview par téléphone. Les études, réalisées entre 1952 et 1986 – la plupart aux États-Unis et certaines en Europe – portaient sur divers sujets. Les indicateurs de qualité des données utilisés étaient la validité de la réponse (établie d'après des études de validation), l'absence de biais dû au caractère socialement désirable, la réponse à une question, la quantité d'information (pour les questions ouvertes ou les listes de vérification) et la similarité des réponses. Dans l'ensemble, l'analyse montre que, s'il existe des différences de qualité entre les deux modes d'interviews, celles-ci sont définitivement très faibles et que d'autres aspects, comme les coûts et la commodité, devraient être pris en considération pour décider de recourir à l'interview téléphonique pour le travail d'enquête. Des conclusions similaires sont tirées pour le Royaume-Uni par Sykes et Collins (1988) d'après quatre études comparatives, pour le Danemark par Körmendi (1988) pour les données sur le revenu d'après une étude de validation fondée sur des données administratives et pour le Canada par Caron et Lavallée (1998), pour une comparaison des données financières provenant de l'Enquête financière sur les fermes.

D'autres études récente sur les effets du mode d'interview se concentrent sur des problèmes et des sujets particuliers, mais aboutissent aux mêmes conclusions. Ainsi, Herzog et Rodgers (1988) présentent les résultats d'une comparaison des modes d'interview dans le cas d'une étude auprès de personnes âgées et n'observent que de légères différences. Foley et Brook (1990) présentent des résultats similaires pour une enquête sur les derniers jours de la vie. Lors d'une étude sur le sujet délicat de la consommation de drogues, Aquilino et Lo Sciuto (1990) obtiennent des résultats presque identiques pour les personnes de race blanche, mais observent certains écarts significatifs pour les personnes de race noire, même après avoir tenu compte de l'effet de variables éventuellement liées au sous- dénombrement téléphonique. Les résultats d'une enquête téléphonique sur la consommation de drogues présentée par Johnson, Fendrich, Shaligram et Garey (1997), qui étayent un modèle des effets d'intervieweur axé sur la distance sociale pourraient expliquer cette situation.

Il y a peu de doutes que l'effet de l'intervieweur sur la qualité des données soit important, que l'enquête soit réalisée sur place ou par téléphone. L'utilisation d'installations centralisées d'interview téléphonique permet de mieux contrôler et surveiller les effets d'intervieweur que dans le cas de l'interview sur le terrain. À cet égard, certaines questions sont traitées par Stokes et Yeh (1988) qui proposent un modèle bayesien des effets d'intervieweur, ainsi que des méthodes d'estimation des paramètres du modèle. Pannekoek (1988) propose un modèle bêta-binomial de la composante de la variance due à l'intervieweur et des méthodes d'estimation de ces paramètres.

Un moyen efficace de réduire les erreurs de réponse dans le cas des enquêtes avec interview sur place consiste à utiliser des dossiers fournis par les répondants pour vérifier et rappeler les renseignements sur le revenu, les assurances, les événements liés à la santé, *etc.* Manifestement, l'extension de cette méthode à l'interview téléphonique pose certains problèmes, puisque l'intervieweur ne peut consulter lui-même les documents et que même le fait de demander aux répondants d'aller chercher ces documents peut causer un arrêt plus fréquemment perturbant dans le cas de l'interview téléphonique que dans celui de l'interview sur place. Cependant, l'utilisation de dossiers par les répondants durant les enquêtes téléphoniques peut contribuer à réduire le biais de réponse. Battaglia, Shapiro et Zell (1996) décrivent une expérience consistant à demander aux répondants d'utiliser les dossiers de vaccination durant l'un des cycles de la US National Immunization Survey et à comparer les renseignements recueillis à ceux des dossiers des vaccinateurs. Bien que quelque 47 % de répondants se soient effectivement servi de leur dossier de vaccination, un biais important de sous-dénombrements persiste, peut-être parce que les rapports de vaccination ne sont pas toujours à jour. Des effets comparables ont été constatés pour les enquêtes réalisées sur place – consulter Brick, Katton, Nixon, Givens et Ezzati-Rice (2000).

4. PROGRÈS TECHNOLOGIQUES COURANTS ET FUTURS

Conjugués à la couverture téléphonique presque complète, le progrès technologique très intense et la diversité des moyens de communication ne cessent d'offrir de nouvelles occasions d'utiliser des moyens neufs de communication pour réaliser les enquêtes. Par ailleurs, certains de ces progrès pourraient compliquer la réalisation des téléenquêtes selon la méthodologie classique appliquée aujourd'hui. Ainsi, la complexité croissante des dispositifs et des algorithmes de filtrage (comme l'évolution des simples répondeurs et dispositifs d'identification des appelants mentionnés à la section 3.3) permettra peut-être aux répondants d'éviter plus facilement de coopérer. Nous allons maintenant examiner les applications courantes, ainsi que les développements et les applications futurs prévus, et commenter les problèmes méthodologiques que pose leur utilisation.

4.1 Courrier électronique et enquêtes par Internet

Le nombre de ménages qui ont accès à Internet a augmenté très rapidement ces dernières années. Par exemple, aux États-Unis, la proportion de ménages raccordés à Internet est passée de 26 % en décembre 1998 à 42 % en août 2000 (NTIA 2000). Dans d'autres pays, la proportion est un peu plus faible, soit 28 % au Royaume-Uni (en août 2000 – OFTEL 2000), 25 % au

Canada, 22 % en Finlande, 7 % en France et 5 % en Belgique en 1999 selon Rouquette (2000), 12 % en Israël, (en 1999 – Bureau central de la statistique 2000) et 11 % en Allemagne (République fédérale d'Allemagne 1999). Malgré cette augmentation rapide, la couverture Internet est encore loin d'être complète. Qui plus est, certaines données laissent entendre qu'une catégorie croissante d'anciens internautes se forme parallèlement à l'augmentation globale de l'utilisation. Selon Katz et Aspden (1998), la proportion d'anciens utilisateurs d'Internet est passée de 8 % en 1995 à 11 % en 1996. Cependant, l'augmentation globale de l'accès a favorisé l'utilisation du courrier électronique et d'Internet pour la réalisation des enquêtes. Quoique la couverture d'une enquête par courrier électronique (ECE) soit comparable à celle d'une enquête Internet et que toutes deux se fondent sur l'utilisation d'un questionnaire à remplir soi-même informatisé (ou QRSMI), il existe une différence fondamentale entre ces deux formes de téléenquêtes. L'enquête par courrier électronique est fort semblable à une enquête par la poste, en ce sens qu'elle consiste à envoyer un questionnaire textuel et à demander au répondant de le retourner dûment rempli. L'avantage par rapport à l'enquête par la poste tient au coût plus faible, ainsi qu'à la facilité et à la simplicité de transmission et de réception. L'enquête Internet se fonde, en général, sur l'interaction entre le répondant et l'instrument d'enquête, grâce à l'utilisation de Java, XML ou d'un instrument comparable. Elle permet de nombreuses améliorations, comme l'utilisation de couleurs et d'animation, et les possibilités qu'elle offre sont multiples en ce qui concerne l'utilisation d'enchaînements complexes des questions et la vérification en temps réel. Les perspectives intéressantes de développement de systèmes novateurs de collecte des données grâce à la création incessante de nouveaux outils Internet ne permettent pas encore de surmonter le problème fondamental que posent les enquêtes par courrier électronique et les enquêtes Internet, c'est-à-dire la couverture absolument insuffisante à l'heure actuelle pour la plupart des populations humaines étudiées (Dillman 2000).

Néanmoins, les enquêtes par courrier électronique et par Internet peuvent et sont utilisées, avec plus ou moins de bonheur, pour certaines populations pour lesquelles la couverture est virtuellement complète, ou conjuguées à d'autres modes de collecte. Ainsi, Couper, Blair et Triplett (1999) font le compte rendu d'une étude expérimentale visant à comparer l'utilisation du courrier électronique et du service postal ordinaire pour réaliser une enquête auprès des employés de plusieurs bureaux gouvernementaux de la statistique aux États-Unis. Les employés échantillonnés ont été affectés au hasard à la collecte des données par courrier électronique ou par la poste, et des méthodes comparables ont été suivies pour le contact préalable et le suivi des sujets. Le taux de réponse a été un peu plus élevé pour l'enquête par la poste que pour celle par courrier électronique, mais la qualité des données (éléments de données manquants) était la même pour les deux modes d'enquête. Lors d'essais sur le terrain de la US National Study of Postsecondary Faculty de 1999, on a donné aux administrateurs ainsi qu'aux professeurs le choix de remplir et de renvoyer par la poste un questionnaire papier traditionnel ou d'utiliser un questionnaire à remplir soi-même informatisé (QRSMI) par Internet (Abraham, Steiger et Sullivan 1998). Bien qu'il soit raisonnable de supposer que presque tous les répondants avaient accès à Internet, 8 % seulement des professeurs répondants et 17 % des administrateurs d'établissements ont choisi le QRSMI. La US National Science Foundation prévoit offrir une option Internet pour la National Survey of Recent College Graduates de 1999, l'hypothèse étant que la plupart des personnes visées par l'enquête connaissent les ordinateurs et ont accès à Internet (Meeks, Lanier, Fecso et Collins 1998). Pour une revue de l'utilisation des QRSMI par les organismes gouvernementaux et les organismes privés d'enquête et des problèmes qu'elle pose, consulter Ramos, Sedivi et Sweet (1998).

Cependant, à l'heure actuelle, la plupart des enquêtes Internet visant des populations générales s'appuient sur l'échantillonnage non probabiliste, principalement par recours à une forme ou l'autre d'auto-sélection. Fischbacher, Chappel, Edwards et Summerton (1999) décrivent la méta-analyse de 28 enquêtes touchant le domaine de la santé réalisées par courrier électronique et par Internet. Nombre de ces enquêtes étant des études épidémiologiques visant des personnes atteintes de maladies particulières, le problème du biais de sélection empêche la généralisation de la plupart des résultats. L'une des enquêtes Internet les plus importantes est la WWW User Survey réalisée par le Graphics Visualization and Usability Center du Georgia Institute of Technology (Kehoe, Petkow, Sutton, Aggarwal et Rogers 1999). Bien que la population étudiée soit, par définition, les internautes, l'absence de tout cadre d'échantillonnage pour cette population sous-entend que les répondants ont dû être sollicités par diverses méthodes (annonces sur Internet et d'autres supports, bandeaux publicitaires, prix d'encouragement en espèces, *etc.*) au lieu d'être échantillonnés avec une probabilité connue. Bien que quelque 20 000 internautes aient participé, les auteurs du rapport mentionnent que les données sont biaisées en faveur des utilisateurs chevronnés et les plus fréquents d'Internet et recommandent que l'on augmente les données en réalisant des enquêtes auprès d'échantillons aléatoires. Pour essayer d'éviter le biais dû à la réalisation des enquêtes auprès d'échantillons de la population de personnes qui ont accès à Internet uniquement, certains organismes commerciaux d'enquête distribuent à tous les membres des panels qu'elles sélectionnent par CA un appareil qui leur permet de se raccorder à Internet au moyen de leur téléviseur pour s'assurer d'obtenir des résultats cohérents (Felson 2001). Cependant, Poynter (2000) prédit que d'ici à l'an 2005, 95 % d'études de marché seront réalisées par la voie d'Internet, mais que 80 % seront réalisées auprès de répondants qui auront

« choisi de participer » plutôt qu'auprès d'un échantillon probabiliste.

Par ailleurs, il existe des preuves que la collecte de données par Internet donne d'assez bons résultats pour les enquêtes auprès des établissements. Nusser et Thompson (1998) décrivent son utilisation pour les National Resources Inventory Surveys du US Department of Agriculture; Rosen, Manning et Harrell (1998) publient les résultats de la collecte de données par Internet auprès des établissements qui participent à la US Current Employment Statistics Survey et Meeks et coll. (1998) décrivent son utilisation pour la collecte de données auprès d'établissements universitaires, d'organismes fédéraux et de sociétés privées pour les US National Science Foundation surveys. En supposant que l'on arrivera à résoudre les problèmes de couverture et d'échantillonnage en ce qui concerne les ménages et les particuliers, il est permis d'espérer que l'on pourra appliquer la collecte de données par Internet aux enquêtes-ménages dans l'avenir.

4.2 Autres questionnaires à remplir soi-même informatisés (QRSMI) et méthodes d'auto-interview assistée par ordinateur (AIAO)

Couper et Nichols (1998) font une distinction entre la collecte au moyen d'un questionnaire à remplir soi-même informatisé (QRSMI) qui ne comporte pas l'intervention d'un intervieweur et l'auto-interview assistée par ordinateur (AIAO) durant laquelle un intervieweur est présent ou délivre l'instrument d'enquête. Donc, tant les enquêtes par courrier électronique que celles par Internet se fondent sur l'utilisation d'un QRSMI appuyé par la technologie des télécommunications. La saisie des données au clavier (SDC), où les répondants entrent les données en se servant du clavier de leur téléphone, et la reconnaissance vocale interactive (RVI) ou la saisie par reconnaissance vocale (SRV) sont d'autres méthodes basées sur un QRSMI. Dans les deux cas, les répondants font eux-mêmes l'appel à leur meilleure convenance pour faire leur déclaration, après une première prise de contact et les deux méthodes ont été mises à l'essai à grande échelle et utilisées avec succès par le US Bureau of Labor Statistics pour recueillir des données auprès des établissements pour son Current Employment Statistics program Werking, Tupek et Clayton (1988), Winter et Clayton (1990), Clayton et Winter (1992). Phipps et Tupek (1991) décrivent une étude de la qualité de la collecte de données par SDC réalisée par vérification des enregistrements, et montrent que la méthode pose peu de problèmes et que les erreurs de réponse diminuent avec l'expérience. Plus récemment, les bureaux américains de la statistique ont entrepris des essais en vue de déterminer s'il est possible d'appliquer ces méthodes axées sur un QRSMI aux enquêtes-ménages. McKay, Robison et Malik (1994) font le comte rendu d'un essai de laboratoire préliminaire de la SDC pour la Current Population Survey. Malakhoff et Appel (1997), quant à eux, décrivent la mise au point d'un prototype de RVI au US Bureau of Census, mais il s'agit d'une opération d'établissement de listes par les employés qui travaillent sur le terrain. Il convient de souligner que, si la SDC est de toute évidence propre aux enquêtes téléphoniques, la RVI peut être appliquée à d'autres modes de collecte.

Les méthodes d'auto-interview assistée par ordinateur (AIAO), qui incluent les modes audio (AIAOA) et vidéo (AIAOV) de collecte, sont considérées depuis longtemps comme des extensions naturelles des enquêtes par la poste qui tirent parti de la technologie moderne (Dillman 2000). On a surtout insisté sur leur utilité pour les enquêtes portant sur des sujets délicats et embarrassants, où l'idée que l'intervieweur soit présent (interview sur place) peut faire hésiter les répondants à participer à l'enquête. Pour une revue des progrès récents visant ces méthodes, consulter Baker (1998), O'Reilly, Hubbard, Lessler, Biemer et Turner (1994), Rogers, Miller, Forsyth, Smith et Turner (1996), ainsi que Tourangeau et Smith (1998). Presque toutes les applications décrites correspondent à des enquêtes pour lesquelles l'instrument est amené au domicile du répondant par un membre du personnel sur le terrain. Certains chercheurs ont déjà essayé d'utiliser le téléphone pour la collecte de données par AIAOA (AIAOA-T) – consulter Turner, Forsyth, O'Reilly, Cooley, Smith, Rogers et Miller (1998). La mise au point longtemps attendue de la vidéotéléphonie pour en faire une forme commune généralisée de service téléphonique pour les ménages ne s'est pas encore concrétisée. Le jour où elle se matérialisera, la vidéotéléphonie rendra possible l'AIAOV téléphonique (AIAOV-T), ce qui aura des conséquences importantes pour le travail d'enquête. L'ajout d'un élément visuel permettra de surmonter plus facilement les problèmes que posent aujourd'hui les enquêtes téléphoniques contrairement aux enquêtes par interview sur place (regard droit dans les yeux de l'intervieweur, utilisation de cartons aide-mémoire ou d'autres aides visuelles). Il faudra sans doute fort longtemps avant que la vidéotéléphonie soit universelle, si bien que, du moins pour le moment, l'AIAOV-T ne pourra servir que de moyen complémentaire de collecte des données.

4.3 Téléphones mobiles

Les problèmes de couverture que pourraient poser les enquêtes par CA axées sur le service téléphonique fixe à cause de la prolifération rapide des téléphones mobiles ont été mentionnés à la section 3.3.1. Dans l'avenir, il est évident qu'il faudra se servir des téléphones mobiles pour rejoindre le nombre sans cesse croissant de ménages qui ne sont pas abonnés au service téléphonique fixe. Le niveau actuel de couverture des téléphones mobiles signifie que les enquêtes par téléphone mobile ne peuvent, en général, être utilisées que pour des populations particulières ou pour compléter les enquêtes par CA dans le cas d'un service fixe. Par exemple, Perone, Matrundola et Soverini (1999) présentent une enquête par téléphone mobile auprès d'une

population à laquelle on a naturellement accès – celle des abonnés au service mobile – afin d'évaluer la satisfaction des consommateurs. Ils constatent que les taux de refus n'excèdent pas ceux observés en cas d'enquête téléphonique à ligne fixe. Cependant, les taux de non contact sont élevés, principalement parce que les abonnés se trouvent en-dehors du champ du signal ou qu'ils éteignent leur téléphone. Un autre problème que posent les enquêtes par téléphone mobile tient au fait que, dans de nombreux cas en Amérique du Nord, l'abonné doit payer pour recevoir des appels (Casady et Lepkowski 1999).

Comme on l'a mentionné plus haut, Cunningham et coll. (1997) décrivent l'utilisation des téléphones mobiles pour interviewer les membres des ménages qui ne possèdent pas le téléphone (principalement dans les régions rurales), le téléphone mobile étant apporté aux répondants par les intervieweurs sur le terrain. Cette méthode a été conçue en vue de réduire au minimum les effets de mode d'interview grâce à des interviews téléphoniques réalisées par les mêmes intervieweurs que ceux qui avaient procédé aux interviews des membres des ménages abonnés au téléphone. Les taux de réponse sont élevés, même si, dans certains cas, l'interview a dû avoir lieu à l'extérieur pour que la réception soit bonne. L'usage le plus intensif du téléphone mobile dans le cadre d'enquêtes-ménages est sans aucun doute celui qui en a été fait en Finlande pour l'Enquête sur la population active (Kuusela et Notkola 1999). Des quelque 97 % d'interviews réalisées par téléphone, plus de 20 % le sont par téléphone mobile. La durée moyenne des interviews par téléphone mobile est légèrement supérieure à celle des interviews par service téléphonique fixe, mais l'écart est sans doute causé par des différences sociodémographiques entre les groupes de répondants.

4.4 Progrès technologiques futurs et leur effet sur la méthodologie des téléenquêtes

La rapidité des progrès technologiques dans le domaine des télécommunications et des systèmes d'information permet difficilement de prédire leur effet sur les activités d'enquête. Ces changements technologiques ne permettront pas nécessairement tous d'augmenter l'application des techniques de télécommunication de pointe aux activités d'enquête. Les problèmes que posent les personnes qui ont décidé de « décrocher » d'Internet (Katz et Aspden 1998) ou du service téléphonique fixe (voir, par exemple, Gabler et Haeder 2000 et Kuusela et Vikki 1999) ont déjà été mentionnés. De surcroît, dans certains domaines, comme l'étude de marché et la statistique officielle, les progrès technologiques permettront peut-être de dépendre moins des enquêtes pour recueillir les renseignements nécessaires aux prises de décision. Par exemple, Baker (1998) et Poynter (2000) prédisent que des techniques telles que l'extraction de données à partir de sources existantes pourraient devenir le mode prédominant de collecte de données pour les études de marché. De façon similaire, Scheuren et Petska (1993) examinent la possibilité de s'appuyer sur des systèmes de dossiers administratifs pour produire des statistiques officielles. Néanmoins, pour certains domaines importants (par exemple, sondages d'opinion et données sur les comportements inobservables), les enquêtes demeureront la source principale de données. Les progrès technologiques offriront de nouvelles perspectives de téléenquête, mais rendront peut-être la méthodologie plus compliquée que celle utilisée aujourd'hui.

L'intégration d'appareils et de méthodes de communication multiples – téléphonie (services fixe et mobile), télécopieur, Internet, courrier électronique, vidéotéléphonie, transmission de données, transmission d'images télévisées, *etc.* – est l'un des progrès attendu dans un avenir proche (Baker 1998). Autrement dit, chaque personne aura accès à divers services de télécommunication, éventuellement au moyen d'un seul instrument physique qui pourrait être un téléphone mobile (par exemple, technologie des téléavertisseurs à zone étendue), un ordinateur personnel ou un téléviseur, ou d'une combinaison de ces appareils. Pareillement, les intervieweurs auront peut-être la possibilité d'entrer en communication avec les répondants par divers moyens. Pour la description de certains problèmes que pourrait poser cette pléthore de moyens d'accès, consulter Ranta-aho et Leppinen (1997). On s'attend à ce que la personne appelée puisse contrôler dans une large mesure si elle recevra la communication et, le cas échéant, de quelle façon. Nombre d'utilisateurs ont déjà cette capacité grâce à des moyens perfectionnés de filtrage, de renvoi automatique d'appels, de transfert de messages, de transmission de messages multiples, *etc.* En revanche, parallèlement, le degré de contrôle que l'appelant exerce sur le mode de transmission diminuera vraisemblablement.

En ce qui concerne les opérations d'enquête, ces progrès signifient que les enquêtes à mode d'interview mixte et, éventuellement, la méthodologie des bases de sondage multiples prédomineront sans doute. Selon nous, même si la couverture globale des télécommunications atteint un point de saturation proche de la couverture universelle, il est peu probable qu'un mode particulier de télécommunications puisse offrir, en soi, une couverture virtuellement complète. En outre, même si un seul mode de télécommunication pouvait assurer une couverture presque complète, rien ne dit qu'une méthode à mode mixte, tenant compte des préférences des répondants, ne soit pas préférable. Le fait que les activités d'enquête dépendent de plus en plus de la coopération volontaire des répondants dicte pratiquement que l'on donne aux répondants le choix du mode de communication. Cependant, on notera que les enquêtes à mode d'interview mixte sont fort coûteuses et qu'à l'heure actuelle, la technologie ne permet pas le simple transfert des questionnaires d'un mode (par exemple, le questionnaire Blaise d'IAO) à un autre mode (par exemple, un questionnaire papier).

Le problème le plus important que poseront, pour la conception des enquêtes, les progrès dans le domaine des

télécommunications, sera sans doute celui du choix d'un cadre de référence pertinent et de la répartition des unités d'échantillonnage entre les modes de collecte. On pense qu'en bout de ligne, chaque personne se verra attribuer un numéro de communication personnel (ou numéro d'identification) unique permanent grâce auquel elle pourra être rejointe de diverses façons (communication écrite, verbale ou visuelle) grâce à divers appareils fixes ou sans fil qui pourraient se trouver à son domicile, à son bureau ou être mobiles. Le choix du mode d'interview dépendra de la décision collective du répondant et de l'appelant. Les libertaires condamneront indubitablement l'idée de ce numéro universel (qui serait essentiellement un numéro d'identité), mais il est probable qu'il devienne éventuellement acceptable, même si de petits groupes d'activistes essayent d'éviter de l'utiliser et même d'entraver son application. En fait, des systèmes normalisés de numéros d'identité universels sont en place et bien acceptés depuis plusieurs décennies dans plusieurs pays d'Europe du Nord et en Israël. Dans ces pays, le numéro d'identité, qui n'est pas considéré comme un renseignement confidentiel, est utilisé à grande échelle à de nombreuses fins administratives et commerciales. Par exemple, en Israël, la loi exige que les chèques personnels portent le numéro d'identification de la personne, son nom, son adresse et son numéro de téléphone.

Une fois que l'on pourra exploiter un système de numéro de communication unique, on pourra appliquer les méthodes types d'échantillonnage. Il se pourrait fort bien que l'accès à des listes complètes de ces numéros – ne contenant peut-être que des renseignements géographiques et autres limités, soit généralisé, comme pour les numéros d'identification utilisés dans de nombreux systèmes nationaux d'enregistrement. Il y a de bonnes raisons de croire que la situation sera la même pour les numéros de communication – au départ, au moins en Europe, plutôt qu'en Amérique du Nord. En effet, les techniques perfectionnées de filtrage pourraient fort bien rendre redondante l'utilisation de numéros de téléphone non publiés. Le filtrage risque, certes, de faciliter la non-réponse, mais la possibilité de transmettre au préalable un message écrit par courrier électronique ou de laisser un message dans une boîte vocale pourrait réduire le problème.

L'échantillonnage à partir du genre de listes susmentionnées serait simple, mais inefficace dans la plupart des cas, puisque les données auxiliaires n'offriraient peut-être qu'un avantage marginal. Bien que les listes puissent permettre de faire la distinction entre les numéros résidentiels et commerciaux, il est douteux qu'elles fournissent des renseignements sur les ménages. Il s'ensuit que l'unité d'échantillonnage et de déclaration devrait être la personne plutôt que le ménage. Il s'agit de toute façon du but poursuivi par de nombreuses enquêtes et il n'est pas certain que le choix du ménage comme unité d'échantillonnage pour les téléenquêtes soit utile, même dans la pratique courante. S'ils sont nécessaires, les renseignements sur les ménages devraient être recueillis auprès des membres des ménages et comprendre ceux sur la taille du ménage pour que l'on puisse pondérer correctement les données sur les caractéristiques du ménage. Si le système de numéros de communication permet d'attribuer un numéro unique à chaque personne, aucun renseignement ne sera nécessaire sur les modes de communication ou leur multiplicité.

S'il n'existe pas de listes de numéros de communication ou que le problème des numéros non publiés persiste, il faudra recourir à une forme de sélection par CA. La méthode suivie ne devrait pas différer beaucoup des méthodes de CA employées à l'heure actuelle. Si l'on suppose que le système de numéros de communication sera effectivement unique et universel et qu'il suivra une certaine logique, on devrait pouvoir élaborer facilement des méthodes efficaces d'échantillonnage. Idéalement, le système de numérotation présentera certains liens avec les données géographiques, grâce à l'adresse permanente de la personne. Sinon, les enquêtes locales, voire même nationales, par CA deviendront extrêmement difficiles à concevoir efficacement. L'obtention de suffisamment de renseignements sur le système de numérotation permettrait de réduire au minimum le nombre de numéros hors du champ de l'enquête.

Puisque le répondant contrôlera vraisemblablement en grande partie le choix du mode de communication, la question de la répartition des unités d'échantillonnage entre les modes de communication ne se posera pour ainsi dire pas. Les concepteurs d'enquête devront préparer toute une gamme d'instruments de collecte adaptés aux divers modes de communication. Cette gamme inclura des instruments textuels, comme des versions pour télécopieur, courrier électronique et Internet du questionnaire, des instruments oraux, comme les interviews orales classiques et automatisées ou une combinaison de celles-ci. L'intégration des données recueillies selon divers modes de collecte en un ensemble uniforme de données posera un défi technologique énorme, mais surmontable.

La situation presque utopique susmentionnée ne se concrétisera sans doute pas avant longtemps et, entre-temps, il faudra élaborer des méthodes appropriées pour résoudre les problème que posera, à court terme, la mise au point des technologies de communication et de leurs applications. Comme il l'est noté à la section 4.3, il faudra tenir compte prochainement de la nécessité de passer des enquêtes téléphoniques fondées uniquement sur le service téléphonique fixe à une combinaison de services mobile et fixe. Fondamentalement, la méthodologie des bases de sondage multiples mise au point pour couvrir à la fois les ménages qui ont le téléphone et ceux qui ne l'ont pas peut être étendue facilement pour résoudre ce problème. Il faut encore créer les bases de sondage et(ou) les méthodes d'échantillonnage par CA appropriées pour les téléphones mobiles, mais les principes fondamentaux existent. Il faudra aussi résoudre le problème que pose la combinaison de données obtenues à partir de téléphones mobiles, qui sont

fondamentalement des appareils personnels, à celles obtenues à partir de téléphones fixes, qui sont essentiellement des appareils de ménage, pour s'assurer que la pondération soit correcte. Pour cela, il faudra recueillir suffisamment de données complètes sur tous les appareils de communication dont dispose le ménage.

Pour conclure, la méthodologie des téléenquêtes, dont l'évolution au cours des dernières décennies a rendu les enquêtes téléphoniques viables et en a fait l'instrument prédominant d'enquête, devra être mise à jour continuellement en vue de s'adapter à l'évolution constante des technologies de télécommunication et de leur utilisation. Cependant, les éléments méthodologiques fondamentaux de cette évolution existent et continueront de permettre d'utiliser des solutions de pointe pour obtenir des données d'enquête de haute qualité.

BIBLIOGRAPHIE

ABRAHAM, S.Y., STEIGER, D.M. et SULLIVAN, C. (1998). Electronic and mail self-administered questionnaires : A comparative assessment of use among elite populations. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 833-841.

ALEXANDER, C.H. (1988). Cutoff rules for secondary calling in a random digit dialing survey. *Telephone Survey Methodology*, (R.M. Groves et coll. - éds.). New York: John Wiley and Sons, 113-126.

AMERICAN STATISTICAL ASSOCIATION (1999). *More About Telephone Surveys*. ASA series: what is a survey? Section on Survey Research Methods [http://www.amstat.org/sections/srms/brochures/telephone.pdf].

ANDERSON, J.E., NELSON, D.E. et WILSON, R.W. (1998). Telephone coverage and measurement of health risk indicators: data from the National Health Interview Survey. *American Journal of Public Health*, 88, 1392-1395.

AQUILINO, W.S., et LO SCIUTO, L.A. (1990). Effects of interview mode on self-reported drug use. *Public Opinion Quarterly*, 54, 362-395.

BAKER, R.P. (1998). The CASIC future. *Computer Assisted Survey Information Collection*, (M.P. Couper et coll. - éds.). NewYork: John Wiley and Sons, 583-604.

BANKS, M.J., et HAGAN, D.E. (1984). Reducing interviewer screening and controlling sample size in a local-area telephone survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 271-273.

BANKS, R., CHRISTIE, C., CURRALL, J., FRANCIS, J., HARRIS, P., LEE, B., MARTIN, J., PAYNE, C. et WESTLAKE, A. (Éds.) (1999). ASC'99 – Leading Survey & Statistical Computing into the New Millennium. *Proceedings of the ASC International Conference*. Association for Survey Computing Chesham, Bucks, UK.

BATTAGLIA, M.P., SHAPIRO, G. et ZELL, E.R. (1996). Substantial response bias may remain when records are used in a telephone survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 452-455.

BENNET, C.T. (1961). A telephone interview: A method for conducting a follow-up study. *Mental Hygiene*, 45, 216-220.

BERCINI, D.H., et MASSEY, J.T. (1979). Obtaining the household roster in a telephone survey: The impact of names and placement on response rates. *Proceedings of the Social Statistics Section, American Statistical Association*, 136-140.

BERRY, S.H., et O'ROURKE, D. (1998). Administrative designs for centralized telephone survey centers: Implications of the transition to CATI. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 457-474.

BIEL, A. L. (1967). Abuses of survey research techniques: the phony interview. *Public Opinion Quarterly*, 31, 298.

BIEMER, P.P. (1983). Optimal dual frame sample design: Results of a simulation study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 630-635.

BINSON, D., CANCHOLA, J.A. et CATANIA, J.A. (2000). Random selection in a national telephone survey: A comparison of the Kish, next-birthday, and last-birthday methods. *Journal of Official Statistics*, 16, 53-59.

BLAIR, J., et CZAJA, R. (1982). Locating a special population using random digit dialing. *Public Opinion Quarterly*, 46, 585-590.

BLANKENSHIP, A.B. (1977a). *Professional Telephone Surveys*. New York: McGraw Hill.

BLANKENSHIP, A.B. (1977b). Listed versus unlisted numbers in telephone-survey samples. *Journal of Advertising Research*, 39-42.

BOTMAN, S.L., et ALLEN, K. (1990). Some effects of undercoverage in a telephone survey of teenagers. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 396-400.

BOTMAN, S.L., MASSEY, J.T. et SHIMIZU, I.M. (1982). Effect of weighting adjustments on estimates from a random-digit-dialed telephone survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 139-144.

BRICK, J.M. (1990). Multiplicity sampling in an RDD telephone survey. *Proceedings of the Section on Survey Research Methodology, American Statistical Association*, 296-301.

BRICK, J.M., et COLLINS, M.A. (1997). A response rate experiment for RDD surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1052-1057.

BRICK, J.M., KALTON, G., NIXON, M., GIVENS, J. et EZZATI-RICE, T. (2000). Statistical issues in a record check study of childhood immunizations. *Proceedings of the 1999 Federal Committee on Statistical Methodology Research Conference: Statistical Policy Working Paper*, 30, 625-634.

BRICK, J.M., et WAKSBERG, J. (1991). Méthode pour éviter l'échantillonnage progressif dans une enquête téléphonique à composition aléatoire. *Techniques d'enquête*, 17, 31-46.

BRICK, J.M., WAKSBERG, J. et KEETER, S. (1996). Utilisation des données sur les interruptions du service téléphonique pour ajuster la couverture. *Techniques d'enquête*, 22, 187-199.

BRICK, J. M., WAKSBERG, J., KULP, D. et STARER, A. (1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly*, 59, 218-235.

BRUNNER, J.A., et BRUNNER, G.A. (1971). Are voluntary unlisted telephone subscribers really different? *Journal of Marketing Research*, 8, 121-124.

BRYANT, B.E. (1975). Respondent selection in a time of changing household composition. *Journal of Marketing Research*, 12, 129-135.

BRYSON, M.C. (1976). The literary digest poll: Making of a statistical myth. *The American Statistician*, 30, 184-185.

BULL, S.B., PEDERSON, L.L. et ASHLEY, M.J. (1988). Intensity of follow up; effects on estimates in a population telephone survey with an extension of Kish's (1965) approach. *American Journal of Epidemiology*, 127, 552-561.

BURKE, J., MORGANSTEIN, D. et SCHWARTZ, S. (1981). Toward the design of an optimal telephone sample. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 448-453.

CAHALAN, D. (1960). Measuring newspaper readership by telephone: two comparisons with face to face interviews. *Journal of Advertising Research*, 1, 2, 1-6.

CAHALAN, D. (1989). Comment: The digest poll rides again! *Public Opinion Quarterly*, 53, 129-133.

CAMPBELL, J., et PALIT, C.D. (1988). Total digit dialing for a small area census by phone. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 549-551.

CANNELL, C.F., GROVES, R.M., MAGILAVY, L.J., MATHIOWETZ, N.A., MILLER, P.V. et THORNBERRY, O.T. (1987). An experimental comparison of telephone and personal health interview surveys. *Vital and Health Statistics*, Series 2, 106, Public Health service.

CARON, P., et LAVALLÉE, P. (1998). Comparison study on the quality of financial data collected through personal and telephone interviews. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 208-213.

CASADY, R.J., et LEPKOWSKI, J.M. (1991). Optimal allocation for stratified telephone survey design. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 111-116.

CASADY, R.J., et LEPKOWSKI, J.M. (1993). Plans d'enquête téléphonique stratifiée. *Techniques d'enquête*, 19, 115-125.

CASADY, R.J., et LEPKOWSKI, J.M. (1998). Telephone sampling. *Encyclopedia of Biostatistics*. New York: John Wiley and Sons, 4498-4511.

CASADY, R.J., et LEPKOWSKI, J.M. (1999). Telephone sampling. *Sampling of Populations: Methods and Applications – third edition*, (P.S. Levy and S. Lemeshow, éds.). New York: John Wiley and Sons, 455-479.

CASADY, R.J., et SIRKEN, M.G. (1980). A multiplicity estimator for multiple frame sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 601-609.

CASRO (1982). *Report of the Council of American Survey Research Organization Completion Rate Task Force*. New York: Audits and Surveys Inc. (Rapport non-publié).

CENTRAL BUREAU OF STATISTICS (2000). *The Household Expenditure Survey 1999*. Special Publication 1147. Jerusalem.

CHAPMAN, D.W., et ROMAN, A.M. (1985). An investigation of substitution for an RDD survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 269-274.

CHOUDHRY, G.H. (1989). Cost-variable optimization of dual frame design for estimating proportions. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 566-571.

CLAYTON, R.L., et WINTER, D. L. S. (1992). Speech data entry: results of a test of voice recognition for survey data collection. *Journal of Official Statistics*, 8, 377-388.

COLLINS, M. (1983). Computer assisted telephone interviewing in the UK. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 636-639.

COLLINS, M. (1999). Editorial: sampling for UK telephone surveys. *Journal of the Royal Statistical Society*, A 162, 1-4.

COLLINS, M., SYKES, W., WILSON, P. et BLACKSHAW, N. (1988). Nonresponse: the UK experience. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 213-232.

COLOMBOTOS, J. (1965). The effects of personal vs. telephone interviews on socially acceptable responses. *Public Opinion Quarterly*, 29, 457-458.

COOMBS, L., et FREEDMAN, R. (1964). Use of telephone interviews in a longitudinal fertility study. *Public Opinion Quarterly*, 28, 112-117.

COOPER, S. L. (1964). Random sampling by telephone: an improved method. *Journal of Marketing Research*, 1, 45-48.

COUPER, M.P., BAKER, R.P., BETHLEHEM, J., CLARK, C.Z F., MARTIN, J., NICHOLLS, W.L., II et O'REILLY, J.M. - (Éds.) (1998). *Computer Assisted Survey Information Collection*. New York: John Wiley and Sons.

COUPER, M.P., BLAIR, J. et TRIPLETT, T. (1999). A comparison of mail and e-mail for a survey of employees in US statistical agencies. *Journal of Official Statistics*, 15, 39-56.

COUPER, M.P., et NICHOLLS, W.L., II (1998). The history and development of computer assisted survey information collection methods. *Computer Assisted Survey Information Collection*, (M.P. Couper, et coll. - éds.). NewYork: John Wiley and Sons, 1-22.

CUNNINGHAM, J.M., WESTERMAN, H.H. et FISCHOFF, J. (1956). A follow-up study of patients seen in a psychiatric clinic for children. *American Journal of Orthopsychiatry*, 26, 602-610.

CUNNINGHAM, P., BERLIN, M., MEADER, J., MOLLOY, K., MOORE, D. et PAJUNEN, S. (1997). Using cellular telephones to interview nontelephone households. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 250-254.

CUNNINGHAM, P., BRICK, J.M. et MEADER, J. (2000). 1999 *NSAF In-Person Survey Methods Report No. 5*. Washington, DC: Urban Institute. [http://newfederalism.urban.org/nsaf/ methodology_rpts/ 1999_Methodology_5.pdf].

CUMMINGS, M.K. (1979). Random digit dialing: a sampling technique for telephone surveys. *Public Opinion Quarterly*, 43, 233-244.

CZAJA, R., BLAIR, J. et SEBESTIK, J.P. (1982). Respondent selection in a telephone survey: a comparison of three techniques, *Journal of Marketing Research*, 19, 381-385.

DECISION ANALYST (1997). More households using answering machines. *News Release, October 15, 1997.* [http://www.decisionanalyst.com/publ_data/1997/ansmachi.htm].

DEKKER, F., et DORN, P.K. (1984). Computer Assisted Telephone Interviewing: A research project in the Netherlands. Article présenté à la *Conference of the Institute of British Geographers*.

DE LEEUW, E.D., et VAN DER ZOOWEN, J. (1988). Data quality in telephone and face to face surveys: a comparative meta-analysis. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 283-299.

DIEHR, P., KOEPSELL, T.D., CHEADLE, A. et PSATY, B.M. (1992). Assessing response bias in random-digit dialing surveys: The telephone-prefix method. *Statistics in Medicine*, 11, 1009-1021.

DILLMAN, D.A. (1978). *Mail and Telephone Surveys: the Total Design Method*. New York: John Wiley and Sons.

DILLMAN D. A. (2000). *Mail and Internet Surveys: the Total Design Method* (2e édition). New York: John Wiley and Sons.

DREW, J.D., CHOUDHRY, G.H. and HUNTER, L.A. (1988). Nonresponse issues in government telephone surveys. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 233-246.

DREW, J.H., et GROVES, R.M. (1989). Adjusting for nonresponse in a telephone subscriber survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 452-456.

DUTKA, S., et FRANKEL, L. R. (1980). Sequential survey design through the use of computer assisted telephone interviewing. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 73-76.

EASTLACK, J.O., JR. (1964). Recall of advertising by two telephone samples. *Journal of Advertising Research*, 4, 25-29.

EASTLACK, J.O., JR., et ASSAEL, H. (1966). Better telephone surveys through centralized interviewing. *Journal of Advertising Research*, 6, 1, 2-7.

FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY. (1984). *The Role of Telephone Data Collection in Federal Statistics*. Statistical Policy Working Paper 12, Washington, D.C.

FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY. (1990). *Computer Assisted Survey Information Collection*, Statistical Policy Working Paper 19, Washington, D.C.

FEDERAL REPUBLIC OF GERMANY (1999). Continuous family budget surveys for January 1999. Statistiches Bundesamt: *Press release*, 20 December, 1999. [http://www.statistik-bund.de/presse/englisch/pm1999/p4350024.htm].

FELSON, L. (2001). Netting limitations: online researchers' new tactics for tough audiences. Marketing News (*American Marketing Association*), 35, 5 [http://www.ama.org/pubs/article.asp?id=4881].

FINK, J.C. (1983). CATI's first decade: The Chilton experience. *Sociological Methods and Research*, 12, 153-168.

FISCHBACHER, C., CHAPPEL, D., EDWARDS, R. et SUMMERTON, N. (1999). The use and abuse of the Internet for survey research. *Proceedings of the Association for Survey Computing 3rd International Conference*, Edinburgh, 501-507.

FITTI, J.E. (1979). Some results from the telephone health interview system. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 244-249.

FOLEY, D.J., et BROCK, D.B. (1990). Comparison of in-person and telephone responses in a survey of the last days of life. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 382-386.

FORD, E.S. (1998). Characteristics of survey participants with and without a telephone; findings from the third National Health and Nutrition Examination Survey. *Journal of Clinical Epidemiology*, 51, 55-60.

FORSMAN, G. (1993). Sampling individuals within households in telephone surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1113-1118.

FORSMAN, G., et DANIELSSON, S. (1997). Can plus digit sampling generate a probability sample? *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 958-963.

FOX, A., et RILEY, J. P.(1996). Telephone coverage, housing quality and rents: RDD survey biases. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 515-519.

FRANKEL, M.R., SRINATH, K.P., BATTAGLIA, M.P., HOAGLIN, D.C., WRIGHT, R.A. et SMITH, P.J. (1999). Reducing nontelephone bias in RDD surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 934-937.

FREEMAN, H.E., et SHANKS, J. M. - Éds. (1983). The emergence of computer-assisted survey research. *Sociological Methods and Research*, 12 (numéro spéciale), 115-230.

FRÉJEAN, M., PANZANI, J-P. et TASSI, P. (1990). Les ménages inscrits en liste rouge et les enquêtes par téléphone. *Journal de la Société de Statistique de Paris*, 131, Nos. 3-4, 86-102.

FREY J. H. (1989). *Survey Research by Telephone* (2e édition). Beverly Hills, CA: Sage Publications.

FRY, H.G., et MCNAIRE, S. (1958). Data gathering by long distance telephone. *Public Health Records*, 73, 831-835.

GABLER, S., et HAEDER, S. (2000). Telephone sampling in Germany. *Paper presented at Fifth International Conference of Social Science Methodology*, Koln.

GENESYS (1996). Unlisted numbers: what's really important. *Genesys News* (Genesys Sampling Systems, Fort Washington, PA), 1-2.

GHOSH, D. (1984). Improving the plus 1 method of random digit dialing. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 285-288.

GIESBRECHT, L.H., KULP, D.W. et STARER, A.W. (1996). Estimating coverage bias in RDD samples with current population survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 503-508.

GLASSER, G.J., et METZGER, G.D. (1972). Random digit dialing as a method of telephone sampling. *Journal of Marketing Research*, 9, 59-64.

GLASSER, G. J., et METZGER, G. D. (1975). National estimates of nonlisted telephone households. *Journal of Marketing Research*, 12, 359-361.

GOKSEL, H., JUDKINS, D.R. et MOSHER, W.D. (1991). Nonresponse adjustments for a telephone follow-up to a national in-person survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 581-586.

GROVES, R.M. (1977). *An Empirical Comparison of Two Telephone Sample Designs*. Rapport non publié Survey Research Center, the University of Michigan, Ann Arbor, MI.

GROVES, R.M., BIEMER, P.P., LYBERG, L.E., MASSEY, J.T., NICHOLLS, W.L., II et WAKSBERG, J. - éds. (1988). *Telephone Survey Methodology*. New York: John Wiley and Sons.

GROVES R.M., et KAHN, R.L. (1979). *Surveys by Telephone: A National Comparison With Personal Interview*. New York: Academic Press.

GROVES, R.M., et LEPKOWSKI, J.M. (1985). Dual frame, mixed mode survey designs. *Journal of Official Statistics*, 1, 264-286.

GROVES, R.M., et LEPKOWSKI, J.M. (1986). An experimental implementation of a dual frame telephone sample design. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 340-345.

GROVES, R.M., et LYBERG, L.E. (1988a). An overview of nonresponse issues in telephone surveys. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 191-211.

GROVES, R.M., and LYBERG, L.E. - Éds. (1988b). Telephone survey methodology. *Journal of Official Statistics* (numéro spéciale), 4, 283-416.

GUNN, W.J., et RHODES, I.N. (1981). Physician response rates to a telephone survey: effects of monetary incentive level. *Public Opinion Quarterly*, 45, 109-115.

HAGAN, D. E., et MEIER C. C. (1983). Must respondent selection procedures for telephone surveys be invasive? *Public Opinion Quarterly*, 47, 547-556.

HARLOW, B.L., CREA, E.C., EAST, M.A., OLESON, B., FRAER, C.J. et CRAMER, D.W. (1993). Telephone answering machines: the influence of leaving messages on telephone interviewing response rates. *Journal of Epidemiology*, 4, 380-383.

HARTGE, P., BRINTON, L.A., ROSENTHAL, J.F., CAHILL, J.I., HOOVER, R.N. et WAKSBERG, J. (1984). Random digit dialing in selecting a population-based control group. *American Journal of Epidemiology*, 120, 825-833.

HAUCK, M., et COX, M. (1974). Locating a sample by random digit dialing. *Public Opinion Quarterly*, 38, 253-260.

HERZOG, A.R., et RODGERS, W.L. (1988). Interviewing older adults: mode comparison using data from a face-to-face survey and a telephone resurvey. *Public Opinion Quarterly*, 52, 84-99.

HOAGLIN, D.C., et BATTAGLIA, M.P. (1996). A comparison of two methods of adjusting for noncoverage of nontelephone households in a telephone survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 497-502.

HOCHSTIM, J.R. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association*, 62, 976-989.

HOGUE, C.R., et CHAPMAN, D.W. (1984). An investigation of PSU cutoff points for a random digit dialing survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 286-291.

HOX, J.J., DE LEEUW, E.D. et KREFT, I.G.G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. *Measurement errors in surveys* (P.P. Biemer, L.E. Lyberg, N.A. Mathiowetz et S. Sudman - éds.). New York: John Wiley and Sons, 439-461.

INGLIS, K.M., GROVES, R.M. et HEERINGA, S.G. (1987). Plans de sondage d'enquêtes téléphoniques auprès de ménages noirs aux États-Unis. *Techniques d'enquête*, 13, 1-17.

JANOFSKY, A.I. (1971). Affective self-disclosure in telephone versus face-toface interviews. *Journal of Humanistic Psychology*, 11, 93-103.

JOHNSON, T., FENDRICH, M., SHALIGRAM, C. et GAREY, A. (1997). A comparison of interviewer effects models in an RDD telephone survey of drug use. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 887-892.

KALSBEEK, W.D., et DURHAM, T.A. (1994). Nonresponse and its effects in a followup telephone survey of low-income women. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 943-948.

KALTON, G. (2000). L'évolution de la recherche sur les enquêtes au cours des 25 dernières années. *Techniques d'enquête*, 26, 3-11.

KATZ, D., et CANTRIL, H. (1937). Public opinion polls. *Sociometry*, 1, 155-179.

KATZ, J.E., et ASPDEN, P. (1998). Internet dropouts in the USA. *Telecommunications Policy*, 22, 4/5, 327-339.

KEETER, S. (1995). Estimating telephone noncoverage bias with a telephone survey. *Public Opinion Quarterly*, 59, 196-217.

KEHOE, C., PITKOW, J., SUTTON, K., AGGARWAL, G. et ROGERS, J.D. (1999). *Results of GVU's Tenth World Wide Web User Survey*. Atlanta, GA: Graphics Visualization and Usability Center, College of Computing, Georgia Institute of Technology. [http://www.gvu.gatech.edu/user_surveys].

KHURSHID, A., et SAHAI, H. (1995). A bibliography on telephone survey methodology. *Journal of Official Statistics*, 11, 325-367.

KISH, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, 44, 380-387.

KOEPSELL, T.D., MCGUIRE, V., LONGSTRETH, Jr., W.T., NELSON, L.M. et VAN BELLE, G. (1996). Randomized trial of leaving messages on telephone answering machines for control recruitment in an epidemiological study. *American Journal of Epidemiology*, 144, 704-706.

KÖRMENDI, E. (1988). The quality of income information in telephone and face to face surveys. *Telephone Survey Methodology* (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 341-356.

KUUSELA, V., et NOTKOLA, V. (1999). Survey quality and mobile phones. Article présenté à l'*International Conference on Survey Nonresponse*, Portland OR.

KUUSELA, V., et VIKKI, K. (1999). Change of telephone coverage due to mobile phones. Article présenté à l'*International Conference on Survey Nonresponse*, Portland OR.

LANDON, E.L., JR., et BANKS, S.K. (1977). Relative efficiency and bias of plus-one telephone sampling. *Journal of Marketing Research*, 14, 294-299.

LARSON, O. N. (1952). The comparative validity of telephone and face-to-face interviews in the measurement of message diffusion from leaflets. *American Sociological Review*, 17, 471-476.

LAVRAKAS, P.J. (1993). *Telephone Survey Methods: Sampling, Selection and Supervision* (2e édition). Newbury Park, CA: Sage Publications.

LAVRAKAS, P.J., BAUMAN, S.L. et MERKLE, D.M. (1993). The last-birthday method and within-unit coverage problems. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1107-1112.

LEPKOWSKI, J.M. (1988). Telephone sampling methods in the United States. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 73-98.

LEPKOWSKI, J.M., et GROVES, R.M. (1984). The impact of bias on dual frame survey design. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 265-270.

LEPKOWSKI, J.M., et GROVES, R.M. (1986a). A two phase probability proportional to size design for telephone sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 357-362.

LEPKOWSKI, J.M., et GROVES, R.M. (1986b). A mean square error model for dual frame, mixed mode survey design. *Journal of the American Statistical Association*, 81, 930-937.

LEUTHOLD, D.A., et SCHEELE, R. (1971). Patterns of bias in samples based on telephone directories. *Public Opinion Quarterly*, 35, 249-257.

LOCANDER, W., SUDMAN, S. et BRADBURN, N. (1976). An investigation of interview method, threat and response distortion. *Journal of the American Statistical Association*, 71, 269-275.

LYBERG, L., et KASPRZK, D. (1991). Data collection methods and measurement error: an overview. *Measurement Errors in Surveys* (P.P. Biemer, L.E. Lyberg, N.A. Mathiowetz et S. Sudman - éds.). New York: John Wiley and Sons, 237-258.

MCCARTHY, W.F., et BATEMAN, D.V. (1988). The use of mathematical programming for designing dual frame surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 652-653.

MCKAY, R.B., ROBISON, E.L. et MALIK, A.B. (1994). Touch-tone data entry for household surveys: research findings and possible applications. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 509-511.

MAFFEO, C., FREY, W. et KALTON, G. (2000). Survey design and data collection issues in the Disability Evaluation Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, forthcoming.

MAKLAN, D., et WAKSBERG, J. (1988). Within household coverage in RDD surveys. *Telephone Survey Methodology* (R. M. Groves, et coll. - éds). New York: John Wiley and Sons, 51-69.

MALAKHOFF, L.A., et APPEL, M.V. (1997). The development of a voice recognition prototype for field listing. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 234-238.

MASON, R.E., et IMMERMAN, F.W. (1988). Minimum cost sample allocation for Mitofsky-Waksberg random digit dialing. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 127-141.

MASSEY, J.T. (1995). Estimating the response rate in a telephone survey with screening. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 673-677.

MASSEY, J T., et BOTMAN, S.L. (1988). Weighting adjustments for random digit dialed surveys. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 143-160.

MASSEY, J.T., O'CONNOR, D. et KROTKI, K. (1997). Response rates in random digit dialing (RDD) telephone surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 707-712.

MEEKS, R.L., LANIER, A.T., FECSO, R.S. et COLLINS, M.A. (1998). Web-based data collection in national science foundation surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 349-353.

MERKLE, D.M., BAUMAN, S.L. et LAVRAKAS, P.J. (1993). The impact of callbacks on survey estimates in an annual RDD survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1070-1075.

MITCHELL, G.H., et ROGERS, E.M. (1958). Telephone interviewing. *Journal of Farm Economics*, 40, 743-747.

MITOFSKY, W. (1970). Sampling of Telephone Households. Mémorandum de CBS non publié.

MOHADJER, L. (1988). Stratification of prefix areas for sampling rare populations. *Telephone Survey Methodology*, (R.M. Groves, et coll.- éds.). New York: John Wiley and Sons, 161-173.

MULLET, G.M. (1982). The efficacy of plus-one dialing: self-reported status. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 575-576.

NATHAN, G., et AFRAMIAN, N. (1996). An experiment with CATI in Israel. Article présenté à l'*InterCasic '96 Conference*, San Antonio, TX.

NATHAN, G., et ELIAV, T. (1988). Comparison of measurement errors for telephone interviewing and home visits by misclassification models. *Journal of Official Statistics*, 4, 363-374.

NICHOLLS, W.L., II (1983). CATI research and development at the Census Bureau. *Sociological Methods and Research*, 12, 191-198.

NICHOLLS, W.L., II (1988). Computer-assisted telephone interviewing: A general introduction. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 377-386.

NICOLAAS, G., LYNN, P. et LOUND, C. (2000). Random digit dialling in the UK: viability of the sampling method revisited. Article présenté à *Fifth International Conference of Social Science Methodology*, Koln.

NORRIS, D.A., et PATON, D.G. (1991). L'Enquête sociale générale canadienne : bilan des cinq premières années. *Techniques d'enquête*, 17, 245-260.

NTIA (2000). *Falling Through the Net, Toward Digital Inclusion.* Washington DC: National Telecommunications and Information Administration.

NUSSER, S., et THOMPSON, D. (1998). Web-based survey tools. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 951-956.

OAKES, R.H. (1954). Differences in responsiveness in telephone versus personal interviews. *Journal of Marketing*, 19, 169.

OKSENBERG, L., et CANNEL, C. (1988). Effects of interviewer vocal characteristics on nonresponse. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 257-269.

OLDENDICK, R.W., BISHOP, G.F., SORENSON, S.B. et TUCHFARBER, A.J. (1988). A comparison of the Kish and last birthday methods of respondent selection in telephone surveys. *Journal of Official Statistics*, 4, 307-318.

OLDENDICK R.W., et LINK, M.W. (1994). The answering machine generation: who are they and what problem do they pose for survey research? *Public Opinion Quarterly*, 58, 264-273.

OFTEL (1999). *Homes Without a Fixed Line Phone - Who Are They?* [http://www.oftel.gov.uk/publications/research/unph0400.htm].

OFTEL (2000). *Consumers' use of Internet.* [http://www.oftel.gov.uk/publications/research/int1000.htm]

O'REILLY, J.M., HUBBARD, M.L., LESSLER, J.T., BIEMER, P.P. et TURNER, C.F. (1994). Audio and video computer assisted self-interviewing; preliminary tests of new technologies for data collection. *Journal of Official Statistics*, 10, 197-214.

O'ROURKE, D., et BLAIR, J. (1983). Improving random respondent selection in telephone surveys. *Journal of Marketing Research*, 20, 428-432.

PALIT, C.D. (1980). A microcomputer based computer assisted interviewing system. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 243-244.

PALIT, C.D. (1983). Design strategies in RDD sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 627-629.

PALIT, C.D., et BLAIR, J. (1986). Some alternatives for the treatment of first phase telephone numbers in a Waksberg-Mitofsky RDD sample. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 363-369.

PALIT, C.D., et SHARP, H. (1983). Microcomputer-assisted telephone interviewing. *Sociological Methods and Research*, 12, 169-189.

PANNEKOEK, J. (1988), Interviewer variance in a telephone survey. *Journal of Official Statistics*, 4, 375-384.

PAYNE, S. L. (1956). Some advantages of telephone surveys. *Journal of Marketing*, 20, 278-281.

PERNEGER, T.V., MYERS, T.L., KLAG, M.J. et WHELTON, P.K. (1993). Effectiveness of the Waksberg telephone sampling method for the selection of population controls. *American Journal of Epidemiology*, 138, 574-584.

PERONE, C., MATRUNDOLA, G. et SOVERINI, M. (1999). A quality control approach to mobile phone surveys; the experience of Telecom Italia Mobile. *Proceedings of the Association for Survey Computing 3rd International Conference*, Edinburgh, 180-187.

PERRY, J. B. (1968). A note on the use of telephone directories as a sample source. *Public Opinion Quarterly*, 32, 691-695.

PHIPPS, P.A., et TUPEK, A.R.(1991). Fiabilité des données introduites au moyen d'un téléphone à clavier. *Techniques d'enquête*, 17, 17-30.

PIAZZA, T. (1993). Meeting the challenge of answering machines. *Public Opinion Quarterly*, 57, 219-231.

POTTER, F.J., MCNEILL, J.J., WILLIAMS, S.R. et WAITMAN, M.A. (1991). List-assisted RDD telephone surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 117-122.

POTTHOFF, R.F. (1987a). Some generalizations of the Mitofsky-Waksberg technique of random digit dialing. *Journal of the American Statistical Association*, 82, 409-418.

POTTHOFF, R.F. (1987b). Generalizations of the Mitofsky-Waksberg technique for random digit dialing: some added topics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 615-620.

POYNTER, R. (2000). *We've Got Five Years.* London: Association for Survey Computing meeting on Survey Research on the Internet.

RAMOS, M., SEDIVI, B.M. et SWEET, E.M. (1998). Computerized self-administered questionnaires. *Computer Assisted Survey Information Collection* (M.P. Couper, et coll. - éds.). NewYork: John Wiley and Sons, 389-408.

RANTA-AHO, M., et LEPPINEN, A. (1997). Matching telecommunication services with user communication needs. *Proceedings of the International Symposium on Human Factors in Telecommunications*, (K. Nordby et L. Grafisk - éds.). Oslo, Norway, 401-408. [http://www.comlab.hut.fi/hft/publications/matcharticle.pdf].

RICH, C.L. (1977). Is random digit dialing really necessary? *Journal of Marketing Research*, 14, 300-305.

ROGERS, T.F. (1976). Interviews by telephone and in person: quality of responses and field performance. *Public Opinion Quarterly*, 40, 51-65.

ROGERS, S.M., MILLER, H.G., FORSYTH, B.H., SMITH, T.K. et TURNER, C.F. (1996). Audio-CASI: the impact of operational characteristics on data quality. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1042-1047.

ROMUALD, K.S., et HAGGARD, L.M.(1994). The effect of varying the respondent selection script on respondent self-selection in RDD telephone surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1299-1304.

ROSEN, R.J., MANNING, C.D. et HARRELL, L.J., Jr. (1998). Web-based data collection in the current employment statistics survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 354-359.

ROSLOW, S., et ROSLOW, L. (1972). Unlisted phone subscribers are different. *Journal of Advertising Research*, 7, 4, 35-38.

ROUQUETTE, C. (2000). La percée du telephone portable et d'Internet. *INSEE Première No. 200. INSEE*, Paris. [http://www.insee.fr/fr/ffc/docs_ffc/ip700.pdf].

ST. CLAIR, J., et MUIR, J. (1997). Household adoption of digital technologies. *Year Book Australia 1997*. Canberra: Australian Bureau of Statistics.

SALMON, C.T., et NICHOLS, J.S. (1983). The next birthday method of respondent selection. *Public Opinion Quarterly*, 47, 270- 276.

SCHEUREN, F., et PETSKA, T. (1993). Turning administrative systems into information systems. *Journal of Official Statistics*, 9, 109-119.

SCHMIEDESKAMP, J. W. (1962). Reinterviews by telephone. *Journal of Marketing*, 26, 28-34.

SEBOLD, J. (1988). Survey period length, unanswered numbers, and nonresponse in telephone surveys. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 247-256.

SHANKS, J.M. (1983). The current status of computer assisted telephone interviewing: recent progress and future prospects. *Sociological Methods and Research*, 12, 119-142.

SHANKS, J.M., NICHOLLS, W.L., II et FREEMAN, H.E. (1981). The California Disability Survey: design and execution of a computer-assisted telephone study. *Sociological Methods and Research*, 10, 123-140.

SHAPIRO, G. M., BATTAGLIA, M. P., HOAGLIN, D. C., BUCKLEY, P. et MASSEY, J. T. (1996). Geographical variation in within-household coverage of households with telephones in an RDD survey. *Proceedings of the Section on Survey Research Methods*, 491-496.

SHURE, G.E., et MEEKER, R.J. (1978). A mini-computer system for multi-person computer-assisted telephone interviewing. *Behavior Methods and Instrumentation*. (Avril) 196-202.

SMITH, C., et FRAZIER, E. L. (1993). Comparison of traditional and modified Waksberg. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 926-931.

SQUIRE, P. (1988). Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly*, 52, 125-133.

STATISTICS CANADA (2000). *Selected Dwelling Characteristics and Household Equipment*. Income Statistics Division. [http://www.statcan.ca/english/Pgdb/People/Families/famil09b.htm].

STOCK, J. S. (1962). How to improve samples based on telephone listings. *Journal of Advertising Research*, 2, 3, 50-51.

STOKES, L., et YEH, M.-Y. (1988). Searching for causes of interviewer effects in telephone surveys. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 357-376.

SUDMAN, S. (1966). New uses of telephone methods in survey research. *Journal of Marketing Research*, 3, 107-120.

SUDMAN, S. (1973). The uses of telephone directories for survey sampling. *Journal of Marketing Research*, 10, 204-207.

SUDMAN, S. (1978). Optimum cluster designs within a primary unit using combined telephone screening and face-to-face interviewing. *Journal of the American Statistical Association*, 73, 300-304.

SURVEY RESEARCH CENTER (2000). *Sample Design for Household Telephone Surveys: A Bibliography* 1949-1996. College Park, MD: University of Maryland. [http://www.bsos.umd.edu/ src/sampbib.html].

SURVEY SAMPLING INC. (1998). Random digit samples – part 1. [http://www.ssisamples.com/ssi.x2o$ssi_gen.search_item?id=119].

STATISTICS NETHERLANDS (1987). *Automation in Survey Processing*. Voorburg/Heerlen: Netherlands Central Bureau of Statistics (CBS Select 4).

SYKES, W.M., et COLLINS, M. (1987). Comparaison entre l'interview téléphonique et l'interview sur place au Royaume-Uni. *Techniques d'enquête* 3, 19-33.

SYKES, W.M., et COLLINS, M. (1988). Effects of mode of interview: experiments in the UK. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 301-320.

THORNBERRY, O.T. JR., et MASSEY, J.T. (1978). Correcting for undercoverage bias in random digit dialed National Health Surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 224-229.

THORNBERRY, O.T. JR., et MASSEY, J.T. (1983). Coverage and response in random digit dialed national surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 654-659.

THORNBERRY, O.T. JR., et MASSEY, J.T. (1988). Trends in United States telephone coverage across time and subgroups. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 25-49.

TORTORA, R.D. (1985). CATI in an agricultural statistics agency. *Journal of Official Statistics*, 1, 301-314.

TOURANGEAU, R., et SMITH, T.W. (1998). Collecting sensitive information with different modes of data collection. *Computer Assisted Survey Information Collection*, (M.P. Cooper, et coll. - éds.). NewYork: John Wiley and Sons, 431-453.

TRAUGOTT, M.W., GROVES, R.M. et LEPKOWSKI, J.M. (1987). Using dual frame designs to reduce nonresponse in telephone surveys. *Public Opinion Quarterly*, 51, 522-539.

TREWIN, D., et LEE, G. (1988). International comparisons of telephone coverage. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 3-24.

TROLDAHL, V.C., et CARTER, R.E. (1964). Random selection of respondents within households in phone surveys. *Journal of Marketing Research*, 1, 71-76.

TUCKEL, P.S., et FEINBERG, B.M. (1991) The answering machine poses many questions for telephone survey researchers. *Public Opinion Quarterly*, 55, 200-217.

TUCKEL, P.S., et O'NEILL, H. (1996), New technology and nonresponse bias in RDD surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 889-894.

TUCKEL, P., et SHUKERS, T. (1997). The effect of different introductions and answering machine messages on response rates. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1047-1051.

TUCKER, C., CASADY, R. et LEPKOWSKI, J. (1992). Sample allocation for stratified telephone sample designs. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 291-296.

TURNER, C.F., FORSYTH, B.H., O'REILLY, J.M., COOLEY, P. C., SMITH, T.K., ROGERS, S.M. et MILLER, H.G. (1998). Automated self-interviewing and the survey measurement of sensitive behaviors. *Computer Assisted Survey Information Collection*, (M.P. Cooper, et coll. - éds.). NewYork: John Wiley and Sons, 455-473.

WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

WAKSBERG, J. (1983). A note on 'Locating a special population using random digit dialing'. *Public Opinion Quarterly*, 47, 576-578.

WAKSBERG J. (1984). *Efficiency of Alternative Methods of Establishing Cluster Sizes in RDD Sampling*. Mémorandum de Westat non publié.

WAKSBERG, J., BRICK, J.M., SHAPIRO, G., FLORES-CERVANTES, I. et BELL, B. (1997). Dual-frame RDD and area sample for household survey with particular focus on low-income population. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 713-718.

WERKING, G., TUPEK, A. R., et CLAYTON, R. L. (1988). CATI and touchtone self-response applications for establishment surveys. *Journal of Official Statistics* 4, 349-362.

WHITE, A. A. (1983). Response rate calculation in RDD telephone health surveys: current practices. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 277-282.

WHITMORE, R. W., MASON, R. E. et HARTWELL, T. D. (1985). Use of geographically classified telephone directory lists in multi-mode surveys. *Journal of the American Statistical Association* 80, 842-844.

WILSON, P., BLACKSHAW, N. et NORRIS P. (1988). An evaluation of telephone interviewing on the British Labour Force Survey. *Journal of Official Statistics* 4, 385-400.

WINTER, D. L. S., et CLAYTON, R. L. (1990). Speech data entry: results of the first test of voice recognition for data collection. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 387-392.

WISEMAN, F. (1972). Methodological bias in public opinion surveys. *Public Opinion Quarterly* 36, 105-108.

XU, M., BATES, B.J. et SCHWEITZER, J.C. (1993). The impact of messages on survey participation in answering machine households. *Public Opinion Quarterly*. 57, 232-237.

Techniques d'enquête, juin 2001
Vol. 27, n° 1, pp. 35-48
Statistique Canada

Estimation composite par régression pour l'Enquête sur la population active du Canada avec plan de sondage à renouvellement de panel

AVINASH C. SINGH, BRIAN KENNEDY et SHIYING WU[1]

RÉSUMÉ

Nous considérons l'estimation composite par régression introduite par Singh (1994, 1996), qui a été appelée au début « estimation composite par régression modifiée » et dont une version (proposée par Fuller 1999) est appliquée à l'Enquête sur la population active du Canada (EPAC) depuis janvier 2000. L'estimateur par régression composite (rc) comporte plusieurs améliorations comparativement à l'estimateur par régression généralisée (gr) et l'estimateur composite *ak* bien connu de Gurney-Daly. Les caractéristiques principales de l'estimateur rc sont les suivantes : a) il augmente considérablement l'efficacité de l'estimation du niveau et de la variation pour les variables étudiées importantes, ce qui produit des séries d'estimations moins instables; b) comme l'estimateur gr, il est calculé de la même façon qu'un estimateur par calage, de sorte que soient satisfaites les valeurs de contrôle utilisées habituellement pour la stratification *a posteriori* dans gr, ainsi que les nouvelles valeurs de contrôle qui correspondent aux variables corrélées provenant de la période d'observation précédente; c) il maintient la cohérence interne des estimateurs sans qu'il soit nécessaire de calculer les estimations partielles par différence. Les innovations les plus importantes qui caractérisent la classe des estimateurs rc consiste à : a) utiliser des matrices de covariances de travail pour estimer les fonctions au lieu de recourir à la modélisation d'une superpopulation pour définir les coefficients de régression des prédicteurs inclus dans l'estimateur gr, b) traiter les contrôles aléatoires (ceux fondés sur les variables corrélées importantes des périodes antérieures) comme étant des constantes, tout en calculant les coefficients de régression de la même façon que pour l'estimation à deux phases et comme étant justifiés par le concept de la matrice de covariances de travail et c) se fonder sur le micro-appariement pour obtenir des données auxiliaires de même niveau que les estimations antérieures pour obtenir une plus forte corrélation avec les variables étudiées à la période courante. Secondairement, nous recommandons d'utiliser une nouvelle version de l'estimateur *ak* qui s'appuie sur le macro-appariement basé sur les prédicteurs des périodes antérieures plutôt que sur les microdonnées classiques afin d'augmenter les gains d'efficacité. L'article présente aussi une justification heuristique intéressante de la caractéristique de lissage des estimations composites fondées sur le concept de l'amortissement. Enfin, nous présentons les résultats empiriques de la comparaison de divers estimateurs basée sur les données de l'EPAC de 1996 pour l'Ontario.

MOTS CLÉS : Régression généralisée; régression modifiée; fonctions d'estimation; calage par régression.

1. INTRODUCTION

Dans le cas des enquêtes répétées avec chevauchement partiel des échantillons, il est bien connu (voir, par exemple Cochran 1977, chapître 12) que l'on peut améliorer les estimations ponctuelles du niveau pour une période et de la variation entre deux périodes par régression de l'estimateur transversal ordinaire (habituellement, estimateur de régression ou simplement estimateur d'Horvitz-Thompson) sur les nouveaux prédicteurs fournis par les observations corrélées faites sur le sous-échantillon chevauchant de la période précédente. Ces méthodes d'estimation rentrent dans la catégorie de l'estimation composite dont une version simple, appelée estimateur composite *k*, a été proposée il y a quelque temps par Hansen, Hurwitz et Madow (1953) et étudiée plus en détail par Rao et Graham (1964), Binder et Hidiroglou (1988), quant à eux, font une excellente revue des articles qui traitent de l'estimation dans le cas d'enquêtes répétées. Il convient de souligner que l'agrégation des estimations sur plusieurs périodes cause une perte d'efficacité , à cause de la plus forte corrélation positive entre les estimations composites ponctuelles successives. Néanmoins, il s'agit probablement d'un faible prix à payer, car ce n'est pas la précision de l'agrégat, mais celle des estimations du niveau et de la variation qu'il faut améliorer. L'estimateur composite *ak* de Gurney et Daly (1965) est une version améliorée de l'estimateur composite *k* obtenue grâce à une réduction supplémentaire de la variance et pour laquelle une autre justification plus simple a été fournie par Wolter (1979).

L'estimateur composite considéré ici a été développé dans le contexte de l'Enquête sur la population active du Canada (EPAC). Cette dernière est une enquête mensuelle fondée sur un plan de sondage avec renouvellement de panel et comptant six panels. Pendant deux mois consécutifs, cinq sixième des ménages sélectionnés forment l'échantillon chevauchant. C'est en janvier 2000 que l'on a commencé à appliquer aux données de l'EPAC une version (proposée par Fuller 1999) des estimateurs composites introduits par Singh (1994, 1996) et appelés au départ estimateurs « composites par régression modifiée », que nous dénommerons ici simplement estimateurs « composites par régression » ou estimateurs rc. Avant janvier 2000, on utilisait pour l'EPAC les estimateurs par régression

[1] Avinash C. Singh et Shiying Wu, Statistical Research Division, Research Triangle Institute, Research Triangle Park, N.C. 27709-2194, U.S.A.; Brian Kennedy, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6.

généralisée (gr) de Cassel, Särndal et Wretman (1976) et de Särndal (1980), qui se fondaient uniquement sur des données transversales (c'est-à-dire, du mois courant). On pensait depuis longtemps qu'il serait possible d'améliorer l'estimateur de l'EPAC en exploitant le concept de l'estimation composite, en ce sens que les estimations du niveau et de la variation seraient plus efficaces et, par conséquent, que les séries résultantes seraient plus stables, autrement dit, moins volatiles ou instables. Les quatre objectifs de la transformation de l'estimateur gr pour passer à l'estimateur rc sont les suivants :

i) le nouvel estimateur devrait augmenter considérablement l'efficacité de l'estimation du niveau et de la variation, de sorte que la série d'estimations devienne plus lisse et moins instable;

ii) on devrait pouvoir le calculer comme un estimateur par calage comparable à l'estimateur gr, de sorte qu'on puisse se servir des logiciels d'estimation existants en n'y apportant que peu de modifications;

iii) les poids calés finals devraient continuer de satisfaire les valeurs des variables démographiques et géographiques de contrôle utilisées habituellement dans l'estimateur gr et satisfaire aussi les nouveaux contrôles fondés sur les variables du mois précédent;

iv) l'estimateur devrait avoir la propriété de cohérence interne voulant que la somme des estimations composites partielles soit égale à l'ensemble; par exemple, la somme des estimations du nombre de personnes occupées (E), de chômeurs (U) et de personnes ne faisant pas partie de la population active (N) devrait être égale au total de la population admissible dans le domaine observé.

Kumar et Lee (1983) ont étudié l'estimateur *ak* dans le contexte de l'EPAC et constaté qu'il ne produisait pas les gains d'efficacité prévus par l'objectif (i). Ils savaient naturellement que les estimateurs *ak* ne satisfaisaient pas l'objectif (iv) car les coefficients « optimaux » a et k utilisés pour combiner plusieurs estimateurs du mois courant (en fait trois d'entre eux, à savoir l'estimateur ordinaire fondé sur le mois courant et deux autres basés sur des prédicteurs provenant du mois précédent) s'avèrent particuliers à la variable, comme E. Une solution (quoiqu'insatisfaisante) consiste à désigner l'une des composantes comme étant moins importante (disons N), puis de l'estimer par différence. Par contre, les objectifs (ii) et (iii) peuvent être atteints par la pondération composite ak proposée par Fuller (1990) et étudiée dans le contexte de la *Current Population Survey* des États-Unis par Lent, Miller, et Cantwell (1994, 1996). L'objectif (ii) est important, surtout pour des variables étudiées non prévues pour lesquelles les coefficients (a,k) ne sont pas connus d'avance. L'estimateur rc permet d'atteindre les quatre objectifs, en particulier les objectifs (i) et (iv), grâce aux trois innovations qui suivent.

i) On peut intégrer l'estimation basée sur le plan de sondage en présence de variables prédictives corrélées dans un cadre de fonctions d'estimation tel que celui défini par Godambe et Thompson (1989), puis exploiter l'idée d'une matrice de covariances de travail, comme celle avancée par Liang et Zeger (1986), pour obtenir une autre solution que la modélisation de la superpopulation pour calculer les coefficients de régression. Comme dans le cas de la régression généralisée gr, les estimations par régression résultantes ne sont sous-optimales que dans le cas de la randomisation du plan de sondage.

ii) On peut traiter les estimations composites calculées d'après l'échantillon complet du mois précédent et utilisées comme contrôles de la régression pour obtenir l'estimation du mois courant comme des constantes en se servant du concept de la covariance de travail pour simplifier les calculs tout en respectant la propriété de cohérence du plan de sondage. Pour estimer la variance, il faut évidemment pouvoir tenir compte de la variation supplémentaire due aux contrôles aléatoires.

iii) Par micro-appariement du sous-échantillon chevauchant du mois courant avec l'échantillon du mois précédent, à partir des données provenant du mois précédent on augmente l'information sur les variables étudiées importantes le mois courant. Ces variables servent alors de covariables supplémentaires qui devraient être fortement corrélées à la variable étudiée du mois courant.

Ces innovations permettent de calculer toutes les estimations en se servant du système de régression généralisée gr, donc de ne pas devoir calculer certains éléments des estimations par différence afin d'assurer la cohérence interne. Le micro-appariement permet de réaliser les gains d'efficacité souhaités. En pratique, il arrive souvent que certains répondants du mois courant compris dans l'échantillon chevauchant aient été des non-répondants le mois précédent et qu'une imputation soit donc nécessaire. Dans le cas de l'EPAC, ces cas ne représentent qu'une faible fraction de l'échantillon et l'on recourt pour remplacer les valeurs manquantes à la méthode hot deck avec des catégories de donneurs définies selon les caractéristiques démographiques et géographiques (régions économiques infraprovinciales), le type de région (rurale/urbaine), la situation d'emploi durant le mois courant et le groupe de branches d'activité. Il convient de souligner que, parfois, l'imputation est nécessaire non pas parce qu'il n'y a pas eu de réponse à la période précédente, mais parce que le ménage a déménagé. Si l'on suppose que le nombre moyen de ménages qui déménagent pour s'installer dans des logements échantillonnés à la période courante t est comparable

au nombre moyen de ménages qui déménagent hors de ces logements à la période t, alors, même si les personnes qui ont déménagé ont d'autres caractéristiques d'emploi que celles qui n'ont pas déménagé, l'imputation de données pour les personnes qui ont déménagé ne devrait, en principe, introduire aucun nouveau biais, puisque l'on tient compte de la situation d'emploi durant le mois courant pour les autres covariables.

Dans nos conclusions, à la section 6, nous proposons une méthode pour diagnostiquer l'effet de cette imputation. Cet effet peut être important dans le cas d'enquêtes où la fraction de valeurs qui manquaient à la période précédente est importante pour les répondants du mois courant qui font partie de l'échantillon chevauchant. Un moyen simple de résoudre ce problème consisterait à remanier le questionnaire de sorte que l'application informatique d'ITAO (interview téléphonique assistée par ordinateur, utilisée couramment de nos jours) indique à l'intervieweur, lorsqu'il procède à l'interview le deuxième mois ou les mois suivants, si le répondant avait répondu ou non le mois précédent. Dans la négative, l'intervieweur poserait alors quelques questions supplémentaires afin de déterminer la situation d'emploi du répondant le mois précédent. Cette solution est comparable à la méthode proposée par Hansen-Hurwitz-Madow pour des enquêtes répétées avec échantillons entièrement non chevauchants, où des questions sont posées à chaque répondant concernant la période courante et la période précédente (voir Cochran 1977, page 355).

Le plan de l'article est le suivant. À la section 2, nous présentons une justification heuristique de l'utilisation du concept d'amortissement pour expliquer pourquoi on s'attend en général à ce que l'estimation composite produise le lissage souhaité de la série d'estimations. À la section 3, nous définissons divers estimateurs et décrivons leur calcul au moyen du système de répression généralisée gr. Nous proposons aussi une nouvelle version de l'estimateur *ak*, représentée par *ak**. Cet estimateur comprend des prédicteurs provenant du mois précédent obtenus par micro-appariement et devrait, en principe, permettre de réaliser des gains d'efficacité importants. À la section 4, nous considérons l'estimation de la variance selon la méthode du jackknife utilisée à l'heure actuelle. À la section 5, nous présentons une comparaison empirique des estimateurs fondée sur des données de l'EPAC de 1996 pour l'Ontario. Enfin, nous présentons nos conclusions à la section 6.

2. LISSAGE DES SÉRIES PAR ESTIMATION COMPOSITE : JUSTIFICATION HEURISTIQUE

Nous présentons ici une justification heuristique intéressante (fondée sur le concept d'amortissement plutôt que sur le concept de réduction) du fait que l'on s'attende à ce que l'estimation composite produise un lissage de la série d'estimations. (Si l'on s'appuie uniquement sur le concept de réduction, la série sera lissée, mais ne recoupera pas assez souvent la série originale. Par contre, en appliquant le concept d'amortissement, on rend compte graduellement au cours du temps de la partie restante après la réduction, ce qui permet le recoupement plus fréquent de la série lissée et de la série originale.) Considérons un plan de sondage avec renouvellement de panel comparable à celui de l'EPAC et représentons par γ la fraction des panels retirée de l'échantillon; dans le cas de l'EPAC, γ est égal à 1/6. Représentons l'estimateur transversal (classiquement, gr) à la période t fondé sur tous les panels (c'est-à-dire l'échantillon complet) par F_t, l'estimateur fondé sur le nouveau panel uniquement (c'est-à-dire le panel ajouté à l'échantillon) par B_t, et celui fondé sur les panels déjà existants (c'est-à-dire le sous-échantillon qui, à la période t chevauche l'échantillon de la période précédente $t-1$) $\bar{B}_t$. Pareillement, représentons l'estimateur fondé uniquement sur le panel supprimé (c'est-à-dire retiré de l'échantillon) par D_t, et celui fondé sur les panels non supprimés (c'est-à-dire le sous-échantillon qui, à la période $t-1$, chevauche l'échantillon de la période courante t) par $\bar{D}_t$. Nous obtenons

$$F_t = \gamma\, B_t + (1-\gamma)\, \bar{B}_t \tag{2.1a}$$

$$F_{t-1} = \gamma\, D_{t-1} + (1-\gamma)\, \bar{D}_{t-1}. \tag{2.1b}$$

Supposons que la série $\{F_t\}$ est trop instable et que nous voulons la lisser. Dans la suite, nous supposons qu'aucun biais n'est introduit par le groupe de renouvellement (Bailar 1975), c'est-à-dire que la valeur prévue est la même pour divers groupes de renouvellement. Donc, F_t est non biaisé, mais peut être instable. Il s'agit-là des conditions classiques de l'estimation composite pour laquelle diverses estimations non biaisés sont combinées de façon optimale pour obtenir une estimation plus efficace. Pour une autre perspective sur l'estimation composite en cas de biais dû au groupe de renouvellement, consulter la discussion à la fin de la présente section. Maintenant, représentons la série lissée par $\{C_t\}$, et considérons l'identité :

$$F_t = C_{t-1} + (F_t - F_{t-1}) + (F_{t-1} - C_{t-1}). \tag{2.2}$$

Nous pouvons interpréter la relation qui précède comme suit. L'estimation C_{t-1} à la période $t-1$ est corrigée en tenant compte de la fluctuation $(F_t - F_{t-1})$ à la période t dans la série F et de l'écart existant $(F_{t-1} - C_{t-1})$ à la période $t-1$. Si nous définissons C_t après les rajustements complets pour ces deux différences, C_t sera identique à F_t et il n'y aura aucun lissage de la série F. Ces résultats laissent entendre qu'on ne devrait expliquer que partiellement les rajustements pour les différences $(F_t - F_{t-1})$ et $(F_{t-1} - C_{t-1})$ lorsque la série C passe de C_{t-1} à C_t. Les parties restantes des différences devraient être amorties graduellement sur les périodes futures. Tous ces rajustements devraient être effectués en s'assurant que

l'estimateur C_t demeure non biaisé. L'espérance de la différence $(F_{t-1} - C_{t-1})$ est nulle si l'on suppose que C_{t-1} et F_{t-1} ne sont pas biaisés (ce qui est le cas si l'on suppose qu'il n'existe aucun biais dû au groupe de renouvellement) et, par conséquent, leur amortissement partiel n'aura pas d'influence sur l'absence de biais des futures estimations C_t. Cependant, en réalité, l'espérance de la différence $F_t - F_{t-1}$ n'est pas nulle et il convient d'être prudent lors de l'amortissement d'une de ses parties. Notons que

$$F_t - F_{t-1} = (\bar{B}_t - \bar{D}_{t-1}) + \gamma(B_t - \bar{B}_t) + \gamma(\bar{D}_{t-1} - D_{t-1}). \quad (2.3)$$

Le premier terme du deuxième membre de l'équation est l'estimation de la variation fondée sur les panels communs, tandis que les deuxième et troisième représentent l'effet du panel ajouté et du panel retiré aux périodes t et $t-1$, respectivement. Les deux derniers termes sont des fonctions nulles (c'est-à-dire que leur espérance est nulle), mais le premier ne l'est pas (heureusement, ce premier terme devrait être stable, car il s'agit de la différence entre deux estimations fortement corrélées). Par conséquent, ce sont les deuxième et troisième termes qui doivent être amortis. Maintenant, écrivons l'équation (2.2) sous la forme

$$\begin{aligned} F_t &= C_{t-1} + (\bar{B}_t - \bar{D}_{t-1}) + \gamma(B_t - \bar{B}_t) \\ &\quad + [\gamma(\bar{D}_{t-1} - D_{t-1}) + (F_{t-1} - C_{t-1})] \\ &= C_{t-1} + (\bar{B}_t - \bar{D}_{t-1}) + \gamma(B_t - \bar{B}_t) \\ &\quad + [(\bar{D}_{t-1} - F_{t-1}) + (F_{t-1} - C_{t-1})] \\ &= C_{t-1} + (\bar{B}_t - \bar{D}_{t-1}) + \gamma(B_t - \bar{B}_t) + (\bar{D}_{t-1} - C_{t-1}). \quad (2.4) \end{aligned}$$

et définissons deux facteurs d'amortissement δ_{1t}, δ_{2t} dont la valeur est comprise entre 0 et 1, puis définissons la série lissée $\{C_t\}$ comme étant

$$C_t = C_{t-1} + (\bar{B}_t - \bar{D}_{t-1}) + \delta_{1t}\gamma(B_t - \bar{B}_t) + \delta_{2t}(\bar{D}_{t-1} - C_{t-1}). \quad (2.5)$$

Le terme qui contient δ_{1t} dans l'équation (2.5) représente la réduction de l'effet du nouveau panel à la période t dont C_t essaye de rendre compte, tandis que le terme qui contient δ_{2t} se rapporte approximativement à la réduction de l'effet du panel de la période précédente $(t-1)$ que C_t essaye de compenser à la période t. En outre, il serait souhaitable de poser $\delta_{2t} < \delta_{1t}$ pour que la série $\{C_t\}$ suive mieux la série $\{F_t\}$, de sorte qu'elles soient caractérisées toutes deux par une même tendance au cours du temps, ce qui revient à accorder plus d'importance à l'effet du nouveau panel à la période courante qu'à celui du panel de la période précédente qui a été supprimé. (En fait, une justification rigoureuse, dans des conditions assez générales, de la contrainte $\delta_{2t} < \delta_{1t}$ provient de considérations d'optimalité en vertu desquelles la variance de C_t est réduite au minimum pour obtenir la meilleure combinaison linéaire possible des trois estimateurs non biaisés, F_t, $C_{t-1} + \bar{B}_t - \bar{D}_{t-1}$ et $F_t + C_{t-1} - \bar{D}_{t-1}$ du total de population du mois courant; voir l'équation (2.8) à la fin de la présente section pour l'expression réelle.) Maintenant, pour montrer le lien avec l'estimation composite bien connue définie à la section suivante, posons $0 < a_t, b_t < 1$, de sorte que $\delta_{1t} = 1 - b_t$, $\delta_{2t} = 1 - b_t - a_t$. Nous avons

$$\begin{aligned} C_t &= C_{t-1} + (\bar{B}_t - \bar{D}_{t-1}) + (1 - b_t)\gamma(B_t - \bar{B}_t) \\ &\quad + (1 - b_t - a_t)(\bar{D}_{t-1} - C_{t-1}). \quad (2.6) \end{aligned}$$

Il est intéressant de noter que si $b_t = 0$, il n'y aura aucun amenuisement de l'effet dû au nouveau panel et la série C se rapprochera, en principe, de la série F, c'est-à-dire que le lissage sera moins important et que les deux séries se recouperont plus souvent. Si $a_t = 0$, l'effet dû au panel de la période précédente qui a été supprimé, représenté par $(\bar{D}_{t-1} - C_{t-1})$, est moins amorti. Autrement dit, la série F sera lissée plus fortement et, en principe, les deux séries se recouperont moins fréquemment. Enfin, si $a_t, b_t > 0$, le comportement de la série C par rapport à la série F sera compris entre les deux comportements susmentionnés. Qui plus est, si la valeur de b_t est grande (proche de 1), le lissage de la série F sera assez prononcé, car les effets du nouveau panel et du panel supprimé seront tous deux fortement amortis. Dans ces situations, on s'attendrait à observer des écarts prolongés entre les séries F et C au cours du temps avant qu'elles ne se recoupent. On notera que les parties du terme $\gamma(B_t - \bar{B}_t)$ qui sont amorties sur les périodes $t, t+1, \ldots$ diminuent à mesure que t augmente. Elles sont données par $b_t\gamma(B_t - \bar{B}_t), (b_{t+1} + a_{t+1})$ $b_t\gamma(B_t - \bar{B}_t), \ldots$. Pareillement, les parties amorties de $(\bar{D}_{t-1} - C_{t-1})$ sont

$$(b_t + a_t)(\bar{D}_{t-1} - C_{t-1}), (b_{t+1} + a_{t+1})(b_t + a_t)(\bar{D}_{t-1} - C_{t-1}), \ldots .$$

De toute évidence, si b_t est grand, l'amortissement complet prend plusieurs périodes. Cependant, comme on l'a expliqué plus haut, cette situation n'introduira pas de biais, car les effets amortis sont des fonctions nulles si l'on suppose qu'il n'y a pas de biais dû au groupe de renouvellement.

L'expression (2.6) peut être écrite sous une forme plus connue de l'estimateur composite, comme suit :

$$\begin{aligned} C_t &= C_{t-1} + (\bar{B}_t - \bar{D}_{t-1}) + (1 - b_t)(F_t - \bar{B}_t) \\ &\quad + (1 - b_t)(\bar{D}_{t-1} - C_{t-1}) + a_t(C_{t-1} - \bar{D}_{t-1}) \quad (2.7a) \\ &= C_{t-1} + (\bar{B}_t - \bar{D}_{t-1}) + (1 - b_t)(F_t - \bar{B}_t + \bar{D}_{t-1} - C_{t-1}) \\ &\quad + a_t(C_{t-1} - \bar{D}_{t-1}) \quad (2.7b) \\ &= F_t + b_t[C_{t-1} - (F_t + \bar{D}_{t-1} - \bar{B}_t)] + a_t(C_{t-1} - \bar{D}_{t-1}) \quad (2.7c) \\ &= F_t + (b_t + a_t)(C_{t-1} - \bar{D}_{t-1} + \bar{B}_t - F_t) + a_t(F_t - \bar{B}_t). \quad (2.7d) \end{aligned}$$

L'expression (2.7d) correspond à l'estimateur ak (voir la section suivante) si $a_t = a$ et $b_t + a_t = k$. En pratique, on

peut déterminer les valeurs de a_t et b_t de façon optimale ou sous-optimale par régression (voir la section suivante). En général, les coefficients de régression partielle a_t, b_t satisfont $0<a_t<b_t<1$, car l'estimateur direct F_t devrait, en principe, être plus positivement corrélé au prédicteur $F_t + (\bar{D}_{t-1} - \bar{B}_t)$, c'est-à-dire $\bar{D}_{t-1} + \gamma(B_t - \bar{B}_t)$ qu'au prédicteur $\bar{D}_{t-1}$; les deux prédicteurs sont des estimations non biaisées du total de population C_{t-1} à la période précédente $t-1$. Il découle de (2.7c) que l'estimateur C_t peut être écrit sous forme de combinaison linéaire des trois estimateurs non biaisés mentionnés plus haut et donnés par

$$C_t = (1-b_t-a_t)F_t + b_t(C_{t-1} + \bar{B}_t - \bar{D}_{t-1}) + a_t(F_t + C_{t-1} - \bar{D}_{t-1}). \quad (2.8)$$

La justification heuristique qui précède s'appuie sur des considérations de réduction de la variance si l'on suppose qu'il n'y a aucun biais dû au groupe de renouvellement lorsque l'on combine trois estimateurs non biaisés du total de population à la période t. Par contre, l'existence d'un biais dû au groupe de renouvellement introduit dans les trois estimateurs un biais dont la grandeur et le signe peuvent différer selon l'estimateur et, dans ce cas, l'estimation composite rajuste chacun des estimateurs de sorte que la valeur corrigée pour chacun soit égale à une valeur commune donnée par l'estimateur composite (par exemple, dans le cas de deux estimateurs $\hat{\theta}_1$ et $\hat{\theta}_2$ de θ, la combinaison linéaire $\lambda\,\hat{\theta}_1 + (1-\lambda)\,\hat{\theta}_2$ peut s'écrire sous la forme $\hat{\theta}_2 + \lambda(\hat{\theta}_1 - \hat{\theta}_2)$ ou $\hat{\theta}_1 + (1-\lambda)(\hat{\theta}_2 - \hat{\theta}_1)$, ce qui sous-entend que les deux estimateurs originaux sont corrigés de façon appropriée pour qu'ils convergent vers une valeur commune). Le poids relatif appliqué lorsque l'on combine les trois estimateurs dépend du critère de variance minimale. Idéalement, il devrait se fonder sur le critère de l'EQM minimale, mais il est difficile de maîtriser le biais, car on ne peut l'estimer. L'estimateur composite n'est manifestement pas dépourvu de biais et l'on peut simplement spéculer que le biais global de l'estimateur est réduit par le recours à un estimateur composite. Selon le même raisonnement si, inversement, on utilise une régression sous-optimale pour créer l'estimateur composite (comme dans l'estimation rc, voir la section suivante), l'effet de l'estimation composite est de rajuster les poids d'échantillonnage dans l'échantillon complet (qui sont généralement des poids gr) de sorte que $F_t - (\bar{B}_t - \bar{D}_{t-1})$ et $\bar{D}_{t-1}$ avec les poids rajustés, devient égal à C_{t-1}; les valeurs de C_{t-1} servent de nouvelles valeurs de contrôle à l'étape du calage. Il s'agit là d'un autre moyen de rajuster les trois estimateurs sur une valeur commune, mais, de nouveau, le biais introduit par l'estimateur composite est inconnu. L'exposé des deux perspectives sous lesquelles on peut examiner l'estimation composite présente des points communs avec l'effet double de la stratification *a posteriori* qui réduit à la fois la variance et le biais (de couverture); à ce sujet, consulter Singh et Folsom (2000).

3. ESTIMATEURS COMPOSITES : NOUVEAUX ET ANCIENS

Nous commençons par l'estimateur transversal du total $\tau_y(t)$ à la période t défini comme étant l'estimateur par régression généralisée gr, qui est donné par

$$\hat{\tau}_{y(gr)}(t) = \sum_{k\in s(t)} y_k(t)\, w_{gr}(t,k), \quad (3.1)$$

$$w_{gr}(t,k) = d(t,k)\left[1 + x_k(t)'\,(X(t)'\,\Delta(t)X(t))^{-1}\,(\tau_x(t) - \hat{\tau}_x(t))\right], \quad (3.2)$$

où les $d(t,k)$ sont les poids de sondage initiaux corrigés pour la non-réponse, $x_k(t)$ est un vecteur p de covariables utilisées pour le calage (ou la stratification a posteriori), $X(t)$ est la matrice $n(t)\,x\,p$ des observations de x, $n(t)$ est la taille de l'échantillon, $\Delta(t)$ est $\mathrm{diag}(d(t,k))$, $\tau_x(t)$ est le vecteur p connu des contrôles utilisés pour le calage et $\hat{\tau}_x(t)$ est le vecteur correspondant des estimations avec facteur d'extension fondées sur les poids d. Si l'on reprend la notation F_t, $\bar{B}_t$, et B_t de la section précédente, F_t peut être considéré comme l'estimateur gr (3.1) et $\bar{B}_t$, comme l'estimateur gr fondé sur les panels déjà existants donnés par

$$\bar{B}_t = (1-\gamma)^{-1} \sum_{k\in s(t|t-1)} y_k(t)\, w_{gr}(t,k), \quad (3.3)$$

où $s(t|t-1)$ est le sous-échantillon de la période t apparié à l'échantillon de la période $t-1$. L'estimateur B_t est également un estimateur gr qui peut être représenté par

$$B_t = \gamma^{-1} \sum_{k\in s(t)-s(t|t-1)} y_k(t)\, w_{gr}(t,k), \quad (3.4)$$

où la somme est faite sur le sous-échantillon définit par le nouveau panel ajouté à la période t.

L'estimateur composite ak, qui utilise des données de niveau macro de la période précédente pour les nouveaux prédicteurs, peut s'écrire

$$C_{t(ak)} = F_t + k(C_{t-1(ak)} - \bar{D}_{t-1} + \bar{B}_t - F_t) + a(F_t - \bar{B}_t)$$

$$= F_t + (k-a)(C_{t-1(ak)} - \bar{D}_{t-1} + \bar{B}_t - F_t) + a(C_{t-1(ak)} - \bar{D}_{t-1}). \quad (3.5)$$

Ici, on obtient les coefficients a, k pour l'estimation du niveau par régression optimale de F_t sur les deux fonctions prédictives nulles, basées sur les données antérieures, à savoir $C_{t-1(ak)} - (F_t + \bar{D}_{t-1} - \bar{B}_t)$ et $C_{t-1(ak)} - \bar{D}_{t-1}$. Donc, a et k dépendent du plan d'échantillonnage, ainsi que de la variable étudiée y; plus précisément, ils ne sont même pas identiques pour les estimations du niveau et de la variation pour une même variable y. Pour l'estimation de la variation, c'est la régression de $F_t - C_{t-1(ak)}$ et non de F_t que l'on calcule de façon optimale sur les prédicteurs. En pratique, on estime a, k en explorant par compartiment l'intervalle (0,1) afin de trouver les valeurs pour lesquelles la variance

de C_t est réduite au minimum. Comme nous l'avons mentionné plus haut, habituellement a est plus petit que k. Pour définir les deux nouvelles fonctions prédictives nulles susmentionnées en se fondant sur les données antérieures, on commence par former deux estimateurs de $\tau_y(t-1)$: l'un est $\bar{D}_{t-1}$, fondé sur les panels de la période $t-1$ conservés (c'est-à-dire le sous-échantillon de la période $t-1$ apparié à l'échantillon de la période t) et l'autre est $F_t+(\bar{D}_{t-1}-\bar{B}_t)$ qui est l'estimateur gr à la période t, corrigé pour la variation de $t-1$ à t estimée d'après l'échantillon commun. De toute évidence, si le plan de sondage ne prévoit aucun chevauchement de panel, les fonctions prédictives nulles perdent leur signification et l'estimation composite ne modifie pas F_t. Pareillement, si le chevauchement est complet, $\bar{B}_t = F_t$ et, de nouveau, l'estimation composite n'a aucun effet sur F_t. À première vue, cela paraît contraire à l'intuition, puisque les données de la période précédente (y_{t-1}) sont corrélées à celles de la période courante (y_t) à cause du chevauchement de l'échantillon. Cependant, le chevauchement complet se résume, en principe, à recueillir sur y, auprès d'un échantillon unique, des données multidimensionnelles dont les éléments correspondent à y à diverses périodes. Selon cette analogie, il n'y a aucune possibilité d'amélioration (dans le cadre de référence axée sur le plan de sondage), puisqu'il n'existe aucun échantillon plus grand pouvant fournir des données supplémentaires. Dans le cas où il n'y a aucun chevauchement, l'information supplémentaire existe, mais elle n'est pas utile, car elle n'est pas corrélée. On notera, cependant, qu'au premier degré, les UPE (unités primaires d'échantillonnage) de l'EPAC restent communes pendant plusieurs années avant d'être supprimées de l'échantillon. Par conséquent, les gains d'efficacité dus au chevauchement partiel sont réalisés principalement grâce à la réduction de la composante de la variance liée à l'échantillonnage de deuxième degré.

De surcroît, soulignons que l'estimateur $C_{t(ak)}$ se fonde sur des données antérieures de façon unidimensionnelle, puisque, pour la variable étudiée y, on se sert uniquement de l'information y_{t-1} provenant de la période précédente. Si l'on se sert pour la variable étudiée y de prédicteurs fondés sur plusieurs variables des périodes précédentes, comme $y_{t-1}, z_{t-1}, \ldots$, alors l'estimation composite devient multidimensionnelle. Cependant, le choix optimal des coefficients (a, k) pour le cas multidimensionnel peut être assez fastidieux.

La catégorie d'estimateurs par régression composite (rc) est donnée par

$$C_{t(rc)} = F_t + b_{t(rc)}\left(\tilde{C}_{t-1(rc)} - \bar{D}^*_{t-1} + \bar{B}_t - F_t\right)$$

$$+ a_{t(rc)}\left(\tilde{C}_{t-1(rc)} - \bar{D}^*_{t-1}\right) \qquad (3.6)$$

où $\tilde{C}_{t-1(rc)}$ représente l'estimateur $t-1$ pour la variable étudiée y après que les poids calés à la période $t-1$ soient de nouveau calés sur les contrôles utilisés pour la stratification *a posteriori* dans le cas de la régression généralisée (gr) à la période t. Donc, $\tilde{C}_{t-1(rc)}$ est une estimation du total de la population à la période t pour la variable y à la période $t-1$. Dans $\bar{D}^*_{t-1}$, l'astérisque signifie que ce terme est fondé sur le sous-échantillon de la période t apparié à l'échantillon de la période $t-1$, mais que l'on a utilisé les poids gr à la période t, puisque les valeurs de y provenant de la période $t-1$ sont augmentées par micro-appariement de façon à ce qu'elles correspondent aux valeurs de l'échantillon à la période t. (À noter que l'estimateur $\bar{D}^*_{t-1}$ contient, en général, des valeurs imputées et qu'il peut être entaché d'un biais dû à l'imputation. Pour le diagnostic et le rajustement de ce biais, voir la section 6.) Les coefficients $b_{t(rc)}$ et $a_{t(rc)}$ sont calculés de la même façon que pour l'estimateur gr donné par l'équation (3.1); des précisions supplémentaires sont données plus loin. Ces coefficients sont sous- optimaux, contrairement à (a, k). Cependant, comme (a, k), ils sont particuliers à y, et, dans le cas de l'estimation multidimensionnelle, dépendent de l'ensemble clé de variables étudiées provenant de la période précédente pour servir de nouveaux contrôles, mais ils peuvent être calculés facilement puisqu'ils sont sous-optimaux. Donc, dans le cas de l'estimation rc, il est assez facile d'introduire un plus grand nombre de prédicteurs. Les prédicteurs $(C_{t-1} - \bar{D}_{t-1})$ et $(C_{t-1} - \bar{D}_{t-1} + \bar{B}_t - F_t)$ peuvent être qualifiés, respectivement, de prédicteur axé sur le niveau et de prédicteur axé sur la variation comme dans Singh, Kennedy, Wu et Brisebois (1997). Il en est ainsi non seulement parce que le premier est la différence de deux estimations de niveau et le second, la différence de deux estimations de la variation, $(C_{t-1} - F_t)$ et $(\bar{D}_{t-1} - \bar{B}_t)$, mais aussi parce que le premier a tendance à augmenter fortement l'efficacité de l'estimation du niveau comparativement au gain que l'on peut obtenir en présence du second et, pareillement, que le second augmente considérablement l'efficacité de l'estimation de la variation comparativement à celle que l'on peut obtenir en présence du premier.

Nous pouvons appliquer l'utilisation des microdonnées de la période précédente pour obtenir les nouveaux prédicteurs pour l'estimation rc à l'estimateur ak et proposer ainsi un nouvel estimateur ak^* donné par

$$C_{t(ak^*)} = F_t + (k^* - a^*)\left(\tilde{C}_{t-1(ak^*)} - \bar{D}^*_{t-1} + \bar{B}_t - F_t\right)$$

$$+ a^*\left(\tilde{C}_{t-1(ak^*)} - \bar{D}^*_{t-1}\right). \qquad (3.7)$$

La variable de contrôle $\tilde{C}_{t-1(ak^*)}$ représente l'estimateur par calage à la période $(t-1)$ pour y après que les poids composites ak^* soient de nouveau calés sur les variables de contrôle utilisées pour la stratification *a posteriori* par régression généralisée à la période t. (Ici les poids

composites ak^* sont les mêmes que les poids composites ak de Fuller (1990) où les estimateurs composites pour un ensemble de variables y importantes servent de contrôles supplémentaires lors de la régression généralisée ordinaire pour obtenir un ensemble de poids calés finals. Ceci permet de calculer l'estimateur composite ak comme s'il s'agissait d'un estimateur par calage.) Les différences principales entre les divers estimateurs décrits plus haut tiennent à la définition des coefficients de régression (optimaux c. sous-optimaux) et à celle des prédicteurs (utilisation de données antérieures de niveau micro c. macro). On peut obtenir les cas spéciaux des estimateurs composites susmentionnés décrits dans Singh, et coll. (1997) en n'utilisant que l'un des deux prédicteurs. Pour $C_{t(ak)}$, si $a = 0$ (c'est-à-dire si l'on se sert uniquement d'un prédicteur axé sur la variation), nous obtenons l'estimateur composite k bien connu que l'on peut qualifier d'estimateur $ak2$ dans le présent contexte. Si $a = k$, c'est-à-dire si l'on utilise uniquement un prédicteur axé sur le niveau, nous obtenons un nouvel estimateur composite $C_{t(ak1)}$ que l'on peut appeler estimateur $ak1$. De la même façon, pour $C_{t(ak*)}$, nous obtenons deux nouveaux estimateurs composites supplémentaires, ak^*1 et ak^*2. Pour $C_{t(rc)}$, si nous utilisons uniquement un prédicteur axé sur le niveau, nous obtenons l'estimateur rc1, appelé précédemment MR1 dans Singh et Merkouris (1995). Si l'on se sert uniquement de prédicteurs axés sur la variation, nous obtenons l'estimateur rc2 appelé antérieurement MR2 dans Singh, et coll. (1997).

Comme nous l'avons mentionné plus haut, l'estimateur rc est calculé comme un estimateur gr de (3.1) et, par conséquent, peut être exprimé sous la forme $\hat{\tau}_{y(rc)}(t) = \sum_{k \in s(t)} y_k(t)\, w_{rc}(t,k)$. Nous étendons la matrice $X(t)$ à la matrice $n(t) \times (p+2q)$, soit $X(t)^*$, où $2q$ représente le nombre de nouveaux prédicteurs, le facteur 2 signifiant que l'on se sert de deux prédicteurs, l'un axé sur le niveau et l'autre sur la variation. Les totaux de contrôle (aléatoires) $\tilde{C}_{t-1(rc)}$ qui correspondent à l'ensemble clé de variables y de la période $t-1$ sélectionnés pour l'estimation composite sont traités comme des constantes (durant le calcul du coefficient de régression) au même titre que les autres contrôles gr (non aléatoires). Maintenant, puisque le prédicteur axé sur le niveau peut être écrit sous la forme

$$\begin{aligned}\bar{D}^*_{t-1} &= (1-\gamma)^{-1} \textstyle\sum_{k \in s(t \cap t-1)} y_k(t-1)\, w_{gr}(t,k) \\ &= \textstyle\sum_{k \in s(t)} (1-\gamma)^{-1} y_k(t-1)\, 1_{k \in s(t \cap t-1)}\, w_{gr}(t,k) \quad (3.8)\end{aligned}$$

la colonne de la matrice $X(t)^*$ qui correspond à ce prédicteur comprend $n(t)$ valeurs de $(1-\gamma)^{-1}$ $y_k(t-1)\, 1_{k \in s(t \cap t-1)}$. De la même façon, le prédicteur axé sur la variation peut s'écrire sous la forme

$$F_t + \bar{D}^*_{t-1} - \bar{B}_t =$$

$$\textstyle\sum_{k \in s(t)} \left(y_k(t) + (1-\gamma)^{-1}(y_k(t-1) - y_k(t))1_{k \in s(t \cap t-1)}\right) w_{gr}(t,k) \quad (3.9)$$

et la colonne correspondante de la matrice $X(t)^*$ comprend les $n(t)$ valeurs de $y_k(t) + (1-\gamma)^{-1}(y_k(t-1) - y_k(t))\, 1_{k \in s(t \cap t-1)}$. Lorsque la matrice $X(t)^*$ est définie, on peut se servir du système de régression généralisée pour calculer les poids $w_{rc}(t,k)$ par calage comme dans (3.2). On notera que l'on peut utiliser les poids calés $w_{rc}(t,k)$ pour estimer toutes les variables étudiées, mais qu'ils ne dépendent explicitement que de l'ensemble clé de variables étudiées choisies comme nouveaux prédicteurs d'après les données antérieures corrélées. Notons aussi que, même si on a défini l'estimateur rc donné par l'équation (3.6) comme étant l'estimateur gr corrigé par régression pour tenir compte des nouveaux prédicteurs, pour faciliter le calcul, il est commode de procéder à un calage par régression généralisée sur les poids de sondage lorsque l'on considère simultanément les anciens et les nouveaux contrôles de calage. De cette façon, le calcul de l'estimateur rc multidimensionnel diffère peu de celui de l'estimateur rc unidimensionnel. Autrement, on pourrait calculer l'estimateur rc sous forme d'estimateur corrigé par régression généralisée comme dans (3.6), mais les coefficients des nouveaux prédicteurs seraient des coefficients de régression partielle et, par conséquent, n'auraient pas la forme type des coefficients de régression généralisée.

Enfin, nous constatons que, dans le cas de l'estimation composite, on s'attendrait à observer des gains d'efficacité plus importants pour les estimations de la variation $(C_t - C_{t-1}\, c.\, F_t - F_{t-1})$ que pour les estimations du niveau $(C_t\ c.\ F_t)$. Pour montrer ceci, considérons une identité simple : $V(F_t - F_{t-1}) = V(F_t) + V(F_{t-1}) - 2\mathrm{Cov}(F_t, F_{t-1})$. Typiquement, $V(F_t) \approx V(F_{t-1}) = \sigma^2_{gr}$ (disons), alors l'équation qui précède peut être réduite à $V(F_t - F_{t-1}) \approx 2\,\sigma^2_{gr}(1-\rho_{gr})$. Pareillement, $V(C_t - C_{t-1}) \approx 2\,\sigma^2_{rc}(1-\rho_{rc})$. Donc, l'efficacité de l'estimation de la variation est approximativement égale à l'efficacité de l'estimation du niveau multipliée par $(1-\rho_{gr})/(1-\rho_{rc})$. Il s'ensuit que si les nouveaux prédicteurs pour l'estimation composite augmentent considérablement la corrélation (positive) entre C_t et C_{t-1}, alors l'efficacité de l'estimation de la variation sera nettement supérieure à celle de l'estimation du niveau.

4. ESTIMATION DE LA VARIANCE

À l'heure actuelle, dans le cas de l'EPAC, on applique la méthode du jackknife avec élimination d'UPE pour calculer la variance de l'estimation par régression généralisée. La méthode du jackknife est valide (pour les enquêtes transversales) si les estimations au niveau de l'UPE ont les mêmes moyennes et variances et que l'on peut traiter la sélection des UPE comme un tirage avec remise. Si la sélection des UPE se fait sans remise, l'estimation de la variance devient prudente si, comme cela est généralement le cas, la covariance (commune) entre les estimations au

niveau de l'UPE est négative. Dans le cas des enquêtes répétées, il est nécessaire d'imposer une troisième contrainte, à savoir que les UPE soient communes (ou liées au cours du temps). Le cas échéant, l'enquête peut être considérée comme transversale si l'on traite le vecteur d'observations (estimations au niveau de l'UPE) faites au cours du temps comme une observation unique recueillie au point final conceptuel dans le temps. Dans le cas du plan de sondage avec renouvellement de panel de l'EPAC, les UPE ne sont pas supprimées de l'échantillon pendant plusieurs années, mais les unités à l'intérieur des UPE sont remplacées tous les six mois. Dans le cas de l'EPAC, chaque UPE correspond à un panel distinct qui est soit un nouveau panel soit un ancien panel. À noter que pour satisfaire les conditions de la méthode du jackknife, il n'est pas nécessaire d'utiliser le même ensemble d'unités pour obtenir les estimations au niveau de l'UPE. La contrainte voulant que les estimations au niveau de l'UPE ait une moyenne et une variance commune dans une strate est raisonnable étant donné que les estimations de panel ont une moyenne et une variance communes. Pour l'estimation composite, même si l'on traite différemment les nouveaux panels et les anciens panels, les estimations composites au niveau du panel devraient avoir une même moyenne et une même variance indépendamment de l'attribution du panel. Cette condition est vérifiée parce que les panels sont attribués au hasard; un panel pourrait être un nouveau panel auquel est associée la probabilité $\gamma = 1/6$ et un ancien panel auquel est associée la probabilité $1 - \gamma = 5/6$. L'estimation inconditionnelle de la variance résultante ne serait pas plus faible que celle obtenue en la subordonnant à l'attribution du panel. Donc, la méthode du jackknife devrait produire une estimation prudente de la variance dans le contexte de l'EPAC. On notera que les considérations qui précèdent pour les mesures de l'incertitude n'englobent pas le biais dû au groupe de renouvellement qui pourrait exister.

5. RÉSULTATS DE L'ÉVALUATION

Les résultats numériques se fondent sur les données de l'EPAC de 1996 pour l'Ontario (voir Singh, et coll. 1997). Les variables auxiliaires pour la régression généralisée sont les totaux de population correspondants à 16 groupes âge-sexe, 11 régions économiques, 10 régions métropolitaines de recensement et 6 panels. Chaque total de contrôle de panel définit un sixième de la population de 15 ans et plus. Les nouveaux totaux de contrôle (30 en tout) utilisés pour la régression composite correspondant uniquement aux prédicteurs axés sur la variation sont : les personnes occupées, les chômeurs et les personnes qui ne font pas partie de la population active selon le groupe âge (jeunes/vieux)-sexe (soit, en tout, 12 groupes), l'emploi selon le groupe de branches d'activité (soit, en tout, 16) et l'emploi selon le travail à temps plein ou à temps partiel (soit, en tout, 2). En fait, ces 30 nouvelles variables de contrôle se réduisent à 28 seulement, à cause de la dépendance linéaire. L'estimateur de régression composite multidimensionnelle comprend ces 28 variables de contrôle supplémentaires, tandis que l'estimateur de régression composite unidimensionnelle ne comprend qu'une seule variable de contrôle supplémentaire. L'efficacité moyenne relative présentée dans divers tableaux est calculée comme étant la variance moyenne de l'estimateur de régression généralisée sur les 12 mois de 1996 divisée par la variance moyenne de l'estimateur composite sur ces 12 mois.

5.1 Prédicteurs de niveau macro c. micro

Pour les estimations du niveau, on calcule la corrélation entre l'estimation du niveau du mois courant (c'est-à-dire F_t) et le prédicteur (c'est-à-dire le prédicteur axé sur le niveau $C_{t-1} - \bar{D}_{t-1}$ au niveau macro), tandis que pour l'estimation de la variation, on calcule la corrélation entre $F_t - C_{t-1}$ et les prédicteurs. Comme il faut s'y attendre, la corrélation est négative, car l'estimation portant sur les panels communs est corrélée positivement à F_t, mais exprimée avec un signe négatif dans le prédicteur. Rappelons que l'estimateur composite utilisé est *ak* pour les prédicteurs de niveau macro et *ak** pour les prédicteurs de niveau micro.

Le tableau 1 montre que, pour les quatre variables importantes (personnes occupées, chômeurs, personnes occupées dans le secteur du commerce et personne occupée dans le secteur des transports et des communications (TRCO)), pour chaque prédicteur axé sur le niveau ou axé sur la variation, le prédicteur de niveau micro donne lieu à une plus forte corrélation que le prédicteur de niveau macro.

Si l'on compare les prédicteurs axés sur le niveau et sur la variation au niveau micro, on constate que ceux axés sur la variation donnent de meilleurs résultats que ceux axés sur le niveau. Les résultats sont comparables pour d'autres variables importantes. Étant donné ces corrélations, les autres résultats d'évaluation présentés plus loin se rapportent uniquement aux versions *ak*2, *ak**2 et rc2 des estimations composites. En vue d'éviter un nombre trop élevé de variables de contrôle supplémentaires, nous n'avons inclus l'estimateur rc ni pour les prédicteurs axés sur le niveau ni pour ceux axés sur la variation.

5.2 *ak c. ak* c.* rc (efficacité comparativement à gr)

Le tableau 2 montre les coefficients optimaux (par exemple, *k* pour l'estimateur *ak*2) et l'efficacité relative correspondante par rapport à l'estimateur gr. Nous avons obtenu les coefficients optimaux par recherche par compartiment en utilisant les données de 1996. (En pratique, cette recherche devrait se fonder sur des données antérieures). Nous constatons que les gains d'efficacité peuvent être considérables lorsque l'on passe de *ak* à *ak**. Les coefficients optimaux diffèrent pour les estimations du niveau et de la variation. Les deux dernières colonnes sous les rubriques « niveau » et « variation » montrent la réduction de l'efficacité si l'on se sert des coefficients optimaux pour

Tableau 1
Corrélation mensuelle moyenne entre les prédicteurs et les estimations composites pour le niveau et la variation (Ontario, 1996)

	Niveau				Variation			
	Prédicteurs axés sur le niveau		Prédicteurs axés sur la variation		Prédicteurs axés sur le niveau		Prédicteurs axés sur la variation	
Variable	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
Personnes occupées	-0,27	-0,35	-0,23	-0,45	-0,35	-0,49	-0,57	-0,84
Chômeurs	-0,26	-0,35	-0,24	-0,33	-0,22	-0,40	-0,39	-0,53
Emploi Commerce	-0,58	-0,55	-0,58	-0,65	-0,65	-0,73	-0,91	-0,96
Emploi TRCO	-0,58	-0,55	-0,60	-0,68	-0,63	-0,70	-0,92	-0,96

Tableau 2
Efficacité relative moyenne des estimateurs *ak* et *ak** par rapport à l'estimateur gr (Ontario, 1996)

	Coeff.				Eff. (Niveau)		Eff. (Variation)		Eff. (Niveau)	Eff. (Variation)
	Optimal-niveau		Optimal-variation		Optimal-niveau		Optimal-variation		Optimal-variation	Optimal-niveau
Variable	*ak*	*ak**	*ak*	*ak**	*ak*	*ak**	*ak*	*ak**	*ak*	*ak**
Personnes occupées	0,42	0,72	0,48	0,95	1,05	1,25	1,28	2,43	0,72	2,21
Chômeurs	0,40	0,50	0,54	0,69	1,06	1,12	1,11	1,29	1,05	1,26
Emploi Commerce	0,79	0,84	0,95	0,98	1,43	1,67	2,36	4,97	0,88	4,22
Emploi TRCO	0,84	0,87	0,95	0,98	1,59	1,88	3,60	7,59	1,11	6,51

le niveau pour estimer la variation et inversement. Les coefficients optimaux pour le niveau semblent donner d'assez bons résultats pour les estimations de la variation, tandis que l'utilisation des coefficients optimaux pour la variation pour estimer le niveau font baisser l'efficacité.

Le tableau 3 donne une comparaison de l'estimateur rc (unidimensionnel et multidimensionnel) à l'estimateur *ak**. Les valeurs possibles des coefficients $b_{i(rc2)}$ sur la période de 12 mois pour l'estimateur unidimensionnel rc2 sont résumées au moyen de la moyenne, du minimum et du maximum. On peut comparer ces valeurs aux coefficients optimaux correspondants pour *ak**. Les coefficients rc semblent représenter un compromis et se situer entre les valeurs du coefficient optimal pour le niveau et du coefficient optimal pour la variation. Les valeurs de l'efficacité de rc pour l'estimation de la variation sont pour ainsi dire équivalentes à celles obtenues pour *ak**, mais sont un peu plus faibles pour les estimations du niveau. Les gains d'efficacité sont faibles au niveau agrégé pour lequel il existe des variables de contrôle pour l'estimateur gr mais sont grands pour les domaines sans contrôle pour gr.

Le tableau 4 présente les pertes possibles d'efficacité pour les estimations obtenues par différences pour maintenir la cohérence interne dans le cas de l'estimateur *ak**. Les résultats indiquent qu'en pratique, il faut se montrer prudent lors du choix des variables pour l'estimation par différence ou de l'utilisation de valeurs intermédiaires (de compromis) des coefficients pour l'estimation *ak** des composantes d'un agrégat.

5.3 Estimation de la variation c. estimation du niveau – efficacité de rc par rapport à gr

Le tableau 5 montre que la relation approximative (voir la section 3) entre l'efficacité de l'estimation de la variation et de l'estimation du niveau est assez bien vérifiée. Le tableau montre que la corrélation des estimations rc obtenue d'un mois à l'autre pour les domaines pour lesquels il n'existe pas de contrôle de la population correspondant dans gr peut être assez forte comparativement à la corrélation observée pour l'estimation gr. Cette forte corrélation se traduit, à son tour, par une efficacité de l'estimation de la variation nettement plus grande que celle de l'estimation du niveau.

5.4 Estimation ponctuelle et écart-type de la différence entre rc et gr

Le tableau 6 montre les estimations mensuelles (et l'é.-t. des estimations du niveau et de la variation) pour la variable (personnes occupées dans le secteur du commerce en Ontario) pour les estimateurs gr et rc. Les valeurs correspondantes pour la différence mensuelle (rc - gr) ne sont pas présentées. Nous constatons qu'en général, la différence entre rc et gr n'est pas significative. Les valeurs de l'efficacité (qui ne sont pas présentées ici) des estimations annuelles moyennes et trimestrielles ont également été calculées pour rc et gr. Comme il fallait s'y attendre, rc pourrait donner lieu à une perte d'efficacité par rapport à gr, à cause de la corrélation sériale. Cependant, en ce qui concerne le coefficient de variation, cette perte d'efficacité semble n'avoir aucune conséquence pratique.

Tableau 3
Efficacité relative moyenne de l'estimateur rc par rapport à l'estimateur gr (Ontario, 1996)

	Coeff.					Eff. (Niveau)		Eff. (Variation)	
	rc - unidimensionnel (niveau ou variation)			*ak**		rc (uni-dimensionnel)	rc (multi-dimensionnel)	rc (uni-dimensionnel)	rc (multi-dimensionnel)
Variable	Moy	Min	Max	Niveau	Variation				
Personnes occupées	0,88	0,81	0,90	0,72	0,95	1,05	1,05	2,39	2,46
Chômeurs	0,60	0,53	0,65	0,50	0,69	1,12	1,12	1,31	1,33
Emploi Commerce	0,96	0,94	0,98	0,84	0,98	1,17	1,22	4,98	5,07
Emploi TRCO	0,95	0,93	0,97	0,87	0,98	1,37	1,42	7,47	7,52

Tableau 4
Efficacité relative moyenne des estimateurs *ak** et rc par rapport à l'estimateur gr, Ontario, 1996 (ordinaire c. différence)

		Niveau			Variation		
Variable		*ak* *Coeff	Eff (*ak* *)	Eff (*rc*)	*ak* *Coeff	Eff (*ak* *)	Eff (*rc*)
Agriculture	(ordinaire)	0,91	2,55	2,32	0,97	4,88	5,22
Agriculture	(différence)	NA	0,63	2,32	NA	3,90	5,22
NILF	(ordinaire)	0,74	1,26	1,07	0,95	1,96	2,01
NILF	(différence)	NA	1,21	1,07	NA	1,95	2,01

Tableau 5
Relation entre les efficacités des estimations du niveau et de la variation pour rc (multidimensionnel) par rapport à gr (Ontario, 1996)

Variable	Eff. Variation	Eff. Niveau	Eff. Variation/Eff. Niveau	$(1-\rho_{gr})(1-\rho_{rc})$	ρ_{gr}	ρ_{rc}
Personnes occupées	2,46	1,05	2,34	2,65	0,77	0,91
Chômeurs	1,33	1,12	1,19	1,21	0,5	0,59
Emploi Commerce	5,07	1,22	4,16	3,80	0,79	0,95
Emploi TRCO	7,54	1,42	5,31	5,66	0,8	0,97

5.5 Séries chronologiques des estimations du niveau

Les figures 1(a) et (b) montrent les estimations du niveau de l'emploi pour l'Ontario pour la période allant de 1988 à 1996 obtenues au moyen de gr et rc avec et sans désaisonnalisation (par la méthode X11-ARIMA). Les figures 2(a) et (b) montrent l'emploi pour le groupe de branches d'activité reprises sous « commerce ». Au niveau provincial, pour l'agrégation au niveau du groupe de branches d'activité, les séries (désaisonnalisée ou non) obtenues par régression généralisée (gr) et par régression composite (rc) sont semblables, parce qu'avant tout, les estimations par régression généralisée sont d'une haute précision. Par contre, au niveau du domaine « commerce », les séries sont assez différentes. (Notons que nous avons choisi cette série particulière parmi les nombreuses que nous avons examinées afin d'illustrer ici le scénario extrême pour ce qui est des écarts entre les séries gr et rc. Pour presque toutes les autres branches d'activité, les deux séries se recoupent assez fréquemment.) Puisque la série gr est fortement instable, l'utilisation de l'estimateur rc peut donner lieu à un lissage considérable. Remarquons aussi qu'à cause du ratio signal-bruit élevé prévu, la série rc désaisonnalisée au niveau du domaine « commerce » paraît nettement plus lisse que la série gr; en fait, il y a fort peu d'écarts entre les séries gr désaisonnalisée et non désaisonnalisée. Nous constatons qu'on observe, en général, une série de périodes consécutives où l'estimation rc est soit plus grande soit plus faible que l'estimation gr. Cette situation est prévisible étant donné les valeurs élevées des coefficients $b_{s(rc)}$ (tableau 3) et la forte corrélation sériale des deux séries (voir le tableau 5). Curieusement, les points de renversement dans les séries gr et rc ont tendance à survenir (approximativement) au même moment, quoi qu'ils paraissent légèrement atténués dans le cas de rc à cause d'une corrélation sériale plus forte pour la série rc. Notons aussi que l'écart entre les deux séries serait plus faible si l'on incluait aussi les variables de contrôle pour les prédicteurs axés sur le niveau.

Tableau 6
Estimations ponctuelles mensuelles au moyen de gr et rc et leurs différences (Ontario, 1996)
(Niveau et variation pour l'emploi dans le secteur du commerce, Ontario, 1996)

Mois	Type	gr		rc		rc-gr	
Janvier	Niveau	886,5	(21,0)	858,9	(17,3)	(27,6)	(23,0)
	Variation	-25,8	(13,2)	-21,0	(5,6)	4,8	(11,4)
Février	Niveau	906,5	(22,9)	867,9	(17,6)	38,6	(24,6)
	Variation	20	(14,2)	9,0	(4,7)	-11,0	(12,5)
Mars	Niveau	927,1	(20,8)	874,1	(18,3)	-52,9	(23,1)
	Variation	20,6	(13,3)	6,2	(4,7)	-14,4	(12,5)
Avril	Niveau	914,8	(20,3)	872,5	(17,7)	-42,3	(22,4)
	Variation	-12,3	(13,4)	-1,6	(5,1)	10,7	(12,5)
Mai	Niveau	912,8	(18,9)	887,6	(17,0)	-25,1	(21,8)
	Variation	-2,1	(13,0)	15,1	(5,7)	17,2	(11,6)
Juin	Niveau	908,1	(17,8)	888,6	(17,2)	-19,5	(21,5)
	Variation	-4,7	(12,3)	0,9	(4,9)	5,6	(11,9)
Juillet	Niveau	899,9	(18,1)	881,2	(17,7)	-18,7	(23,0)
	Variation	-8,2	(12,8)	-7,4	(6,7)	0,8	(10,7)
Août	Niveau	913,9	(16,9)	888,1	(18,3)	-25,8	(22,6)
	Variation	14	(11,5)	6,9	(5,3)	-7,1	(10,3)
Septembre	Niveau	886,6	(20,4)	876,4	(19,7)	-10,2	(23,1)
	Variation	-27,3	(12,6)	-11,8	(6,3)	15,6	(11,1)
Octobre	Niveau	898,6	(22,9)	889,3	(19,3)	9,3	(26,1)
	Variation	12,1	(13,4)	12,9	(6,6)	0,9	(11,8)
Novembre	Niveau	911,2	(20,3)	902,3	(19,3)	-8,9	(25,9)
	Variation	12,6	(13,9)	13,0	(7,0)	0,4	(12,6)
Décembre	Niveau	917,9	(20,5)	916,3	(19,0)	-1,5	(26,0)
	Variation	6,7	(12,5)	14,0	(6,1)	7,4	(10,9)

Nota : Les écarts-types sont donnés entre parenthèses.

6. CONCLUSIONS

L'estimateur par régression généralisée (gr) utilisé antérieurement dans le cadre de l'EPAC produisait des séries instables d'estimations de la variation et de diverses estimations au niveau de domaines. L'estimateur de régression composite (rc) produit des séries d'estimations plus lisses (qui, à leur tour, rendent les estimations de la variation plus stables). La méthode de régression composite s'écarte de l'estimation composite *ak* classique de plusieurs façons, les différences les plus importantes étant le recours à un micro-appariement pour la collecte des données de la période précédente au niveau de l'unité pour les panels communs et au calage par régression (comme dans le cas de l'estimateur gr) pour produire un ensemble de poids finals applicables à toutes les variables étudiées. Nous avons étudié trois versions de l'estimateur rc. Bien que cet article porte principalement sur l'estimateur rc2, c'est-à-dire l'estimateur avec prédicteur axé sur la variation (à cause du lissage résultant souhaité de la série d'estimation), nous avons constaté (bien que les résultats ne soient pas présentés ici) que les estimations du niveau de certaines variables importantes peuvent encore être améliorées (comparativement aux estimations rc2) en incluant les prédicteurs axés sur le niveau correspondant. Donc, en pratique, une bonne stratégie consisterait à utiliser un mélange comptant principalement des prédicteurs axés sur la variation et quelques prédicteurs axés sur le niveau.

La version de l'estimateur rc appliquée à l'heure actuelle pour l'EPAC, qui a été proposée par Fuller (1999), peut être formulée comme suit

$$C_{t(rc\alpha)} = F_t + b_{t(rc\alpha)}[(1-\alpha)(\tilde{C}_{t-1(rc\alpha)} - \bar{D}^*_{t-1} + \bar{B}_t - F_t) + \alpha(\tilde{C}_{t-1(rc\alpha)} - \bar{D}^*_{t-1})] \qquad (6.1)$$

où α est fixé (à, disons, 1/3, mais, de façon générale, pourrait être particulier à y) et le coefficient $b_{t(rc\alpha)}$ est calculé au moyen du système de régression généralisée comme s'il faisait partie de la classe rc d'estimations. On obtient une interprétation simple de l'équation (6.1) en la comparant à l'équation (3.7) de l'estimateur *ak**. Commençons par écrire (3.7) sous la forme

$$C_{t(ak^*)} = F_t + k^*[(1 - a^*/k^*)(\tilde{C}_{t-1(ak^*)} - \bar{D}^*_{t-1} + \bar{B}_t - F_t)$$

$$+ (a^*/k^*)(\tilde{C}_{t-1(ak^*)} - \bar{D}^*_{t-1})]. \qquad (6.2)$$

Maintenant, dans (6.1), α peut être considéré de façon grossière comme le quotient des deux coefficients optimaux a^* *et* k^*, et le facteur k^* à l'extérieur des crochets de (6.2) peut être remplacé par le coefficient de régression (sous-optimal) $b_{t(rc\alpha)}$. Donc, $C_{t(rc\alpha)}$ n'est pas l'équivalent de l'estimateur ak^* optimal, mais on pourrait retenir une certaine optimalité (si l'on faisait en sorte que α soit particulier à y) en fixant la contribution relative des prédicteurs axés sur la variation et sur le niveau. Notons cependant que le problème de l'incohérence interne mentionné dans l'introduction pourrait se poser si α est lié à y. D'autres caractéristiques intéressantes de cette version sont, d'une part, que l'on peut choisir la valeur de α de façon qu'elle soit strictement non nulle (de façon à éviter les écarts prolongés entre les séries gr et rc) et, d'autre part, que le nombre de variables de contrôle supplémentaires nedouble pas lorsque l'on inclut à la fois les prédicteurs axés sur le niveau et ceux axés sur la variation, ce qui permet d'introduire un plus grand nombre de variables de contrôle ainsi qu'un plus grand nombre de degrés de liberté dans l'estimation de la variance.

Figure 1 (a) Emploi Ontario, non-désaisonnalisé

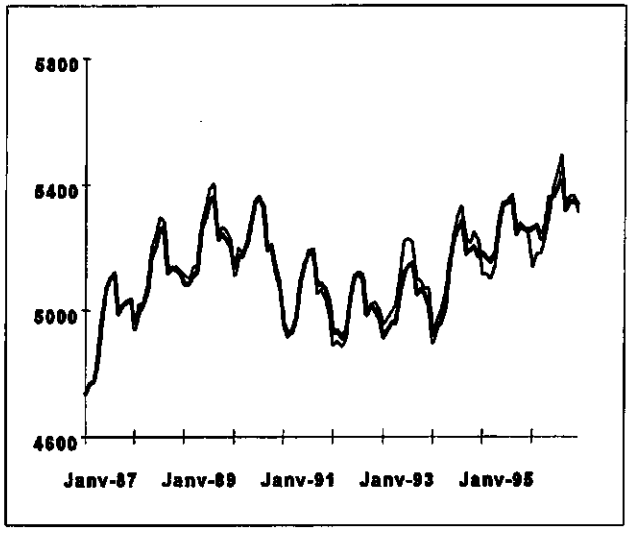


Figure 1 (b) Emploi Ontario, désaisonnalisé

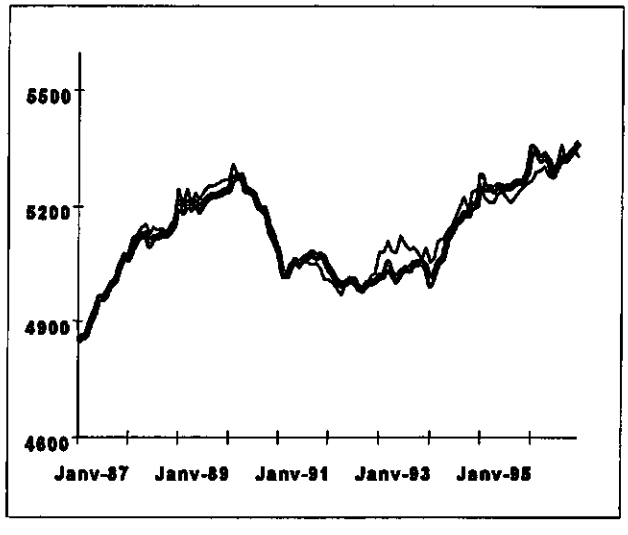


Figure 2 (a)
Emploi dans le commerce ,Ontario non-désaisonnalisé

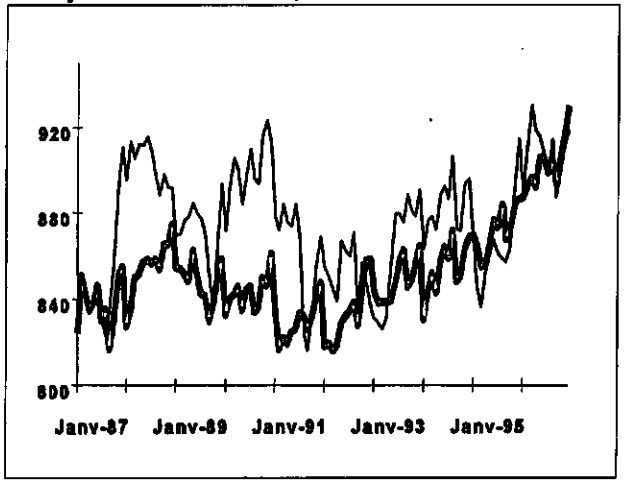
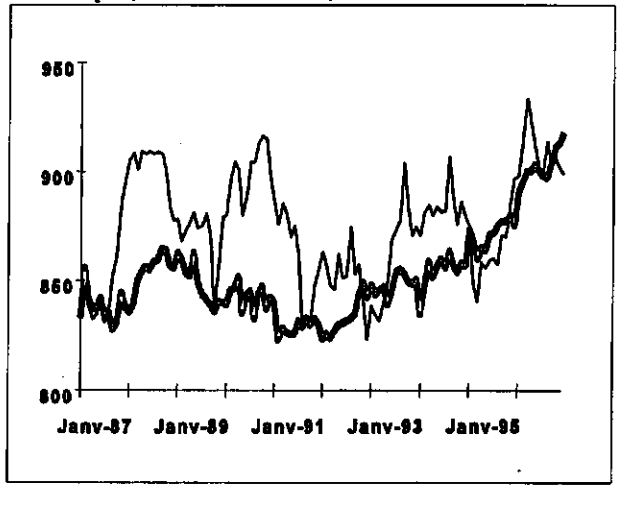


Figure 2 (b)
Emploi, dans le commerce , Ontario désaisonnalisé



La simple vérification qui suit peut être exécutée pour diagnostiquer l'effet du biais dû à l'imputation des données sur la situation d'emploi du mois précédent pour compenser la non-réponse de certains répondants du mois courant. L'idée fondamentale est de calculer un facteur multiplicatif de redressement du biais applicable à l'estimateur $\bar{D}^*_{t-1}$ qui comprend les valeurs imputées. Le facteur est défini comme étant le quotient des deux estimations gr de la caractéristique du mois précédent fondées sur le sous-échantillon apparié. Le dénominateur est une estimation gr pour le mois précédent (comprenant des valeurs imputées), tandis que le numérateur est une estimation gr pour le mois précédent (ne comprenant pas de valeur imputée) et ils sont tous deux calculés selon une méthode qui n'est pas tout à fait typique. Pour le numérateur, nous utilisons les répondants de la période $t-1$ avec les réponses qu'ils ont données à la période $t-1$ et, après correction des poids de sondage pour la non-réponse, nous construisons l'estimateur gr avec les variables de contrôle pour la période t. Pour le dénominateur, nous supposons que les sous-ensembles formés par chaque paire de sous-échantillons appariés aux périodes $t-1$ et t (ici l'appariement des sous-échantillons est fait l'un en fonction de l'autre, l'un en avance dans le temps et l'autre en retard) n'ont pas de contrepartie à cause de la non-réponse, donc sont statistiquement remplaçables l'un par l'autre. Puis, nous remplaçons les répondants de la période $t-1$ qui n'ont pas répondu à la période t par les non-répondants à la période $t-1$ qui ont répondu à la période t, avec les réponses qui leur ont été imputées à la période $t-1$, ainsi que les poids de sondage. Ensuite, nous procédons de nouveau à la correction des poids pour la non-réponse et pour la stratification a posteriori de gr (au moyen des variables de contrôle pour t) pour cet échantillon complet modifié à la période $t-1$. Les poids gr ainsi obtenus sont utilisés pour calculer le dénominateur mentionné plus haut. Nous pouvons maintenant examiner la série chronologique de ce facteur sur plusieurs mois pour diagnostiquer le biais dû à l'imputation. S'il n'est pas jugé proche de 1, alors nous pouvons traiter la moyenne du facteur sur plusieurs mois comme un redressement non aléatoire multiplicatif du biais appliqué à $\bar{D}^*_{t-1}$. En pratique, au lieu de rajuster $\bar{D}^*_{t-1}$, il serait préférable, pour faciliter les calculs, de rajuster la nouvelle variable de contrôle $\bar{C}_{t-1(rc)}$ (de l'équation 3.6) pour la caractéristique correspondante en lui appliquant l'inverse du facteur multiplicatif susmentionné. Autrement, on pourrait éviter tout bonnement l'imputation si l'on pouvait modifier le questionnaire pour obtenir les données nécessaires sur la période antérieure telle que nous le proposons dans l'introduction.

Dans leur étude, Lent, Miller et Cantwell (1994, 1996) considèrent l'estimateur composite pondéré *ak* pour la *Current Population Survey* des États-Unis comme un remplacement possible de l'estimateur *ak* utilisé à l'heure actuelle si a=0,2 et k=0,4. Selon notre expérience concernant l'estimateur *ak**, l'estimateur composite pondéré *ak** pourrait être une meilleure solution de rechange si l'on considère les gains d'efficacité.

REMERCIEMENTS

Le gros de ces travaux de recherche a été réalisé pendant que les premier et troisième auteurs travaillaient à Statistique Canada. Les auteurs remercient M. Sheridan, J.D. Drew, J. Gambino et particulièrement M.P. Singh pour leurs encouragements et plusieurs discussions fort fructueuses. Ils sont reconnaissants à Jon Rao et Wayne Fuller de leurs commentaires et suggestions. Ils remercient aussi J.M. Levesque, P. Lorenz et, tout spécialement, T. Merkouris (grâce auquel ces travaux ont débuté) pour l'aide qu'ils leur ont apportée pour analyser et interpréter les résultats. Enfin, ils remercient l'examinateur et le rédacteur en chef adjoint Harold Mantel pour leurs commentaires constructifs lors de la révision de l'article. Les travaux de recherche du premier auteur ont été financés en partie par une bourse du CRSNG obtenue à l'Université Carleton aux termes d'un professorat de recherche adjoint.

BIBLIOGRAPHIE

BAILAR, B.A. (1975). The effect of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-29.

BINDER, D.A., et HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, 6 : Sampling, Elsevier Science, NY, 187-211.

CASSEL, C.M., SÄRNDAL, C.-E. et WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

COCHRAN, W.G. (1977). *Sampling Techniques*. 3e édition. John Wiley and Sons.

FULLER, W.A. (1990). Analyse d'enquête à passages répétés. *Techniques d'enquête*, 16, 177-190.

FULLER, W.A. (1999). The Canadian Regression Composite Estimation. Manuscript non-publié.

GODAMBE, V.P., et THOMPSON, M.E. (1989). An extension of quasi-likelihood estimation (avec discussion). *Journal Statistical Planning and Inference*, 22, 137-172.

GURNEY, M., et DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 247-257.

HANSEN, M.H., HURWITZ, W.N. et MADOW, W.G. (1953). *Sample Survey Methods and Theory*. 2, John Wiley and Sons.

KUMAR, S., et LEE, H. (1983). Évaluation de l'application d'estimateurs composites à l'enquête sur la population active du Canada. *Techniques d'enquête*, 9, 196-221.

LENT, J., MILLER, S. et CANTWELL, P. (1994). Composite weights for the current population survey. *Proceedings of the Survey Research Methods Section, American Statistical Assocation*, 867-872.

LENT, J., MILLER, S. et CANTWELL, P. (1996). Effects of composite weights on some estimates from the Current Population Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, I, 130-139.

LIANG, K.-Y., et ZEGER, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

RAO, J.N.K., et GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.

SÄRNDAL, C. -E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

SINGH, A.C. (1994). Sampling-design-based estimating functions for finite population totals. Invited paper, *Abstracts of the Statistical Society of Canada, Annual Meeting, Banff*, Alberta, 8-11, page 48.

SINGH, A.C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1, 120-129.

SINGH, A.C., et FOLSOM, R.E. Jr. (2000). Bias corrected estimating functions approach for variance estimation adjusted for poststratification. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 610-615.

SINGH, A.C., KENNEDY, B., WU, S. et BRISEBOIS, F. (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 300-305.

SINGH, A.C., et MERKOURIS, P. (1995). Composite Estimation by modified regression for repeated surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 420-425.

WOLTER, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

Techniques d'enquête, juin 2001
Vol. 27, n° 1, pp. 49-56
Statistique Canada

Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada

WAYNE A. FULLER et J.N.K. RAO[1]

RÉSUMÉ

L'enquête sur la population active du Canada est une enquête mensuelle menée auprès de ménages sélectionnés en fonction d'un plan stratifié à plusieurs degrés. L'échantillon de ménages est divisé en six panels (groupes de renouvellement). Un panel reste dans l'échantillon pendant six mois consécutifs, pour être ensuite éliminé de l'échantillon. Par le passé, un estimateur de régression généralisée, fondé uniquement sur les données du mois courant, a été mis en oeuvre à l'aide d'un programme de poids de régression. Dans le présent exposé, les auteurs examinent des procédures d'estimation composite de régression faisant appel à des données d'échantillon tirées de périodes précédentes et pouvant être mises en oeuvre dans le cadre d'un programme de poids de régression. Singh (1996) a proposé un estimateur composite appelé MR2 qu'on peut calculer en ajoutant des variables x au programme courant de poids de régression. L'estimateur de Singh est considérablement plus efficace que l'estimateur de régression généralisée pour un changement d'une période, mais non pas pour le niveau courant. De plus, l'estimateur de niveau peut s'écarter de l'estimateur de régression généralisée de façon appréciable, et cet écart peut persister longtemps. Les auteurs proposent un estimateur « compromis », en utilisant un programme de poids de régression et le même nombre de variables x que MR2, qui est plus efficace pour le niveau et pour le changement que l'estimateur de régression généralisée fondé uniquement sur les données du mois courant. L'estimateur proposé tient compte également du problème de dérive et s'applique à d'autres enquêtes faisant appel à un échantillonnage par renouvellement.

MOTS CLÉS : Échantillonnage; groupes de renouvellement; combinaison d'estimateur.

1. INTRODUCTION

L'estimation composite est un concept utilisé dans l'échantillonnage d'enquête pour décrire des estimateurs de période courante utilisant l'information de périodes antérieures d'une enquête périodique comportant un plan de renouvellement. Lorsque certaines unités sont observées au cours de certaines périodes, mais non pas toutes, on peut en profiter pour améliorer les estimations pour toutes les périodes.

Statistique Canada, le Bureau of the Census des États-Unis et d'autres bureaux de la statistique font appel à un plan de renouvellement pour les enquêtes sur la population active. L'Enquête sur la population active courante du Canada (EPA) est une enquête mensuelle menée auprès de quelque 59 000 ménages sélectionnés en fonction d'un plan d'échantillonnage stratifié à plusieurs degrés. L'unité d'échantillonnage ultime est le ménage, et un échantillon de ménages est divisé en six panels (groupes de renouvellement). Un groupe de renouvellement reste dans l'échantillon pendant six mois consécutifs, pour ensuite être éliminé complètement de l'échantillon. Ainsi, les cinq sixièmes de l'échantillon de ménages sont communs pour deux mois consécutifs. Singh, Drew, Gambino et Mayda (1990) et Gambino, Singh, Dufour, Kennedy et Lindeyer (1998) ont présenté des descriptions détaillées du plan de l'EPA. Dans la Current Population Survey (CPS) des États-Unis, l'échantillon est constitué de huit groupes de renouvellement. Un groupe de renouvellement reste dans l'échantillon pendant quatre mois consécutifs, quitte l'échantillon pour les huit mois suivants, puis retourne à l'échantillon pour quatre mois consécutifs. Il est alors éliminé complètement de l'échantillon. Il y a donc un chevauchement de 75 % d'un mois à l'autre dans l'échantillon, et un chevauchement de 50 % d'une année à l'autre dans l'échantillon (Hansen, Hurwitz, Nisselson et Steinberg 1955).

Patterson (1950), suivant les travaux initiaux de Jessen (1942), a décrit le fondement théorique du plan et de l'estimation des enquêtes répétées, à l'aide de procédures des moindres carrés généralisés. Pour la CPS, Hansen et coll. (1955) ont proposé un estimateur plus simple, appelé l'estimateur composite K. Gurney et Daly (1965) ont présenté un estimateur composite K amélioré, appelé l'estimateur composite AK avec deux facteurs de pondération A et K. Breau et Ernst (1983) ont comparé d'autres estimateurs à l'estimateur composite K pour la CPS. Rao et Graham (1964) ont étudié des schémas de remplacement optimal pour l'estimateur composite K. Eckler (1955) et Wolter (1979) ont étudié des schémas de renouvellement à deux niveaux comme celui utilisé dans la Retail Trade Survey des États-Unis. Yansaneh et Fuller (1998) ont étudié l'estimation récursive optimale pour des enquêtes répétées. Fuller (1990) et Lent, Miller, Cantwell et Duff (1999) ont élaboré la méthode des poids composites pour la CPS. On obtient les poids composites en appliquant la méthode itérative aux poids du plan en fonction de totaux de contrôle particuliers qui englobent les totaux de population de

[1] Wayne A. Fuller, Iowa State University, 221 Snedecor Hall, Ames, IA, U.S.A.; J.N.K. Rao, Université Carleton, Ottawa, (Ontario) K1S 5B6.

variables auxiliaires et les estimations composites K pour des caractéristiques d'intérêt y. À l'aide des poids composites, les utilisateurs peuvent préparer des estimations à partir de fichiers de microdonnées pour le mois courant sans avoir recours aux données des mois précédents.

Les auteurs mentionnés ci-dessus ont fait appel à la stratégie traditionnelle fondée sur le plan, en supposant que les totaux inconnus à chaque occasion étaient des paramètres fixes. D'autres auteurs (Scott, Smith et Jones 1977; Jones 1980; Binder et Dick 1989; Bell et Hillmer 1990; Tiller 1989 et Pfeffermann 1991) ont élaboré des estimations pour des enquêtes répétées en supposant que les valeurs sous-jacentes réelles étaient une réalisation d'une série chronologique.

Statistique Canada a considéré l'estimation composite K et AK pour l'Enquête sur la population active à plusieurs reprises au cours des 25 dernières années (Kumar et Lee 1983), mais n'a pas adopté l'estimation composite. On a plutôt calculé, à l'aide d'un programme de poids de régression, un estimateur de régression généralisée fondé uniquement sur les données du mois courant. Lorsque l'estimation composite a été considérée dans les années 1990, il y avait de fortes pressions pour que l'on élabore une procédure d'estimation composite faisant appel au programme d'estimation existant. Singh (1996) a proposé une méthode ingénieuse appelée régression modifiée (MR) pour surmonter la difficulté. Cette méthode donne lieu à un estimateur composite appelé MR2 qui utilise le programme existant de poids de régression. Singh a suggéré la création de variables x servant de variables de contrôle dans le programme de régression. Grâce aux variables créées et à l'estimateur de la période précédente, on utilise le programme existant de poids de régression pour établir des poids de régression qui définissent l'estimateur pour la période courante. On inclut également des variables de contrôle comportant des totaux de population connus.

Une étude empirique de l'estimateur MR2 a permis de cerner plusieurs caractéristiques de la procédure. Tout d'abord, la variance estimative d'un changement d'une période est fortement réduite. Deuxièmement, la variance estimative du niveau est souvent semblable à celle de l'estimateur direct. Troisièmement, l'estimateur de niveau peut s'écarter appréciablement de l'estimateur direct, et cet écart peut se prolonger longtemps.

Dans le présent exposé, les auteurs examinent l'efficacité des estimateurs MR2 théoriquement dans le cadre d'un montage simplifié. Ils utilisent également un estimateur « compromis » qui entraîne des gains appréciables en ce qui concerne l'efficacité tant pour le niveau que pour le changement relativement à l'estimateur qui n'utilise que les données du mois courant. L'estimateur composite tient compte également du problème de « dérive » mentionné ci-dessus, et peut être mis en oeuvre en fonction du programme existant de poids de régression. Gambino, Kennedy et Singh (2000) ont évalué l'efficacité des estimations composites pour les données de l'EPA, à l'aide d'une méthode jackknife d'estimation de la variance. Bell (2000) a comparé plusieurs estimateurs composites à l'aide de données tirées de l'enquête sur la population active de l'Australie.

2. ESTIMATION DE RÉGRESSION COMPOSITE

On utilise deux types d'observations dans l'estimation composite : les observations qui se font uniquement au moment t courant et les observations qui se font tant au moment courant qu'au moment précédent, $t-1$. Parfois, l'information des observations précédentes est comprimée dans la ou les estimations pour la ou les périodes précédentes. Soit w_i, le poids d'échantillonnage pour l'observation i au moment t, et A_t, l'ensemble d'éléments comportant des observations pour les deux périodes t et $t-1$, et B_t, l'ensemble d'éléments observé uniquement au cours de la période courante t. Dans un tel contexte initial, i est l'indice pour un même répondant. S'il n'y a pas de non- réponse, l'ensemble A_t pour l'EPA est constitué d'individus dans les cinq panels qui se trouvaient dans l'échantillon au cours de la période précédente, appelés panels de chevauchement. En l'absence de non-réponse, l'ensemble B_t pour l'EPA comporte des individus observés d'abord pendant la période courante, appelés le premier panel. Supposons que

$$\sum_{i \in A_t} w_i + \sum_{i \in B_t} w_i = \hat{N}_t = \text{total de population estimatif.}$$

Soit θ_t, la fraction de l'échantillon qui chevauche au moment t :

$$\theta_t = \hat{N}_t^{-1} \sum_{i \in A_t} w_i. \tag{2.1}$$

Dans l'EPA, θ_t est de 5/6 environ et presque constante au fil des ans. Par souci de simplicité, l'indice t de A_t, B_t et θ_t est souvent omis.

2.1 Estimateur

L'estimateur MR2 de Singh (1996) fait appel à la variable de contrôle

$$\begin{aligned} x_{1ti} &= \theta_t^{-1}(y_{t-1,i} - y_{ti}) + y_{ti} \quad \text{si } i \in A_t \\ &= y_{ti} \quad \text{si } i \in B_t, \end{aligned} \tag{2.2}$$

dans le programme de régression, où y_{ti} est la valeur d'une caractéristique d'intérêt, y, pour l'élément i au moment t. Compte tenu de la non-réponse dans l'EPA, la proposition originale de Singh faisait appel à l'imputation pour les données manquantes et posait $\theta = 5/6$, après l'imputation pour les données manquantes. Dans notre discussion initiale, nous utilisons la θ_t définie en (2.1), en supposant l'absence de non-réponse de façon que l'imputation ne soit pas requise. À noter que le « microappariement » de fichiers de données individuels aux moments $t-1$ et t est nécessaire

pour le calcul de x_{1ti} et de l'estimateur MR2 résultant. Des variables de contrôle supplémentaires ayant la forme (2.2) associées à d'autres variables y de même que des variables auxiliaires comportant des totaux de population connus sont également comprises dans l'estimation de régression. Les variables auxiliaires de l'EPA englobent des variables démographiques, comme l'âge, le sexe et l'emplacement.

Les variables x particulières en (2.2) sont conçues de façon que le total estimatif de x_1 soit un estimateur du total de y à la période précédente. Ainsi, le total de contrôle pour x_1 dans la procédure de régression est l'estimateur du total de y à la période précédente.

Soit $\hat{\mu}_{t-1}$, l'estimateur de la moyenne de y pour la période $t-1$, et $\bar{y}_{m,t-1}$ et $\bar{y}_{mt}$, les moyennes des panels appariés aux moments $t-1$ et t respectivement, et $\bar{y}_t$, la grande moyenne de tous les panels d'échantillon au moment t, et $\bar{y}_{B,t}$ la moyenne du premier panel au moment t. Supposons que l'échantillon de taille n est divisé en g panels de taille égale, et notons la fraction d'échantillonnage appariée à l'aide de θ. Afin de simplifier la discussion, nous considérons une seule variable y. On peut dès lors écrire l'estimateur MR2 de Singh (1996) au moment t, construit avec x_{1ti}, sous la forme d'un estimateur de régression comme suit

$$\hat{\mu}_t = \bar{y}_t + (\bar{x}_{CNt} - \bar{x}_{Ct})\hat{\beta}_{Ct} + \left[\hat{\mu}_{t-1} - (\bar{y}_{m,t-1} - \bar{y}_{m,t} + \bar{y}_t)\right] b_t, \quad (2.3)$$

où $\bar{x}_{CNt}$ est la moyenne de population du vecteur de variables auxiliaires, comme l'âge et le sexe, au moment t, $\bar{x}_{Ct}$ est la moyenne d'échantillon pondérée des variables auxiliaires et $(\hat{\beta}'_{Ct}, b_t)'$ est le vecteur de coefficients de régression pour la régression de y_t sur (x_{Ct}, x_{1t}).

On peut écrire

$$y_{t,i} = \hat{y}_{t,i,(r)} + d_{t,i,(r)},$$

où $\hat{y}_{t,i,(r)}$ est la valeur prévue dans la régression de $y_{t,i}$ sur x_{Cr} et $d_{t,i,(r)}$ est l'écart de la valeur prévue de régression. On a alors

$$\begin{aligned} x_{1t} &= \theta^{-1}(\hat{y}_{t-1,i,(t-1)} + d_{t-1,i,(t-1)} - \hat{y}_{t,i,(t)} - d_{t,i,(t)}) \\ &\quad + \hat{y}_{t,i,(t)} + d_{t,i,(t)} && \text{si } i \in A_t \\ &= \hat{y}_{t,i,(t)} + d_{t,i,(t)} && \text{si } i \in B_t. \end{aligned}$$

Pour les variables démographiques X_{Cti}, il est raisonnable de penser que $\hat{y}_{t-1,i,(t-1)}$ est proche de $\hat{y}_{t-1,i,(t)}$ et proche de $\hat{y}_{t,i,(t)}$. Par conséquent, la partie de x_{1t} qui est orthogonale à x_{Ct} est proche de

$$\begin{aligned} x_{d,1,t} &= \theta^{-1}(d_{t-1,i,(t-1)} - d_{t,i,(t)}) \\ &\quad + d_{t,i,(t)} && \text{si } i \in A_t \\ &= d_{t,i,(t)} && \text{si } i \in B_t. \end{aligned}$$

Ainsi, le coefficient de régression partielle b_t est proche du coefficient de régression pour la régression de $d_{t,i,(t)}$ sur $x_{d,1,t}$, et la valeur dépend de la corrélation entre $d_{t,i,(t)}$ et $d_{t-1,i,(t-1)}$. Un modèle simple pour $d_{t,i,(t)}$ qui a été utilisé dans le passé et que nous adoptons dans notre analyse est l'hypothèse selon laquelle $d_{t,i,(t)}$ est la somme d'un μ_t fixe et d'une erreur qui est une autorégression d'ordre un avec un paramètre ρ.

Afin de simplifier la présentation, nous parlons du modèle d'échantillonnage aléatoire simple sans x_{Ct}. Les résultats s'étendent au cas général si l'on considère le paramètre ρ comme étant la corrélation partielle entre y_t et y_{t-1} après rajustement pour x_{Ct}.

Dans le modèle autorégressif avec ρ fixe, une coordonnée à l'origine et pas d'autre $\mathbf{x}_{Ct}$ dans le modèle, il est possible de montrer que b_t converge pour ce qui est de la probabilité vers

$$b_0 = \rho \lim_{n \to \infty} b_t = \theta\rho\left[2 - \theta - 2(1-\theta)\rho - (1-\theta)\sigma_y^{-2}\Delta_t^2\right]^{-1},$$

où $\Delta_t^2 = (\mu_t - \mu_{t-1})^2$. Si nous supposons que $(1-\theta)\sigma_y^{-2}\Delta_t^2$ est faible relativement aux autres termes, nous obtenons

$$b_0 \doteq \theta\rho\left[2 - \theta - 2(1-\theta)\rho\right]^{-1}. \quad (2.4)$$

Pour l'EPA, $b_0 = (7-2\rho)^{-1}5\rho$.

On obtient d'autres représentations pour l'estimateur $\hat{\mu}_t$, en omettant x_{Ct}, grâce à la formule $\bar{y}_t = \theta\bar{y}_{mt} + (1-\theta)\bar{y}_{Bt}$. Ainsi

$$\begin{aligned} \hat{\mu}_t &= (1-b)\bar{y}_t + [\hat{\mu}_{t-1} + (\bar{y}_{mt} - \bar{y}_{m,t-1})]b \\ &= \theta\left[\bar{y}_{mt} + (\hat{\mu}_{t-1} - \bar{y}_{m,t-1})b\right] \\ &\quad + (1-\theta)(\hat{\mu}_{t-1} - \bar{y}_{m,t-1}y_{m,t-1} + \bar{y}_{mt})b + (1-\theta)(1-b)\bar{y}_{B,t} \\ &= [\theta + (1-\theta)b][\bar{y}_{m,t} + (\hat{\mu}_{t-1} - \bar{y}_{m,t-1})b^*] \\ &\quad + (1-\theta)(1-b)\bar{y}_{B,t} \\ &= \lambda_A[\bar{y}_{m,t} + (\hat{\mu}_{t-1} - \bar{y}_{m,t-1})b^*] + (1-\lambda_A)\bar{y}_{B,t}, \end{aligned} \quad (2.5)$$

où

$$1 - \lambda_A \approx (1-\theta)(1-b_0)$$

et

$$b^* \approx \left[\theta + (1-\theta)b_0\right]^{-1} b_0. \quad (2.6)$$

La première expression à la droite de l'égalité de (2.5) donne l'estimateur MR2 comme combinaison linéaire de l'estimateur direct $\bar{y}_t$ et de l'estimateur de différence $\hat{\mu}_{t-1} + (\bar{y}_{mt} - \bar{y}_{m,t-1})$, c'est-à-dire sous la forme d'un estimateur composite. L'expression finale de (2.5) donne l'estimateur comme combinaison linéaire d'un estimateur de « type régression » fondé sur les panels chevauchants et la moyenne des premiers panels.

2.2 Un autre estimateur

Il est possible de définir d'autres variables de régression à utiliser dans l'estimation composite de régression. Nous présentons une variable de régression particulière ci-dessous. L'estimateur de régression associé n'est pas proposé comme estimateur ultime, mais l'estimateur est membre d'une classe pour laquelle des calculs d'efficacité sont donnés. Un autre estimateur que l'estimateur MR2 de Singh (1996) est décrit à la section 5.

Définissons une variable égale à la valeur de la période précédente si l'individu se trouve dans l'échantillon chevauchant, et égale à la moyenne estimative pour la période précédente si l'individu se trouve dans le premier échantillon. La variable de régression est

$$\begin{aligned} x_{2,ti} &= y_{t-1,i}, \quad \text{si } i \in A_t \\ &= \hat{\mu}_{t-1} \quad \text{si } i \in B_t. \end{aligned} \tag{2.7}$$

Si l'on utilise cette variable dans un estimateur de régression, la moyenne de contrôle est $\hat{\mu}_{t-1}$, l'estimateur de la période précédente, puisque la moyenne pour la variable créée sert à estimer la moyenne pour la période $t-1$. Singh (1996) a utilisé une variable $\tilde{x}_{2ti}$ semblable à x_{2ti}. Dans la variable de Singh, le $\hat{\mu}_{t-1}$ en (2.7) est $\bar{y}_{m,t-1}$ si $i \in B_t$.

Considérons un estimateur de régression construit à l'aide de x_{2ti} et rappelons que la moyenne de contrôle de x_{2t} est $\hat{\mu}_{t-1}$. L'estimateur de régression utilisant x_{2t} peut s'écrire

$$\hat{\mu}_{\text{reg},t} = \bar{y}_t + (\hat{\mu}_{t-1} - \bar{x}_{2,t})\hat{\beta}, \tag{2.8}$$

où $\hat{\beta}$ est le coefficient de régression pour la régression de y_t sur x_{2t} (par souci de simplicité on laisse tomber l'indice t de $\hat{\beta}_t$), $\bar{y}_t$ est la moyenne de l'échantillon de y au moment t, et $\bar{x}_{2,t}$ est la moyenne de l'échantillon de $x_{2,ti}$ pour tous les panels de l'échantillon au moment $t-1$. Le coefficient de régression $\hat{\beta}$ est, approximativement, la régression de y_t sur x_{2t} dans l'ensemble A. Le coefficient n'est pas exactement le coefficient de régression pour l'ensemble A puisque $\bar{y}_{m,t-1}$ n'est pas égale à $\hat{\mu}_{t-1}$, mais la différence entre les deux estimateurs est habituellement faible. Singh (1996) a nommé l'estimateur de régression construit à l'aide de $\tilde{x}_{2ti}$ l'estimateur MR1.

Si l'on utilise $\bar{y}_t = \theta\bar{y}_{mt} + (1-\theta)\bar{y}_{Bt}$, l'estimateur de régression de μ_t utilisant x_{2t} comme variable de contrôle est donné par

$$\hat{\mu}_t = (1-\theta)\bar{y}_{B,t} + \theta\{\bar{y}_{m,t} + (\hat{\mu}_{t-1} - \bar{y}_{m,t-1})\hat{\beta}\}. \tag{2.9}$$

L'expression entre les accolades en (2.9) est l'estimateur de régression de μ_t utilisant l'estimateur $\hat{\mu}_{t-1}$ et uniquement les données tirées de l'échantillon apparié A. À noter que l'estimateur de régression

$$\hat{\mu}_{m,t} = \bar{y}_{m,t} + (\hat{\mu}_{t-1} - \bar{y}_{m,t-1})\hat{\beta}, \tag{2.10}$$

où $\hat{\beta}$ est la régression de y_t sur y_{t-1} dans l'ensemble A, est l'estimateur linéaire optimal pour μ_t fondé sur $\hat{\mu}_{t-1}$ et les données de l'ensemble A. À noter que $\beta = \rho$ si les variances sont les mêmes pour les deux périodes. Ci-dessous, nous posons souvent $\beta = \rho$.

Le fait d'utiliser la variable x_{2t} donne l'estimateur optimal, $\hat{\mu}_{mt}$, selon les données de l'ensemble A, sans toutefois combiner cet estimateur à la moyenne de l'ensemble B d'une façon optimale. Comme on peut le constater en (2.10), le poids donné à la moyenne de l'ensemble B est $1-\theta$. En général, ce poids est trop grand puisque la variance de l'estimateur de régression est inférieure à la variance de la moyenne simple.

3. ESTIMATION OPTIMALE

La façon choisie de combiner l'estimateur de régression pour l'ensemble A à la moyenne de l'ensemble B dépend d'une la fonction objective et de la variance de $\hat{\mu}_{t-1}$. Nous présentons des calculs à titre d'exemple fondés sur quelques hypothèses de simplification. Pour simplifier, exprimons $V\{\hat{\mu}_{t-1}\}$ comme un multiple de la variance du premier panel,

$$V\{\hat{\mu}_{t-1}\} = q_t^{-1} V\{\bar{y}_{B,t}\}. \tag{3.1}$$

Supposons

$$V\{\bar{y}_t\} = g^{-1} V\{\bar{y}_{B,t}\}, \tag{3.2}$$

$$\text{Cov}\{\hat{\mu}_{t-1}, (\bar{y}_{m,t} - \bar{y}_{m,t-1}\beta)\} = 0, \tag{3.3}$$

et

$$\text{Cov}\{\bar{y}_{B,t}, (\bar{y}_{m,t} - \bar{y}_{m,t-1}\beta)\} = 0, \tag{3.5}$$

où g est le nombre de groupes de renouvellement (panels). L'hypothèse (3.1) est raisonnable si les panels originaux comportent une fonction de covariance dont celle d'un processus autorégressif d'ordre un représente une bonne approximation. Pour l'EPA, les covariances nulles en (3.4) et en (3.5) et l'hypothèse (3.2) sont uniquement des approximations puisque $\bar{y}_{B,t}$ ne se fonde pas sur un échantillon entièrement indépendant.

Nous écrivons l'estimateur de régression en fonction du chevauchement sous la forme

$$\hat{\mu}_{m,t} = \bar{y}_{m,t} - \bar{y}_{m,t-1}\beta + \hat{\mu}_{t-1}\beta$$

et, suivant les hypothèses, nous obtenons

$$V(\hat{\mu}_{mt}) \doteq [g^{-1}\theta^{-1}(1-\rho^2) + q_t^{-1}\rho^2]\, V\{\bar{y}_{B,t}\}. \tag{3.6}$$

Pour l'EPA, $g = 6$ est le nombre de panels. Considérons maintenant un estimateur qui est une combinaison linéaire de $\hat{\mu}_{m,t}$ et de $\bar{y}_{B,t}$,

$$\begin{aligned} \hat{\mu}_t &= \lambda\hat{\mu}_{mt} + (1-\lambda)\bar{y}_{B,t} \\ &= \lambda(\bar{y}_{m,t} - \bar{y}_{m,t-1}\beta + \hat{\mu}_{t-1}\beta) + (1-\lambda)\bar{y}_{B,t}, \end{aligned} \tag{3.7}$$

où $0 \le \lambda \le 1$ est à déterminer. Afin de réduire la variance du niveau courant au minimum, étant donné $\hat{\mu}_{t-1}$ avec une variance $q_t^{-1} V\{\bar{y}_{B,t}\}$, on réduit

$$V\{\hat{\mu}_t\} = V\{\lambda \hat{\mu}_{mt} + (1-\lambda)\bar{y}_{B,t}\}$$

$$= \lambda^2 V\{\hat{\mu}_{mt}\} + (1-\lambda)^2 V\{\bar{y}_{B,t}\}, \qquad (3.8)$$

au minimum relativement à λ. Suivant les hypothèses (3.3), (3.4) et (3.5), la valeur optimale de λ pour le niveau courant est

$$\lambda_{\text{opt}} = [g^{-1}\,\theta^{-1}(1-\rho^2) + q_t^{-1}\rho^2 + 1]^{-1}.$$

Toutefois, si l'on compte utiliser l'estimateur longtemps, il faut bien comprendre que seules certaines valeurs de q_t sont possibles à la longue. La valeur de λ choisie pour l'estimation de μ_t détermine la variance de $\hat{\mu}_t$ et, par conséquent, elle détermine la variance qui sera intégrée à l'estimateur de μ_{t+1}. Si nous supposons que $\beta = \rho$, nous avons

$$V\{\hat{\mu}_t\} = \left\{g^{-1}\,\theta^{-1}\lambda^2(1-\rho^2) + q_t^{-1}\lambda^2\rho^2 + (1-\lambda)^2\right\} V\{\bar{y}_{B,t}\}$$

ou

$$q_{t+1}^{-1} = g^{-1}\,\theta^{-1}\lambda^2(1-\rho^2) + (1-\lambda)^2 + \lambda^2\rho^2 q_t^{-1}. \qquad (3.9)$$

Ainsi, pour une valeur donnée de λ, la valeur extrême pour q_t^{-1} est

$$\lim_{t\to\infty} q_t^{-1} = (1-\lambda^2\rho^2)^{-1}[g^{-1}\,\theta^{-1}\lambda^2(1-\rho^2) + (1-\lambda)^2]. \qquad (3.10)$$

Ce résultat correspond à celui donné par Cochran (1977), page 352, équation (12.86).

Le tableau 1 contient des valeurs des variances extrêmes à mesure que le nombre de périodes devient grand, pour un choix de valeurs de ρ et de λ, où $\theta = 5/6$ et $g\theta = 5$ pour l'EPA. Les variances sont normalisées de façon que la variance de l'estimateur direct fondé sur la moyenne de six panels soit de 1,00. Ainsi, les entrées sont de six fois la valeur extrême en (3.10). Si la corrélation est de 0,95 et si la valeur de λ est posée égale à 0,96, la variance à long terme du niveau courant est de 70 % de celle de l'estimateur direct. Si la valeur de λ est posée égale à 0,90, la variance à long terme est de 58 % de celle de l'estimateur direct lorsque $\rho = 0{,}95$.

La première ligne du tableau 1 est réservée pour $\lambda = 5/6 = \theta$. C'est là la valeur de λ qui correspond à l'utilisation de x_{2t} dans un estimateur de régression. La variance comportant $\lambda = 5/6$ est toujours plus faible que celle de l'estimateur direct à cause de l'amélioration associée à l'utilisation de l'estimateur de régression $\hat{\mu}_{m,t}$. Si $\rho \neq 0$, l'estimateur de régression comportant x_{2t} entraîne une réduction significative de la variance relativement à l'estimateur direct, $\bar{y}_t$, qui utilise uniquement des données courantes.

Tableau 1
Variances de niveau extrêmes normalisées : Schéma de renouvellement de l'EPA

	ρ				
λ	0,70	0,80	0,90	0,95	0,98
0,833	0,897	0,840	0,743	0,665	0,600
0,840	0,895	0,836	0,340	0,650	0,581
0,860	0,894	0,830	0,714	0,614	0,527
0,880	0,903	0,835	0,705	0,588	0,481
0,900	0,921	0,851	0,711	0,575	0,444
0,920	0,951	0,882	0,736	0,582	0,420
0,940	0,992	0,928	0,785	0,617	0,420
0,960	1,046	0,994	0,867	0,698	0,465
0,980	1,115	1,083	0,997	0,861	0,619
0,990	1,155	1,138	1,087	0,998	0,803
0,995	1,177	1,168	1,140	1,089	0,960

La valeur optimale de λ est une fonction de ρ et elle augmente lentement à mesure que ρ augmente. Pour $\rho = 0{,}0$, la valeur optimale de λ est de 0,833; pour $\rho = 0{,}7$ la valeur optimale de λ est de 0,85 environ, pour $\rho = 0{,}95$ la valeur optimale de λ est de 0,91 environ et pour $\rho = 0{,}98$ la valeur optimale de λ est de 0,93 environ.

Nous considérons maintenant l'estimateur MR2 (2.3) qui s'écrit sous la forme

$$\hat{\mu}_t = \lambda_A[\bar{y}_{m,t} + (\hat{\mu}_{t-1} - \bar{y}_{m,t-1})b^*] + (1-\lambda_A)\bar{y}_{B,t},$$

où λ_A et b^* sont définis en (2.6). Même si l'estimateur MR2 n'est pas un membre de la classe (3.7), dans la mesure où b^* est « proche » de ρ, il est « proche » d'un membre de la classe. Par exemple, si $\rho = 0{,}95$, on a $b_0 \doteq 0{,}9314$ et $b^* \doteq 0{,}9422$. Si $\rho = 0{,}.90$, on a $b_0 \doteq 0{,}8659$ et $b^* \doteq 0{,}8853$.

Si nous utilisons la valeur extrême b_0 de b, nous avons $(1-\lambda_A) = (1-\theta)(1-b_0)$, où b_0 est donné par (2.4). Dès lors, $\lambda_A = 0{,}9375$, 0,9568, 0,9776, 0,9886 et 0,9954 pour $\rho = 0{,}70$, 0,80, 0,90, 0,95 et 0,98, respectivement. D'après le tableau 1, les variances normalisées de $\hat{\mu}_t$ pour ces valeurs de λ_A sont de 0,986, 0,982, 0,978, 0,976 et 0,975, pour $\rho = 0{,}70$, 0,80, 0,90, 0,95 et 0,98, respectivement. Ainsi, l'estimateur MR2 pour le niveau courant offre une efficacité de niveau qui est essentiellement la même que pour l'estimateur direct, $\bar{y}_t$. L'efficacité de l'estimateur MR1 est celle pour $\lambda = 0{,}833$ dans le tableau 1 et toujours supérieure à celle de $\bar{y}_t$.

4. VARIANCE D'UN CHANGEMENT D'UNE PÉRIODE

Étant donné $\hat{\mu}_{t-1}$, $\bar{y}_{m,t-1}$, $\bar{y}_{m,t}$ et $\bar{y}_{B,t}$, l'estimateur optimal de μ_{t-1} n'est plus $\hat{\mu}_{t-1}$, parce que $\bar{y}_{m,t}$ comporte des renseignements au sujet de μ_{t-1}. Toutefois, on n'a pas

l'habitude de réviser l'estimateur de μ_{t-1}. En l'absence de révision, l'estimateur de changement est $\hat{\mu}_t - \hat{\mu}_{t-1}$.

En l'absence de révision de $\hat{\mu}_{t-1}$ et compte tenu des conditions (3.2) à (3.5), la variance de $\hat{\mu}_t - \hat{\mu}_{t-1}$, où $\hat{\mu}_t$ est défini en (3.7), est la suivante

$$\begin{aligned} V\{\hat{\mu}_t - \hat{\mu}_{t-1}\} &= V\{\lambda[\bar{y}_t + (\hat{\mu}_{t-1} - \bar{x}_{2,t})\rho] \\ &\quad + (1-\lambda)\bar{y}_{B,t} - \hat{\mu}_{t-1}\} \\ &= [g^{-1}\theta^{-1}\lambda^2(1-\rho^2) + (1-\lambda)^2 \\ &\quad + (\rho\lambda - 1)^2 q_t^{-1}]V\{\bar{y}_{B,t}\}. \end{aligned} \tag{4.1}$$

Le tableau 2 contient des variances extrêmes normalisées du changement estimatif, $\hat{\mu}_t - \hat{\mu}_{t-1}$, pour un choix de valeurs de g et de λ, avec $g\theta = 5$, Les entrées du tableau sont les variances extrêmes du changement estimatif divisées par la variance du changement d'après les éléments communs, $V\{\bar{y}_{m,t} - \bar{y}_{m,t-1}\}$. La variance du changement fondée sur les éléments communs est de $2\theta^{-1}(1-\theta)(1-\rho)\,V\{\bar{y}_{B,t}\}$.

Tableau 2

Variances extrêmes normalisées en cas de changement d'une période sans révision : Schéma de renouvellement de l'EPA

			ρ		
λ	0,70	0,80	0,90	0,95	0,98
0,833	1,039	1,168	1,550	2,312	4,595
0,840	1,024	1,142	1,492	2,189	4,277
0,860	0,989	1,079	1,345	1,872	3,454
0,880	0,963	1,029	1,223	1,607	2,756
0,900	0,947	0,993	1,127	1,391	2,181
0,920	0,940	0,970	1,055	1,222	1,723
0,940	0,942	0,959	1,007	1,100	1,379
0,960	0,953	0,961	0,982	1,024	1,146
0,980	0,972	0,975	0,980	0,991	1,021
0,990	0,985	0,986	0,987	0,990	0,998
0,995	0,992	0,993	0,993	0,994	0,996

Les tableaux 1 et 2 indiquent clairement le risque de ne pas réviser l'estimation de $\hat{\mu}_{t-1}$. Par exemple, si $\rho = 0{,}95$, la variance d'un changement sans révision d'une période est réduite au minimum avec $\lambda \doteq 0{,}99$, mais la variance de niveau est réduite au minimum avec $\lambda \doteq 0{,}91$. Une valeur compromis de $\lambda = 0{,}95$ donne une variance de niveau qui est supérieure de 14 % environ à la valeur optimale et une variance de changement qui est supérieure de 7 % environ à la plus petite variance du tableau 2.

Si nous utilisons les valeurs de λ_A associées à l'estimateur MR2, les entrées du tableau 2 sont de 0,940, 0,960, 0,979, 0,989 et 0,996 pour $\rho = 0{,}70$, 0,80, 0,90, 0,95 et 0,98, respectivement. Ainsi, en l'absence de révision, et si l'on laisse de côté la différence entre b_0 et ρ, l'estimateur MR2 est presque optimal comme estimateur de changement, contrairement à l'estimateur MR1, où l'estimateur MR1 correspond à $\lambda = 0{,}833$ du tableau 2.

5. UN ESTIMATEUR COMPROMIS

D'après le tableau 2, l'efficacité de l'estimateur de changement MR2 pour l'EPA fondé sur x_{1t}, pour le cas de non-révision, est assez bonne. L'estimateur de changement sans révision MR1 fondé sur x_{2t} offre une efficacité relativement médiocre puisqu'il est membre de la classe (3.7) avec $\lambda = \theta = 0{,}8333$. Par contre, l'estimateur de niveau MR1 fondé sur x_{2t} est supérieur à l'estimateur MR2 fondé sur x_{1t}, et il y a des membres de la classe (3.7) qui sont nettement supérieurs à l'estimateur de niveau MR2.

Puisque la valeur de λ dans l'estimateur MR2 est relativement grande et puisque la valeur de λ pour l'estimateur MR1 est relativement faible, nous pouvons créer des approximations des membres les plus intéressants de la classe (3.7) comme combinaisons linéaires de (2.10) et de (2.5). Notons

$$x_{3,ti} = \alpha x_{1,ti} + (1-\alpha)x_{2,ti}, \tag{5.1}$$

où $0 \le \alpha \le 1$ est un nombre fixe. L'estimateur de régression fondé sur x_{3ti} donne une approximation d'un membre de la classe (3.7) avec

$$\lambda = \alpha\lambda_A + (1-\alpha)\theta, \tag{5.2}$$

où λ_A est défini en (2.6). Si $\rho = 0{,}95$,

$$\lambda = \alpha(0.9886) + (1-\alpha)(5/6),$$

pour le schéma de renouvellement de l'EPA avec $\theta = 5/6$ et $b_0 = (7-2\rho)^{-1}5\rho$; $\lambda = 0{,}95$ si $\alpha = 0{,}75$.

Nous choisissons α pour avoir la combinaison souhaitée de $\bar{y}_{B,t}$ et de l'« estimateur de régression » fondé sur des observations dans l'ensemble A. Si l'on ne révise pas l'estimateur de μ_{t-1}, la combinaison préférée dépend de l'importance relative accordée à la variance de niveau et à la variance de changement.

Le tableau 3 indique la variance de l'estimateur MR2 ($\alpha = 1$) relativement à la variance de l'estimateur construit à l'aide de $\alpha = 0{,}75$ et à la variance de l'estimateur construit à l'aide de $\alpha = 0{,}65$. Une entrée du tableau 3 pour $\hat{\mu}_t$ est l'expression (3.10) évaluée à λ_A de (2.6) et ρ, divisée par (3.10) évaluée à λ de (5.2) et ρ. Une entrée pour $\hat{\mu}_t - \hat{\mu}_{t-1}$ est l'expression (4.1) évaluée à λ_A de (2.6) et ρ, divisée par (4.1) évaluée à λ de (5.2) et ρ. Ce sont là des approximations d'efficacités réelles car on utilise ρ pour le coefficient de x_3. Le tableau 3 indique clairement que l'estimateur compromis est légèrement inférieur à l'estimateur MR2 pour un changement d'une période, mais qu'il est bien supérieur à l'estimateur MR2 pour le niveau. Par exemple, avec $\rho = 0{,}95$ et $\alpha = 0{,}65$, l'efficacité relative de l'estimateur compromis est de 1,62 pour le niveau et de 0,87 pour un changement d'une période.

Pour de plus grandes valeurs de ρ, la variance de changement est bien plus petite que la variance de niveau. Ainsi, pour $\rho = 0{,}95$, la variance de niveau et la variance de changement pour $\alpha = 1{,}00$ sont de 1,00 et de 0,12 environ,

respectivement, tandis que la variance de niveau et la variance de changement pour $\alpha = 0{,}75$ sont de 0,67 et de 0,13 environ, respectivement, exprimées en unités communes.

La valeur α plus petite offre l'avantage que l'estimateur composite est plus proche de l'estimateur direct. Ainsi, le biais éventuel associé à l'estimateur composite devrait être plus petit pour une valeur α plus petite.

Tableau 3
Efficacités approximatives d'estimateurs compromis relativement à $\alpha = 1$

			$\alpha = 0{,}75$		$\alpha = 0{,}65$	
ρ	b_0	$1-\lambda_A$	$\hat{\mu}_t$	$\hat{\mu}_t - \hat{\mu}_{t-1}$	$\hat{\mu}_t$	$\hat{\mu}_t - \hat{\mu}_{t-1}$
0,70	0,625	0,0625	1,052	0,999	1,069	0,995
0,80	0,741	0,0432	1,099	0,994	1,129	0,984
0,90	0,865	0,0224	1,238	0,975	1,303	0,946
0,95	0,931	0,0114	1,502	0,936	1,616	0,875
0,98	0,972	0,0046	2,177	0,833	2,321	0,712

6. PROBLÈME DE DÉRIVE

Comme il a été noté dans l'introduction, l'estimateur MR2 pourrait s'écarter de l'estimateur direct de façon appréciable, et cet écart pourrait s'étendre sur une longue période. Nous allons maintenant expliquer ce phénomène. Nous pouvons exprimer l'écart de l'estimateur de régression compromis $\hat{\mu}_t$, en fonction de x_{3ti}, vis-à-vis de la moyenne réelle μ_t sous la forme

$$\hat{\mu}_t - \mu_t = (\lambda\rho)^t(\hat{\mu}_0 - \mu_0) + \sum_{j=0}^{t-1} (\lambda\rho)^j$$

$$[\lambda \bar{r}_{m,t-j} + (1-\lambda)(\bar{y}_{B,t-j} - \mu_{t-j})], \qquad (6.1)$$

où μ_0 est la moyenne au début du processus et

$$\bar{r}_{mt} = \bar{y}_{mt} - \mu_t - \rho(\bar{y}_{m,t-1} - \mu_{t-1}).$$

Si ρ est proche de un et si nous utilisons $\lambda = 1$, l'erreur $\hat{\mu}_t - \mu_t$ se comporte plus ou moins comme une marche aléatoire pouvant entraîner de longues périodes au cours desquelles $\hat{\mu}_t - \mu_t$ porte le même signe. Par contre, si $\alpha < 1$ et si $\rho = 1, \lambda < 1$ et l'erreur $\hat{\mu}_t - \mu_t$ manifeste moins de dérive. Par exemple, si $\alpha = 0{,}70$, la corrélation entre des erreurs adjacentes $\hat{\mu}_t - \mu_t$ ne sera pas supérieure à 0,95 suivant l'hypothèse (3.2)-(3.5). Pour l'estimateur MR2, $\lambda \to 1$ à mesure que $\rho \to 1$ et, par conséquent, l'estimateur MR2 peut manifester une dérive pour ρ proche de un.

7. CONCLUSION

Par souci de simplicité, nous avons souvent supposé un échantillonnage aléatoire simple afin d'obtenir des résultats théoriques. Des résultats semblables sont valables pour des plans complexes et des variables auxiliaires supplémentaires, si l'on considère p comme une autocorrélation partielle. De plus, nous avons utilisé des variables x_3 correspondant à une variable y seulement, mais plusieurs variables y peuvent servir à construire les variables x correspondantes en vue de l'estimation de régression. Gambino, Kennedy et Singh (2001) ont mené une étude empirique avec des données de l'EPA en faisant appel à plusieurs variables x_3 comportant un α commun, et ont obtenu un α compromis en vue de l'EPA.

À la section 2.1, nous avons supposé l'absence de non-réponse afin que l'imputation ne soit pas requise. Dans l'EPA, toutefois, il peut y avoir une non-réponse pour un élément y soit au moment $t-1$, soit au moment t ou aux deux à la fois. Gambino, Kennedy et Singh (2001) décrivent en détail les méthodes d'imputation utilisées concrètement dans l'EPA.

REMERCIEMENTS

Les recherches de Wayne Fuller ont été appuyées en partie par un accord de coopération (43-3AEU-3-80088) conclu entre l'Iowa State University, le National Agricultural Statistics Service et le Bureau of the Census des États-Unis. Nous remercions Harold Mantel d'avoir lu le manuscrit soigneusement, car ainsi des améliorations ont pu être apportées.

BIBLIOGRAPHIE

BELL, P. (2001). Comparaison d'autres estimateurs pour l'Enquête sur la population active. *Techniques d'enquête*, 27, 57-68.

BELL, W.R, et HILLMER, S.C. (1990). Estimation dans les enquêtes à passages répétés au moyen de séries chronologiques. *Techniques d'enquête*, 16, 205-227.

BINDER, D.A., et DICK, J.P. (1989). Enquêtes répétées – Modélisation et estimation. *Techniques d'enquête*, 15, 31-48.

BREAU, P., et ERNST, L. (1983). Alternative estimators to the current composite estimator. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 397-402.

COCHRAN, W.G. (1977). *Sampling Techniques*. 3e édition. New York : John Wiley and Sons.

ECKLER, A.R. (1955). Rotation sampling. *Annals of Mathematical Statistics*, 26, 664-180.

FULLER, W.A. (1990). Analyse d'enquêtes à passages répétés. *Techniques d'enquête*, 16, 177-190.

GAMBINO, J.G., KENNEDY, B. et SINGH, M.P. (2001). Estimation composite par régression pour l'Enquête sur la population active du Canada : Évaluation et application. *Techniques d'enquête*, 27, 69-79.

GAMBINO, J.G., SINGH, M.P., DUFOUR, J., KENNEDY, B. et LINDEYER, J. (1998). *Méthodologie de l'Enquête sur la population active du Canada*. Numéro de catalogue 71-526, Statistique Canada.

GURNEY, M., et DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the American Statistical Association, Section on Social Statistics*, 242-257.

HANSEN, M.H., HURWITZ, W.N., NISSELSON H. et STEINBERG, J. (1955). The redesign of the census current population survey. *Journal of the American Statistical Association*, 50, 701-719.

JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin*, 304, 54-59.

JONES, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, B*, 42, 221-226.

KUMAR, S., et LEE, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. *Techniques d'enquête*, 9, 1-28.

LENT, J., MILLER, S.M., CANTWELL, P.J. et DUFF, M. (1999). Effect of composite weights on some estimates from the Current Population Survey. *Journal of Official Statistics*, 14, 431-448.

PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, B*, 12, 241-255.

PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 9, 163-175.

RAO, J.N.K., et GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.

SCOTT, A.J., SMITH, T.M.F. et JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *Revue Internationale de Statistique*, 45, 13-28.

SINGH, A.C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 120-129.

SINGH, M.P., DREW, J.D., GAMBINO, J. et MAYDA, F. (1990). *Méthodologie de l'Enquête sur la population active du Canada*. Numéro de catalogue 71-526, Statistique Canada.

TILLER, R. (1989). A Kalman filter approach to labor force estimation using survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 16-25.

WOLTER, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

YANSANEH, I.S., et FULLER, W.A. (1998). Méthode optimale d'estimation récursive pour les enquêtes répétitives. *Techniques d'enquête*, 24, 33-42.

Techniques d'enquête, juin 2001
Vol. 27, n° 1, pp. 57-68
Statistique Canada

Comparaison d'autres estimateurs pour l'Enquête sur la population active

PHILIP BELL[1]

RÉSUMÉ

L'auteur examine un choix d'estimateurs applicables à une enquête auprès des ménages périodique comportant un chevauchement contrôlé entre les enquêtes successives. L'auteur décrit la façon d'améliorer le MELSB (meilleur estimateur linéaire sans biais) fondé sur une fenêtre fixe de moments en appliquant la technique de la régression généralisée. Cet estimateur amélioré est comparé à l'estimateur *AK* de Gurney et Daly (1965) et à l'estimateur de régression modifiée de Singh, Kennedy, Wu et Brisebois (1997), à l'aide de données de l'enquête sur la population active de l'Australie.

MOTS CLÉS : Estimateur composite; meilleur estimateur linéaire sans biais; régression modifiée; enquêtes répétées.

1. INTRODUCTION

L'auteur examine un choix d'estimateurs applicables à une enquête-ménage périodique comportant un chevauchement contrôlé entre les enquêtes successives. Le thème commun des estimateurs est le recours à des données de moments antérieurs afin d'améliorer les estimations courantes, en bénéficiant de corrélations dans l'échantillon chevauchant. L'auteur considère tous les estimateurs de ce genre comme des estimateurs composites.

Les estimateurs sont évalués en fonction de l'enquête sur la population active (EPA) de l'Australie. Dans l'EPA, on contrôle le chevauchement en divisant l'échantillon du premier degré (régions géographiques) en huit « groupes de renouvellement » parmi lesquels on sélectionne des logements. Chaque mois, les mêmes logements sont tirés de sept des groupes de renouvellement, de nouveaux logements étant sélectionnés à même le groupe restant. L'échantillon est constitué de personnes âgées de 15 ans ou plus demeurant dans les logements sélectionnés.

Ce plan d'échantillonnage entraîne un chevauchement élevé de l'échantillon entre deux mois consécutifs au sein des sept « groupes de renouvellement appariés ». Le fait d'utiliser uniquement des données de ces groupes de renouvellement plutôt que de l'échantillon tout entier permet de diminuer l'erreur d'échantillonnage pour une estimation du mouvement d'un mois à l'autre. À l'aide de techniques d'estimation composite, on peut exploiter cette situation de façon à préparer des estimations comportant une erreur d'échantillonnage plus faible.

À la section 2, l'auteur présente l'EPA de l'Australie et son estimateur courant de « régression généralisée ». L'auteur aborde également la question du biais d'accoutumance (appelé biais de groupe de renouvellement par Bailar 1975).

À la section 3, l'auteur présente l'estimateur « composite *AK* » proposé par Gurney et Daly (1965). Cette méthode est utilisée dans la Current Population Survey des États-Unis depuis de nombreuses années. Un prolongement appelé « pondération composite *AK* » est utilisé depuis quelques années; il a été proposé par Fuller (1990) et étudié par Lent, Miller et Cantwell (1994, 1996).

À la section 4, l'auteur présente la méthode « de régression modifiée » de l'estimation composite (Singh et Merkouris 1995; Singh 1996). Dans le présent exposé, l'auteur examine l'estimateur MR2 de Singh et coll. (1997), qui permet la plus forte réduction de l'erreur d'échantillonnage. L'auteur présente également une variante de cette méthode suggérée par Fuller (1999) en vue de l'Enquête sur la population active du Canada.

À la section 5, l'auteur présente un « meilleur estimateur linéaire sans biais » (MELSB) fondé sur des données tirées d'une « fenêtre » comportant un nombre fixe de mois successifs. Cet estimateur a d'abord été présenté par Jessen (1942) dans le cas de 2 occasions. Un MELSB fondé sur toutes les occasions d'une longue série semble peu pratique, bien que Yansaneh et Fuller (1998) en aient élaboré une approximation récursive. L'auteur améliore le MELSB à fenêtre fixe décrit par Bell (1998) à l'aide de la technique de régression généralisée.

À la section 6, on trouve les résultats de l'application des différentes méthodes à l'estimation des personnes ayant un emploi et des personnes sans emploi dans l'EPA. Les erreurs-types sont estimées pour des indicateurs à plus long terme comme la tendance et le mouvement de la tendance, de même que pour des estimations du niveau mensuel et de son mouvement. L'auteur examine les biais éventuels, de même que les indications d'un changement de profil saisonnier.

L'auteur termine l'exposé en comparant les avantages et le inconvénients des différents types d'estimateurs en fonction de l'EPA. L'estimateur MELSB amélioré est jugé efficace et, lorsqu'il est appliqué à l'EPA, il n'est pas exposé à un biais appréciable.

[1] Philip Bell, Bureau of Statistics de l'Australie,courriel : philip.bell@abs.gov.au

2. ESTIMATEURS COURANTS POUR L'ENQUÊTE SUR LA POPULATION ACTIVE

2.1 Aperçu de l'EPA

L'EPA comporte un plan d'échantillonnage à plusieurs degrés, le premier étant un échantillon de petites régions géographiques appelées CD (« Census collector's Districts ». On tire un nouvel échantillon de CD tous les cinq ans, et les CD sont classés dans huit « groupes de renouvellement ». Les logements tirés d'un CD restent dans l'échantillon durant huit enquêtes, puis ils sont remplacés par d'autres logements tirés du même CD. On appelle renouvellement ce remplacement des logements, tous les logements d'un groupe de renouvellement étant remplacés en même temps. Des intervieweurs cherchent à recueillir des données pour toutes les personnes admissibles dans les logements sélectionnés.

Dans l'EPA, la situation d'activité de la personne (ayant un emploi, sans emploi, ne relevant pas de la population active) revêt un intérêt particulier. Le nombre de personnes se rapportant à chaque situation d'activité, pour diverses catégories de personnes, est un élément clé qu'il s'agit d'estimer dans l'enquête. Les nombreux utilisateurs des données d'enquête accordent une importance encore plus grande aux estimations du mouvement des chiffres d'un moment au suivant. On peut affirmer que les indications à plus long terme de la direction de la série sont encore plus importantes (par exemple le mouvement de la tendance X11 ou d'une tendance semblable plus lisse (Bell 1999)).

Le plan d'échantillonnage garantit que la probabilité inconditionnelle de sélection $\pi_{t,i}$ est connue pour chaque personne échantillonnée i à un moment t. Cela permet d'obtenir un estimateur simple pour un total de population dû à Horvitz et Thompson (1952). Si Y_t est l'élément de population qu'il s'agit d'estimer à un moment t, et si y_{ti} est le même élément signalé par la i^e unité à un moment t, l'estimateur de Horvitz-Thompson est

$$\hat{y}_t^{\mathrm{H}} = \sum_i w_{ti}^{\pi} y_{ti} \tag{1}$$

pour $w_{ti}^{\pi} = \pi_{ti}^{-1}$, connus comme les poids de sélection.

2.2 L'estimateur de régression généralisée (RG)

La régression généralisée est une méthode de rajustement ou de « calage » d'une série de poids unitaires dont l'addition donne une série d'attributs de la population appelés repères. Pour un choix approprié de repères, les poids résultants donnent une meilleure estimation en tenant compte de renseignements externes.

Dans l'EPA, nous commençons par un calage des poids de Horvitz-Thompson de façon que leur addition donne des repères démographiques indiquant le nombre de personnes dans la population pour 560 strates après sélection (14 régions géographiques classées selon le sexe et 20 groupes d'âge). Les poids d'une strate après sélection particulière sont répartis proportionnellement de façon que leur addition donne le repère de strate. Cet estimateur par quotient stratifié après sélection est un cas particulier de l'estimateur de régression généralisée ou estimateur RG.

Soit x_{ti} un vecteur de ligne de variables auxiliaires pour l'unité i à un moment t, et $\hat{x}_t = \Sigma_i b_{ti} x_{ti}$ des estimations du vecteur de ligne correspondant de valeurs repères X_t, d'après certains poids initiaux b_{ti}. L'estimateur RG fondé sur ces poids initiaux est alors donné par

$$\hat{y}_t^{G} = \hat{y}_t + (X_t - \hat{x}_t)\hat{\beta} \tag{2}$$

$$\text{pour } \hat{\beta} = \left(\sum_i b_{ti} x_{ti}' x_{ti}\right)^{-1} \sum_i b_{ti} x_{ti}' y_{ti}. \tag{3}$$

Donc, $\hat{y}_t^{G} = \sum_i w_{ti}^{G} y_{ti}$ pour

$$w_{ti}^{G} = b_{ti}\left(1 + (X_t - \hat{x}_t)\left(\sum_i b_{ti} x_{ti}' x_{ti}\right)^{-1} x_{ti}'\right) \tag{4}$$

Dans une estimation par quotient stratifiée après sélection, les vecteurs de ligne x_{ti} contiennent des zéros sauf dans la colonne qui correspond à la strate après sélection de l'unité, et b_{ti} sont les poids de sélection w_{ti}^{π}. Dans ce cas, les paramètres de régression sont simplement les moyennes de strates après sélection, estimées à l'aide des poids de sélection.

2.3 Estimations par groupe de renouvellement

Chaque groupe de renouvellement est constitué d'un échantillon représentatif de logements, et peut donc fournir une estimation distincte. On numérote les groupes de renouvellement à tel moment selon le nombre de fois que les logements du groupe de renouvellement ont été échantillonnés. On écrit R(t, i) = r si l'unité i se trouve dans le groupe de renouvellement échantillonné pour la r-^{e} fois à un moment t. L'estimation de Horvitz-Thompson de Y_t d'après le groupe de renouvellement r est

$$\hat{y}_t^{\mathrm{H}r} = \sum_{i:R(t,i)=r} 8 w_{ti}^{\pi} y_{ti}. \tag{5}$$

On peut utiliser la régression généralisée pour améliorer ces estimateurs en calant les poids de façon que leur addition donne une série de repères. Malheureusement, la taille plus faible de l'échantillon dans un groupe de renouvellements unique risque d'exiger l'utilisation d'un plus petit nombre de repères que pour le cas global. Dans la situation de l'EPA, l'auteur du présent exposé a appliqué une même étape de régression généralisée à l'ensemble de l'échantillon de façon que, pour tout l'échantillon, l'addition des poids donne les repères pour les 540 strates après sélection courantes, tandis que, dans chaque groupe de renouvellement, l'addition des poids donne un huitième des repères pour 71 strates après sélection groupées. Les poids résultants, lorsqu'ils sont appliqués à un groupe de renouvellements donné r et multipliés par huit, donnent les estimations de groupe de renouvellement $\hat{y}_t^{\mathrm{R}r}$.

2.4 Biais d'accoutumance

Idéalement, les estimations de groupe de renouvellement devraient comporter la même espérance Y_t, mais dans la pratique elles comportent des espérances légèrement différentes et donc un biais différent. La différence est due en partie à des pratiques de collecte; par exemple, les logements échantillonnés pour la première fois sont interviewés dans le cadre d'une visite sur place, tandis que d'autres groupes de renouvellement sont surtout interviewés par téléphone. Il n'est pas clair quel groupe de renouvellement est influencé le moins par ce genre de biais « d'accoutumance ». L'estimation globale comporte un biais d'accoutumance qui est un mélange des biais de chaque groupe de renouvellement. Nous avons recours à de bonnes techniques d'enquête pour réduire ce biais au minimum. À noter que tous les estimateurs composites recevront un apport différent des groupes de renouvellement, et comporteront donc différents biais d'accoutumance.

3. ESTIMATION COMPOSITE *AK*

3.1 Estimateur composite *AK*

L'estimateur composite *AK* (Gurney et Daly 1965) est conçu de façon à insister davantage sur le mouvement des groupes de renouvellement appariés (c'est-à-dire les groupes de renouvellement dans lesquels les mêmes logements ont été sélectionnés pour le mois courant et le mois précédent). L'estimateur compte trois composantes. La première est une moyenne des estimations de groupe de renouvellement pour les données du mois courant (moment t). La deuxième est l'estimateur composite *AK* du mois précédent, plus une estimation de mouvement fondée uniquement sur les groupes de renouvellement appariés. La troisième composante est la différence entre les estimations du groupe de renouvellement non apparié et des groupes appariés. La part de chaque composante qu'il convient d'utiliser est donnée par deux paramètres A et K, comme suit :

$$\hat{y}_t^{AK} = (1 - K)\frac{1}{8}\sum_{r=1}^{8} \hat{y}_t^{Rr} + K\left(\hat{y}_{t-1}^{AK} + \frac{1}{7}\sum_{r=2}^{8} \hat{y}_t^{Rr} - \frac{1}{7}\sum_{r=1}^{7} \hat{y}_{t-1}^{Rr}\right) + A\left(\hat{y}_t^{R1} - \frac{1}{7}\sum_{r=2}^{8} \hat{y}_t^{Rr}\right). \quad (6)$$

3.2 Choix des valeurs paramétriques

Le paramètre clé est K, qui indique la part de l'estimation nouvelle qui se fonde sur le mouvement des groupes de renouvellement appariés. La valeur optimale de A et de K à utiliser dépend de la variable qu'il s'agit d'estimer. Des valeurs K plus élevées se prêtent davantage à l'emploi qu'au chômage, puisque l'emploi comporte une plus forte corrélation d'un mois à l'autre .

L'addition des estimations composites *AK* des personnes ayant un emploi, sans emploi et « ne relevant pas de la population active » ne donne pas exactement la population totale à moins que l'on n'utilise les mêmes paramètres pour toutes les estimations. On choisit donc un compromis pour A et K. Les résultats présentés ici se fondent sur $A = 0{,}06$ et $K = 0{,}7$. On a trouvé ces valeurs en faisant l'essai d'une gamme de valeurs de A et de K et en choisissant les valeurs qui donnent des estimations optimales des personnes ayant un emploi. Dans la présente étude, aucune valeur de A et de K n'a donné des estimations des personnes sans emploi appréciablement meilleures que ces valeurs.

Notre étude empirique n'a pas manifesté un niveau particulièrement bon d'erreurs d'échantillonnage pour l'estimateur *AK*. Cela est peut-être attribuable au calage fin qui a servi à obtenir les estimations de groupe de renouvellement; il est possible que le recours à des catégories plus vastes améliore le bilan des erreurs d'échantillonnage.

3.3 Propriétés de l'estimateur *AK*

L'estimateur *AK* insiste davantage sur le mouvement au sein des groupes de renouvellement appariés. Ainsi, le groupe de renouvellement qui contient des logements admissibles pour la première fois contribue moins que dans l'estimateur RG. L'estimateur *AK* comporte donc un biais d'accoutumance différent de celui de l'estimateur RG.

L'estimateur *AK* est récursif : il faut l'estimateur du mois précédent pour préparer celui du mois courant. Cela est peu pratique lorsqu'il s'agit de préparer des estimations pour un nouvel élément ou une nouvelle catégorie. De plus, le besoin d'utiliser les mêmes valeurs de A et de K pour tous les éléments peut entraîner des estimations sous-optimales pour un élément donné.

À cause de ces préoccupations, la Current Population Survey des États-Unis a été changée en fonction d'une variante appelée « pondération composite *AK* » (Lent, Miller et Cantwell 1994). Dans la pondération composite *AK*, on prépare des estimations distinctes de l'emploi et du chômage pour un certain nombre de catégories publiées importantes, à l'aide de la stratégie composite *AK* avec des paramètres optimaux pour l'estimation en question. Les données courantes sont alors calées de façon que l'addition des poids unitaires donne ces estimations *AK* de même que des repères démographiques. Toutes les estimations sont alors préparées à partir du fichier de données courant à l'aide de ces nouveaux « poids composites *AK* ».

Le fait de pouvoir préparer toutes les estimations sous forme de somme pondérée des données d'un même mois est un grand avantage de la stratégie de pondération composite *AK*. Un autre avantage est que les plus importantes estimations sont des estimations composites *AK* avec un choix de *AK* presque optimal. Un inconvénient est que seules les plus importantes estimations sont de véritables estimations composites. Toutes les autres estimations (y

compris les estimations de personnes ne relevant pas de la population active) ne sont typiquement pas très améliorées relativement aux estimations RG standard (Lent, Miller et Cantwell 1996).

4. ESTIMATION DE RÉGRESSION MODIFIÉE

4.1 Aperçu de la régression modifiée

La méthode de régression modifiée est une autre façon de fournir des estimations composites que l'on peut obtenir comme agrégats pondérés du fichier de données d'enquête courant. La méthode s'applique à un ensemble d'éléments clés déterminé à l'avance, pour lequel elle réalise un nombre particulièrement faible d'erreurs d'échantillonnage.

La technique de régression modifiée fait appel à une régression généralisée pour le fichier de données du mois courant après avoir annexé de nouvelles variables auxiliaires z_{ti} à chaque unité i à un moment t. Ici z_{ti} est un vecteur de ligne comportant un élément pour chacun des éléments clés. Pour ceux-ci, on a des « pseudo-repères » Z_t fondés sur les estimations du mois précédent pour les éléments clés. L'estimateur de régression modifiée est alors donné par une étape de régression généralisée qui applique les repères démographiques aussi bien que les pseudo-repères.

$$\hat{y}_t^{\mathrm{M}} = \hat{y}_t^{\mathrm{H}} + \left((X_t, Z_t) - (\hat{x}_t^{\mathrm{H}}, \hat{z}_t^{\mathrm{H}})\right)\beta_t^{\mathrm{M}} \tag{7}$$

pour

$$\beta_t^{\mathrm{M}} = \left(\sum_i w_{ti}^{\pi}(x_{ti}, z_{ti})'(x_{ti}, z_{ti})\right)^{-1} \sum_i w_{ti}^{\pi}(x_{ti}, z_{ti})' y_{ti} \tag{8}$$

donc, $\hat{y}_t^{\mathrm{M}} = \sum_i w_{ti}^{\mathrm{M}} y_{ti}$ pour

$$w_{ti}^{\mathrm{M}} = w_{ti}^{\pi}\left\{1 + \left((X_t, Z_t) - (\hat{x}_t^{\mathrm{H}}, \hat{z}_t^{\mathrm{H}})\right)\right.$$
$$\left.\left(\sum_i w_{ti}^{\pi}(x_{ti}, z_{ti})'(x_{ti}, z_{ti})\right)^{-1}(x_{ti}, z_{ti})'\right\}. \tag{9}$$

La clé de la méthode est la définition des variables auxiliaires. Soit D, l'ensemble des unités des groupes de renouvellement appariés (ceux comportant des logements sélectionnés aux deux moments) à un moment t. Soit y_{ti}^*, le vecteur d'éléments clés pour l'unité i à un moment t et Y_t^*, les totaux de population correspondants. Pour $i \in$ D, notons $y_{t-1,i}^*$, la valeur du mois précédent pour le vecteur d'éléments clés ou, si aucune valeur n'a été déclarée, imputons $y_{t-1,i}^*$ – j'ai utilisé $y_{t-1,i}^* = y_{t,i}^*$ comme l'a suggéré Singh (1996).

J'examine des estimations de régression modifiée pour z_{ti} du type ci-dessous, pour $a \in [0, 1]$:

$$z_{ti} = (1-a)\frac{8}{7}y_{t-1,i}^* + a\left(y_{ti}^* - \frac{8}{7}(y_{ti}^* - y_{t-1,i})\right) \quad \text{pour } i \in \mathrm{D}$$
$$= a y_{t,i}^* \quad \text{pour } i \notin \mathrm{D}. \tag{10}$$

Compte tenu de cette définition, nous avons

$$\hat{z}_t^{\mathrm{H}} = (1-a)\hat{y}_{t-1}^{+\mathrm{HD}} + a(\hat{y}_t^{*\mathrm{H}} - (\hat{y}_t^{*\mathrm{HD}} - \hat{y}_{t-1}^{+\mathrm{HD}})), \tag{11}$$

où $\hat{y}_{t-1}^{+\mathrm{HD}} = 8/7 \sum_{i \in \mathrm{D}} w_{t,i}^* y_{t-1,i}^{\pi}$ et $\hat{y}_t^{*\mathrm{HD}} = 8/7 \sum_{i \in \mathrm{D}} w_{t,i}^{\pi} y_{t,i}^*$ sont des estimations de Y_{t-1}^* et de Y_t^* respectivement en fonction d'unités en D seulement et des poids de sélection du mois courant. Pour $a = 0$, $\hat{z}_t^{\mathrm{H}}$ est simplement l'estimation $\hat{y}_{t-1}^{+\mathrm{HD}}$. Pour $a = 1$, $\hat{z}_t^{\mathrm{H}}$ est l'estimation de Horvitz-Thompson du mois courant moins une estimation de mouvement fondée sur les groupes de renouvellement appariés $\hat{y}_t^{*\mathrm{HD}} - \hat{y}_{t-1}^{+\mathrm{HD}}$. Les valeurs $a = 0$ et $a = 1$ donnent les méthodes MR1 et MR2 respectivement de Singh et coll. (1997). L'utilisation d'une valeur a intermédiaire a été suggérée par Fuller (1999).

Un pseudo-repère Z_t approprié serait une estimation de Y_{t-1}^* rajustée de façon à correspondre aux poids du mois courant. Suivant Singh et coll. (1997), j'ai utilisé une étape de régressons généralisée afin de rajuster l'estimateur de régression modifiée du mois précédent pour que l'addition donne les repères du mois courant :

$$Z_t = \hat{y}_{t-1}^{*\mathrm{M}} + (X_t - \hat{x}_{t-1}^{\mathrm{M}})\beta_t^{\mathrm{adj}} \tag{12}$$

$$\text{pour } \beta_t^{\mathrm{adj}} = \left(\sum_i w_{t-1,i}^{\mathrm{M}} x_{t-1,i}' x_{t-1,i}\right)^{-1} \sum_i w_{t-1,i}^{\mathrm{M}} x_{t-1,i}' y_{t-1,i}^*. \tag{13}$$

À noter que $Z_t \approx \hat{y}_{t-1}^{*\mathrm{M}}$ puisque $\hat{x}_{t-1}^{\mathrm{M}} = X_{t-1} \approx X_t$. Cela termine la définition des estimateurs de régression modifiée.

4.2 Propriétés des estimateurs de régression modifiée

Le mouvement $\hat{y}_t^{*\mathrm{HD}} - \hat{y}_{t-1}^{+\mathrm{HD}}$ en (11) se fonde en réalité uniquement sur l'échantillon apparié (c'est-à-dire les unités qui déclarent aux deux moments), puisque les autres unités des groupes de renouvellement appariés D contribuent zéro au mouvement (pour l'imputation utilisée ici). Il peut en résulter des estimateurs de régression modifiée ayant une erreur d'échantillonnage moins élevée qu'un estimateur *AK*, car ce « mouvement de l'échantillon apparié » n'est pas influencé par les unités qui ne sont pas présentes pour les deux mois.

Malheureusement, cela entraîne également une possibilité de biais lorsque des personnes qui ne sont pas représentées dans l'échantillon apparié manifestent un comportement différent de celui des personnes de

l'échantillon apparié. Ce peut être effectivement le cas : l'échantillon apparié exclut les personnes qui ont changé de logement entre les deux mois, et il est possible que le changement de logement soit lié à un changement d'emploi. Ce « biais de l'échantillon apparié » vient s'ajouter à tout biais d'accoutumance.

Une autre difficulté est liée à l'estimateur MR2 (donc, $a = 1$). Si la k-ième variable clé $y^*_{ti,k}$ comporte une forte corrélation d'un mois à l'autre, elle comportera également une forte corrélation avec la k-ième nouvelle variable auxiliaire $z_{ti,k}$. Pour une telle variable, l'élément de β_t^M qui correspond à $z_{ti,k}$ aura une certaine valeur y_t proche de un. Si l'on utilise (7), (11) et $Z_t \approx \hat{y}^{*M}_{t-1}$, l'estimateur MR2 adopte la forme

$$\hat{y}^{*M}_{t,k} \approx \left(1 - y_t)\hat{y}^{*H}_{t,k} + y_t(\hat{y}^{*M}_{t-1,k} + (\hat{y}^{*HD}_{t,k} - \hat{y}^{+HD}_{t-1,k})\right) + \text{autres termes.} \tag{14}$$

Dans ce cas, il est possible que le mouvement de l'échantillon apparié à un moment donné influence fortement les estimations lors de nombreux moments subséquents. De plus, un biais faible dans le mouvement aura tendance à s'accumuler avec le temps. Ce danger a été reconnu par Fuller (1999) et qualifié de « problème de dérive ». C'est ce qui l'a poussé à suggérer la forme d'estimateur donnée ici, avec une valeur de a inférieure à 1.

En résumé, la régression modifiée offre des avantages semblables à ceux de la stratégie de pondération composite AK, mais avec une erreur d'échantillonnage possiblement inférieure. La méthode n'est pas difficile à appliquer, et elle permet d'éviter la nécessité d'un calage distinct des groupes de renouvellement en fonction des repères.

5. MEILLEURE ESTIMATION LINÉAIRE SANS BIAIS (MELSB)

5.1 MELSB pour une fenêtre fixe

On obtient l'estimateur MELSB pour une fenêtre fixe (noté $\hat{y}_t^B$) en choisissant une combinaison linéaire « optimale » des estimations de groupe de renouvellement $\hat{y}_t^{Rr}$ (définie en 2.3) à partir d'une fenêtre de $l + 1$ mois, comme suit :

$$\hat{y}_t^B = \sum_{s=t-l}^{t} \sum_{r=1}^{8} a_{sr} \hat{y}_s^{Rr} \tag{15}$$

où les paramètres a_{sr} sont choisis de façon à réduire $(\hat{y}_t^B)$ au minimum moyennant les contraintes $\Sigma_{r=1}^8 a_{sr} = 1$ pour $s = t$ et $\Sigma_{r=1}^8 a_{sr} = 0$ pour $s = t - l, ..., t - 1$. Ces contraintes garantissent que $\hat{y}_t^B$ sera sans biais pour Y_t, à la condition que les estimations de groupe de renouvellement soient sans biais, c'est-à-dire que $E(\hat{y}_s^{Gr}) = Y_s$ pour $s = t - l, ..., t$.

La réduction au minimum suppose que les variances et les covariances des estimations de groupe de renouvellement soient connues. Dans la pratique, celles-ci sont estimées en fonction de données chronologiques. Le problème se laisse alors écrire sous une forme matricielle : nous cherchons à choisir le vecteur de colonne a (avec les éléments a_{sr} pour $s = t - l, ..., t$ et $r = 1, ..., 8$) de façon à réduire une forme quadratique $a'Va$ au minimum sous réserve des contraintes $C'a = c$. Le résultat standard pertinent (Rao 1973, page 65) est que le minimum survient pour $a = V^{-1}Cq$ où q est une solution de $(C'V^{-1}C)q = c$. Dans la présente étude, la matrice V a été remplacée par une matrice de corrélation, suivant l'hypothèse selon laquelle toutes les estimations de groupe de renouvellement dans la fenêtre comportaient la même variance.

5.2 Structure de corrélation des estimations de groupe de renouvellement

Puisque différents modes de corrélation donnent diverses estimations MELSB, le choix d'un profil de corrélation comporte des aspects semblables à ceux du choix des paramètres A et K pour la stratégie composite AK. Il est souhaitable d'utiliser la même combinaison linéaire pour toutes les estimations afin de garantir l'additivité des estimations.

J'ai adopté un modèle à quatre paramètres pour le mode de corrélation :

$$\begin{aligned} \operatorname{corr}(\hat{y}_t^{Gr}, \hat{y}_s^{Gr'}) &= \rho^W_{|t-s|} \quad \text{pour } r - r' = t - s \\ &= \rho^B_{|t-s|} \quad \text{pour } r - r' = t - s + 8m \\ &\qquad \text{pour la nombre entier } m \neq 0 \\ &= 0 \quad \text{autrement.} \end{aligned} \tag{16}$$

Ainsi, la corrélation entre les estimations pour un décalage k du même groupe de renouvellement est ρ_k^W si le groupe de renouvellement comporte les mêmes logements aux deux moments, et ρ_k^B autrement. Il n'y a pas de corrélation entre les estimations de différents groupes de renouvellement. Un modèle à quatre paramètres est utilisé :

$$\rho_k^W = (1 - r_U^2)(\theta_P^k r_P^2 + \theta_B^k(1 - r_P^2)) \text{ et} \tag{17}$$

$$\rho_k^B = (1 - r_U^2)\,\theta_B^k(1 - r_P^2). \tag{18}$$

Bell et Carolan (1998) ont discuté ce modèle. Les valeurs paramétriques utilisées dans cet exposé étaient $\theta_P = 0{,}87697$, $\theta_B = 0{,}94$, $r_U = 0{,}3101$ et $r_P = 0{,}90456$. Ces valeurs sont le résultat d'un ajustement du modèle en fonction d'autocorrélations estimatives pour les estimations de groupe de renouvellement de la proportion d'employés.

Il importe de noter que les estimations de type MELSB sont sans biais peu importe l'exactitude du modèle de corrélation supposé. Le modèle utilisé ici cherche à être optimal pour des estimations des personnes ayant un

emploi, mais fonctionne bien aussi pour des personnes sans emploi. L'utilisation d'autres valeurs pour les paramètres du modèle n'a donné aucune amélioration appréciable de l'erreur-type pour les personnes sans emploi.

5.3 Estimations MELSB améliorées

Une difficulté liée aux estimations MELSB ci-dessus est qu'il faut des estimations RG au niveau du groupe de renouvellement. La taille plus petite de l'échantillon au niveau du groupe de renouvellement risque de limiter les repères utilisables, comme on l'a vu pour l'*AK*. Pour le MELSB, toutefois, une autre stratégie est possible.

On définit l'estimateur B1 en établissant un estimateur MELSB fondé sur les estimateurs de Horvitz-Thompson au niveau du groupe de renouvellement, puis en appliquant la technique de régression généralisée afin d'améliorer cet estimateur. Il s'agit de procéder comme suit. On définit $y_{ti}^{\#} = a_{tR(t,i)} y_{ti}$ et $x_{ti}^{\#} = a_{tR(t,i)} x_{ti}$, où $a_{tR(t,i)}$ est le multiplicateur MELSB applicable au groupe de renouvellement dans lequel se trouve l'unité i à un moment t. On peut alors écrire l'estimateur MELSB fondé sur les estimateurs de Horvitz-Thompson sous la forme

$$\hat{y}_t^{BH} = \sum_{s=t-l}^{t} \sum_i w_{si}^{\pi} y_{si}^{\#}. \tag{19}$$

En calant en fonction des repères, on obtient l'estimateur MELSB amélioré B1 :

$$\hat{y}_t^{B1} = \hat{y}_t^{BH} + (X_t - \hat{x}_t^{BH})\hat{\beta} \tag{20}$$

$$\text{pour } \hat{\beta} = \left(\sum_{s=t-l}^{t} \sum_i w_{si}^{\pi} x_{si}^{\#\prime} x_{si}^{\#} \right)^{-1} \sum_{s=t-l}^{t} \sum_i w_{si}^{\pi} x_{si}^{\#\prime} y_{si}^{\#}. \tag{21}$$

$$\text{donc, } \hat{y}_t^{B1} = \sum_{s=t-l}^{t} \sum_i w_{si}^{B1} y_{si} \tag{22}$$

$$\text{pour } w_{si}^{B1} = w_{si}^{\pi} a_{sR(s,i)} \left\{ 1 + (X_t - \hat{x}_t^{BH}) \left(\sum_{u=s-l}^{s} \sum_j w_{uj}^{\pi} a_{uR(uj)} x_{uj}' x_{uj} \right)^{-1} a_{sR(s,i)} x_{si}' \right\}. \tag{23}$$

Propriétés des estimateurs MELSB et B1

Les estimations de type MELSB et B1 sont des sommes de données unitaires pondérées pour une fenêtre de mois. Chaque estimation n'exige que des données de cette fenêtre et se laisse préparer indépendamment des estimations pour les mois précédents, de sorte que la méthode n'est pas récursive.

Le même mois de données contribuera différents poids à l'estimation pour différents moments. Une unité contribuera un poids appréciable à son estimation du mois courant, et un poids proche de zéro, souvent négatif, aux estimations d'autres mois. Le travail de préparation d'un tableau est le même que pour la RG multipliée par la taille de la fenêtre. Il existe également une possibilité d'estimations négatives pour de petites cellules ne comportant aucune unité courante.

À noter que, dans l'estimateur B1, on ne force pas l'addition des poids appliqués aux mois autres que le mois courant de donner une valeur de zéro. En vertu des hypothèses du modèle, l'estimation $\hat{y}_t^{B1}$ demeure inconditionnellement sans biais, puisque $\hat{y}_t^{BH}$ et $\hat{x}_t^{BH}$ sont sans biais pour Y_t et X_t respectivement. Dans la pratique, le mois courant contribue 99,5 % environ du poids total. J'estime que le biais qui en résulte est faible et sans danger (entraînant un léger lissage des estimations avec le temps).

Pour toute estimation dans laquelle il existe une corrélation appréciable des données d'un mois à l'autre, les estimations MELSB et B1 devraient comporter une erreur d'échantillonnage moindre que l'estimation RG. C'est là un avantage théorique relativement à une méthode servant à améliorer un ensemble d'estimations déterminé à l'avance (comme la régression modifiée ou l'*AK* avec pondération composite). Dans la pratique, cet avantage n'est pas nécessairement très important, car pour l'EPA on s'intéresse surtout à un petit nombre d'estimations bien définies.

L'utilisateur doit également déterminer la période de temps ou la « fenêtre » des estimations. Le fait d'utiliser trop de moments suppose beaucoup de temps d'ordinateur, tandis qu'un nombre trop faible réduit les avantages. La fenêtre de sept mois utilisée ici a été suffisante pour offrir presque tous les avantages ; des fenêtres plus petites comportent des erreurs-types appréciablement plus élevées.

6. COMPARAISON DES MÉTHODES

6.1 Méthode de comparaison

Les estimations pour la période allant de juillet 1993 à janvier 1999 ont été préparées en fonction de données allant de janvier 1993 à janvier 1999. On a obtenu des estimations classées selon le mois, l'État, le sexe, l'état matrimonial et la situation par rapport au marché du travail. On a également obtenu des estimations pour le mouvement de décalage un, la moyenne trimestrielle et le mouvement de moyennes trimestrielles successives.

L'erreur-type de ces estimations a été calculée à l'aide de la technique « jackknife avec suppression d'un groupe » (Kott 1998). Les unités géographiques qui constituent la première étape du tirage de l'échantillon ont été divisées systématiquement en $G = 30$ groupes, et « des groupes répétés » ont été établis en fonction de l'échantillon complet à l'exclusion des unités de l'un de ces groupes. Chaque estimation étudiée a également été préparée pour chacun des groupes répétés G. Lorsqu'on écrit e pour l'estimation et $e_{(g)}$ pour l'estimation de la répétition g, l'estimation de type jackknife avec suppression d'un groupe de l'erreur-type est donnée par

$$SE_{(e)} = \sqrt{\frac{G-1}{G}\sum_{g=1}^{G}(e_{(g)}-e)^2}. \qquad (24)$$

On a obtenu les estimations et les erreurs-types pour chacun des estimateurs ci-dessous (énumérés en abrégé pour faciliter la suite de l'exposé) :

RG : L'estimation de régression généralisée utilisée couramment dans l'EPA

AK : L'estimation *AK* avec K=0,7, A = 0,06

ME : L'estimateur MELSB fondé sur une fenêtre de 7 mois

B1 : Le MELSB amélioré fondé sur une fenêtre de 7 mois

MR2 : L'estimateur MR2 (régression modifiée avec $a = 1$)

MF : La variante de Fuller de la régression modifiée ($a = 0{,}7$)

Les estimateurs de régression modifiée supposent un choix des variables clés qu'il s'agit d'optimiser. Pour préparer les estimations de régression modifiée du présent rapport, des variables z ont été préparées pour les estimations de l'emploi et du chômage pour chaque État et chaque sexe. Cela donne au total 32 variables auxiliaires supplémentaires, en plus des 540 repères de strate après sélection habituels utilisés dans la régression généralisée.

6.2 Différences par rapport à l'estimation RG

On peut utiliser l'estimateur RG courant comme base de comparaison pour les autres estimateurs. Au lieu de présenter des diagrammes d'estimations de niveau, je montre les différences des autres estimations par rapport aux estimations RG courantes. Les diagrammes 1 et 2 indiquent ces différences pour les estimations des personnes ayant un emploi et des personnes sans emploi respectivement. Afin de situer l'ampleur de ces différences, notons que les erreurs-types publiées pour l'estimation courante étaient de 25 200 pour les personnes ayant un emploi et de 7 900 pour les personnes sans emploi en janvier 1999 (semblables pour les autres mois).

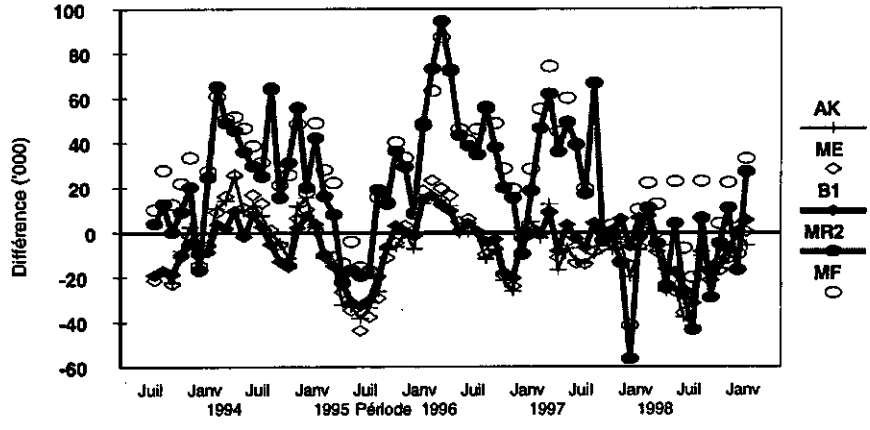


Diagramme 1. Différence des autres estimations par rapport à l'estimation RG, personnes ayant un emploi (milliers), juillet 1993 à janvier 1999.

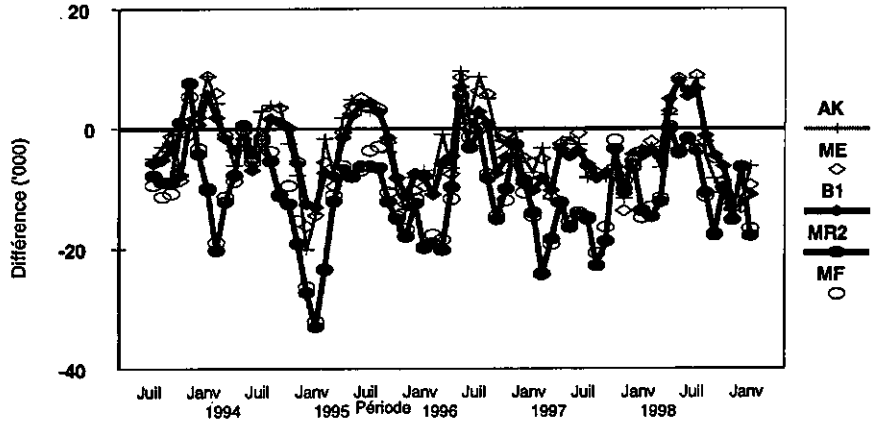


Diagramme 2. Différence des autres estimations par rapport à l'estimation RG, personnes sans emploi (milliers), juillet 1993 à janvier 1999.

Les estimations *AK*, ME et B1 sont assez semblables, puisque, pour l'ensemble des trois méthodes, l'apport d'une unité dépend de son groupe de renouvellement. Dans les deux diagrammes, les estimateurs *AK*, ME et B1 semblent donner, en moyenne, des valeurs plus faibles que les estimations RG. Cela indique un changement du biais d'accoutumance, résultat du fait que l'on accorde un poids moindre au groupe de renouvellement échantillonné pour la première fois. Les estimations varient vers le haut et vers le bas relativement à leur différence moyenne pour de courtes périodes.

Les estimations MR2 et MF tendent à différer des autres estimations, puisqu'elles soulignent l'apport des unités de l'échantillon apparié. Pour les personnes ayant un emploi, les estimateurs MR2 et MF sont appréciablement plus grands en moyenne que les estimations RG, jusqu'à septembre 1997. Il y a alors une baisse des différences correspondant à l'intégration d'un nouvel échantillon à partir de septembre 1997. Pour des raisons qui ne sont pas claires, au cours de cette période l'échantillon apparié manifeste un comportement différent de celui de l'échantillon global. Cela influe sur la différence entre ces séries de régression modifiée et la série RG. Ce qui risque d'être inquiétant est le fait que le changement de niveau influe sur le niveau de la série MR2 au cours d'une période subséquente considérable, manifestation possible du « problème de dérive ».

Pour les personnes sans emploi, les estimations MR2 et MF tendent à être inférieures aux estimations RG. Il n'y a aucune indication d'un « problème de dérive » pour les personnes sans emploi, ce qui n'a rien de surprenant compte tenu des corrélations moindres mises en jeu.

6.3 Différences moyennes selon le mois civil

Afin de quantifier le changement probable de biais suivant le passage à un nouvel estimateur, on a calculé la différence moyenne au cours de la période de chaque estimation relativement à l'estimation RG. Il est possible que la différence soit saisonnière, et l'on a donc obtenu des moyennes distinctes pour chaque mois de l'année civile, de même que pour l'ensemble. On trouvera dans le diagramme 3 les différences moyennes au cours de la période allant de juillet 1993 à janvier 1999 pour les personnes ayant un emploi.

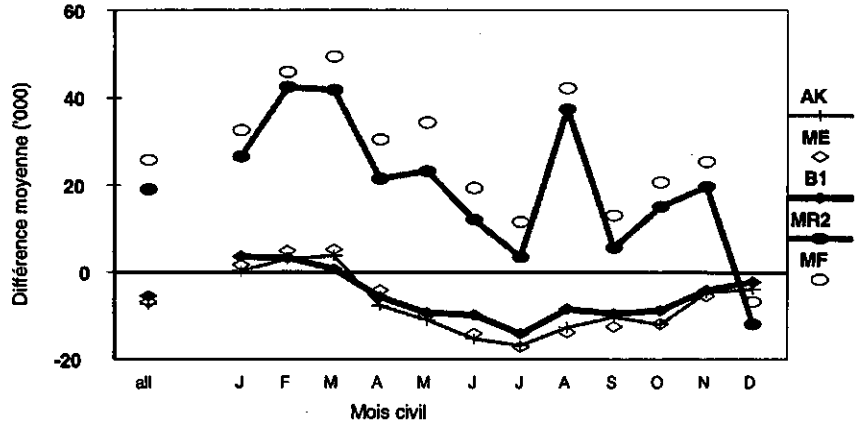


Diagramme 3. Différence moyenne par rapport à l'estimation RG, globalement et selon le mois civil, personnes ayant un emploi (milliers).

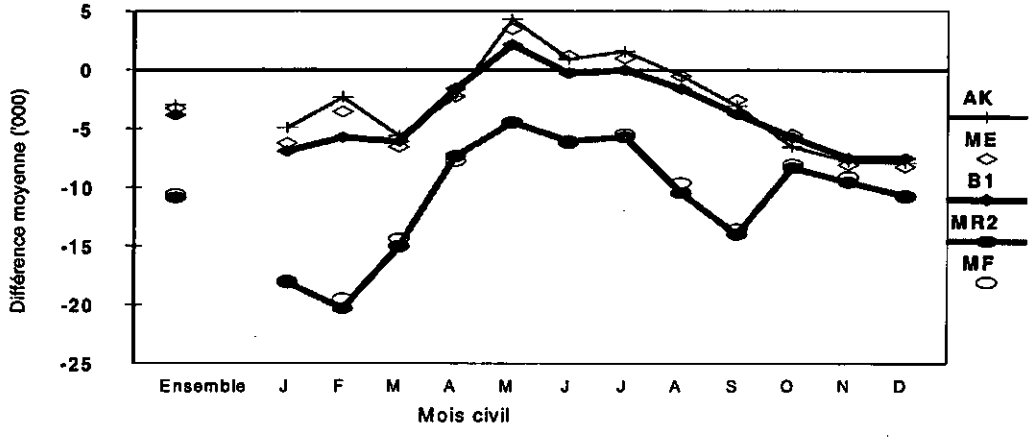


Diagramme 4. Différence moyenne par rapport à l'estimation RG, globalement et selon le mois civil, personnes sans emploi (milliers).

Le diagramme indique que les estimations de personnes ayant un emploi auraient été supérieures en moyenne si l'on avait utilisé l'estimateur MR2 ou MF. Cette différence vers le haut pour les estimateurs de régression modifiée est peut-être en réalité un aspect de la période particulière, puisque la différence semble avoir disparu depuis septembre 1997.

L'autre aspect des estimations MR2 et MF est le fait que la différence pour les personnes ayant un emploi est très saisonnière. Par exemple, le mouvement de décembre à janvier des estimations MR2 est supérieur de 40 000 environ au mouvement des estimations RG. Cela semble indiquer que l'échantillon apparié tend à manquer des personnes qui avaient un emploi en décembre, mais non pas en janvier. Le même caractère saisonnier se manifeste lorsqu'on observe directement des estimations de l'échantillon apparié. Le mouvement des groupes de renouvellement appariés ne manifeste pas cet important biais saisonnier.

Pour les estimations *AK*, ME et B1, les différences manifestent un certain caractère saisonnier, mais elles sont beaucoup plus faibles.

Le diagramme 4 indique les différences moyennes des diverses estimations par rapport à l'estimation RG pour les personnes sans emploi au cours de la même période. Ici, il semble y avoir une différence négative pour tous les estimateurs, mais elle est moins marquée pour les estimations *AK*, ME et B1 que pour les estimations MR2 et MF. Le changement de caractère saisonnier pour un passage des estimateurs RG à MR2 et MF est encore une fois plus marqué que pour le passage aux autres estimateurs.

6.4 Erreurs-types

Les erreurs-types (ET) des estimations dans l'ensemble, selon l'état matrimonial et selon le sexe, sont présentées dans les diagrammes qui suivent. On les obtient comme pourcentage des estimations de l'ET pour la même estimation à l'aide de la méthode RG (c'est-à-dire les ET de l'EPA courante), et on établit alors la moyenne de ces pourcentages pour la période en fonction de laquelle ils ont été préparés (juin 1993 à janvier 1999 pour les estimations de niveau). Les diagrammes 5, 6, 7 et 8 indiquent les ET tant pour les personnes ayant un emploi que pour les personnes sans emploi en fonction du niveau, du mouvement, de la moyenne trimestrielle et du mouvement de la moyenne trimestrielle respectivement.

Pour toutes ces estimations, l'estimateur B1 de type MELSB comporte une erreur d'échantillonnage plus faible que les estimateurs *AK* ou ME. Puisque l'estimation B1 semble manifester un biais et un caractère saisonnier de biais comparables, les estimateurs *AK* et ME utilisés dans la présente étude ne semblent pas pouvoir concurrencer l'estimateur B1.

Par contre, les estimateurs de régression modifiée MR2 et MF manifestent des erreurs d'échantillonnage nettement inférieures à celles de l'estimateur B1 pour les personnes ayant un emploi, pour les estimations dans l'ensemble et pour les estimations selon le sexe. Ce sont là des estimations clés de la régression modifiée – d'autres estimations clés comme les estimations selon l'État ont aussi montré de meilleures erreurs-types. Les estimations selon l'état matrimonial ne sont pas des estimations clés, et elles comportent des erreurs-types plus élevées pour les estimations MR2 et MF que pour l'estimateur B1.

En ce qui concerne les personnes sans emploi, l'amélioration des ET pour un recours aux estimateurs MR2 et MF est moins constante, et elle disparaît complètement pour les estimations de la moyenne trimestrielle. L'estimateur B1 réduit l'erreur-type de façon plus constante, même si les gains sont moins élevés pour les personnes sans emploi que pour les personnes ayant un emploi.

6.5 Série désaisonnalisée et série de tendances

Le Bureau of Statistics de l'Australie fait appel au progiciel X11 (Shiskin, Young et Musgrave 1967) pour préparer des estimations désaisonnalisées visant à éliminer des séries divers effets de l'année civile. Le progiciel permet également de préparer une tendance, qui est un indicateur du comportement sous-jacent de la série.

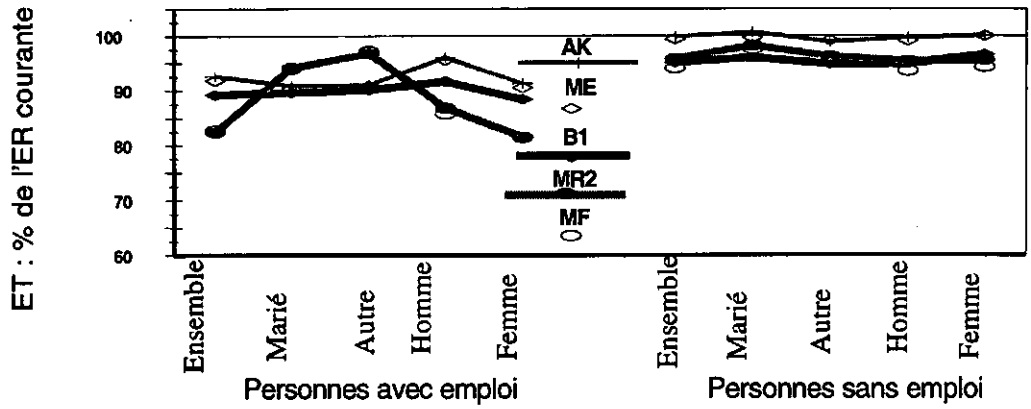


Diagramme 5. Erreur-type du niveau (% de l'ET courante)

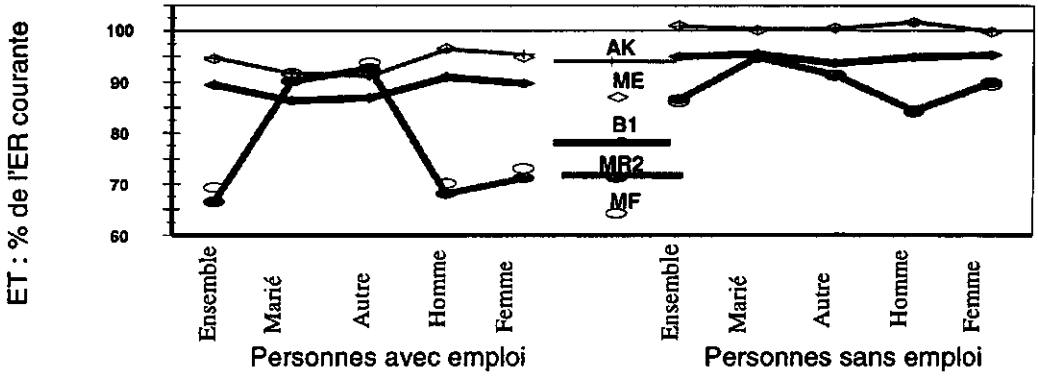


Diagramme 6. Erreur-type du mouvement (% de l'ET courante)

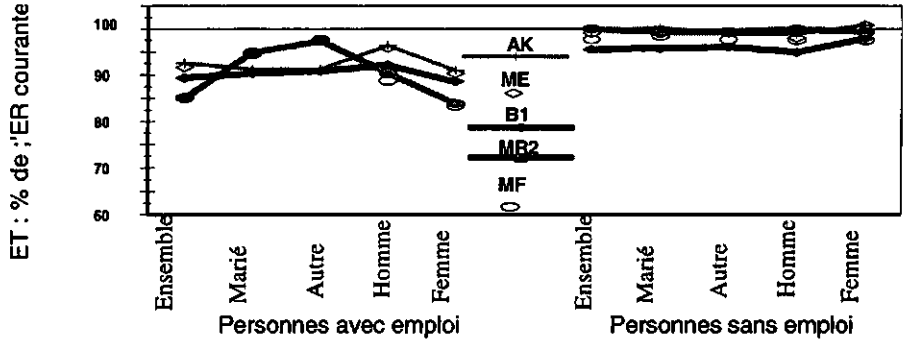


Diagramme 7. Erreur-type de la moyenne trimestrielle (% de l'ET courante)

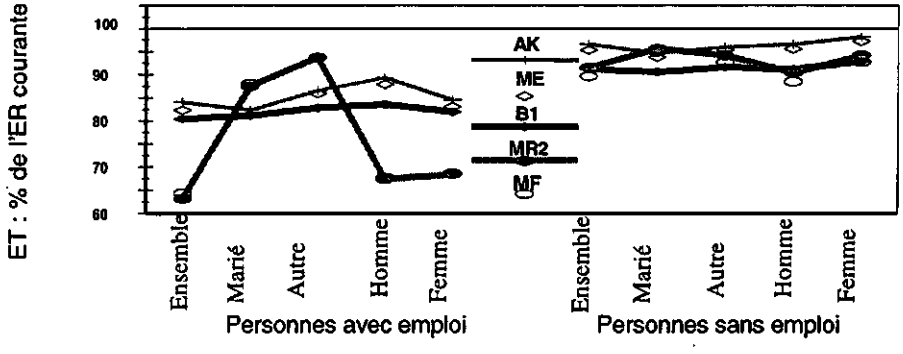


Diagramme 8. Erreur-type du mouvement de la moyenne trimestrielle (% de l'ET courante)

La valeur de la tendance pour tel moment est révisée à mesure que les données pour des moments ultérieurs sont accessibles. J'ai estimé l'erreur-type des estimations de tendance à la fin de la série (tendance de fin) et pour les mêmes moments lorsque douze autres mois de données sont disponibles (tendance à mi-chemin). Les révisions de la tendance (ou du mouvement de tendance) sont définies comme la différence entre la valeur à mi-chemin et la valeur de fin de la tendance (ou du mouvement de tendance). L'ampleur de la révision dépend de la forme de la série réelle et de l'erreur d'échantillonnage des séries estimées. La révision de la tendance quadratique moyenne pour une série d'estimations non biaisées est la somme de deux composantes : la révision de la tendance quadratique moyenne qui serait survenue même en l'absence d'erreur d'échantillonnage, et la variance de l'estimation de la révision. Ainsi, l'erreur-type de la révision est une mesure de la composante erreur d'échantillonnage de la révision de la tendance quadratique moyenne (voir Bell 1999).

Les chiffres désaisonnalisés se prêtent à une révision semblable. Je présente les erreurs-types pour le niveau et le mouvement des estimations désaisonnalisées à la fin de la série. Les erreurs-types étaient très semblables pour les révisions ultérieures de ces estimations.

La technique jackknife avec suppression d'un groupe a permis de préparer des estimations de l'erreur-type pour les différentes estimations de tendance et désaisonnalisées. Cette technique suppose la préparation de versions répétées

des estimations. Malheureusement, l'étude a fourni des valeurs répétées pour la série chronologique uniquement pour les moments entre juin 1993 et janvier 1999. Chacune de ces séries chronologiques répétées a été complétée par les neuf années précédentes de données chronologiques afin qu'il y ait suffisamment de données pour l'application du progiciel X11. Puisque les séries de tendances et désaisonnalisées répétées se fondent sur les mêmes valeurs avant juillet 1993, l'estimation de type jackknife de l'ET aura tendance à sous-estimer légèrement l'ET réelle, surtout pour les moments au début de la série. Afin de réduire cet effet au minimum, on établit la moyenne des mesures du changement de l'erreur d'échantillonnage pour les mois à compter de janvier 1995 uniquement (et uniquement jusqu'à janvier 1998, de façon que l'on puisse utiliser les douze mois jusqu'à janvier 1999 pour l'estimation des révisions).

Le tableau 1 présente ces erreurs-types moyennes pour diverses mesures de tendance et désaisonnalisées, relativement à celles que l'on peut obtenir de l'estimateur RG courant, tant pour les personnes ayant un emploi que pour les personnes sans emploi. Dans ce même tableau on trouve les chiffres correspondants pour le niveau, le mouvement, la moyenne trimestrielle et le mouvement de la moyenne trimestrielle, en fonction des diagrammes 5 à 8.

Tableau 1
Erreur-type comme pourcentage de l'erreur-type de l'estimateur RG courant

	AK	ME	B1	MR2	MF
Personnes ayant un emploi :					
Niveau	93	92	89	82	83
Mouvement	95	95	89	66	69
Moyenne trimestrielle	93	92	89	85	85
Mouvement de la moyenne trimestrielle	84	82	80	63	64
Désaisonnalisé	94	92	90	87	88
Mouvement de la désaisonnalisation	96	95	91	68	71
Tendance de fin	93	91	89	88	88
Mouvement de la tendance de fin	86	84	82	65	67
Révision de la tendance	88	85	83	66	68
Révision du mouvement de la tendance	89	86	84	67	69
Personnes sans emploi :					
Niveau	100	99	95	96	94
Mouvement	101	101	95	87	86
Moyenne trimestrielle	100	99	95	100	98
Mouvement de la moyenne trimestrielle	97	95	91	92	90
Désaisonnalisé	100	99	95	96	95
Mouvement de la désaisonnalisation	102	102	95	87	86
Tendance de fin	99	98	95	99	97
Mouvement de la tendance de fin	97	95	92	93	91
Révision de la tendance	97	95	91	91	89
Révision du mouvement de la tendance	97	95	92	92	90

J'estime que, pour de nombreuses applications, les indicateurs les plus importants sont ceux qui indiquent la direction sous-jacente de la série à la fin courante, c'est-à-dire le mouvement de la moyenne trimestrielle et le mouvement de la tendance. Une réduction de l'erreur-type pour ces éléments rend la direction sous-jacente de la série à la fin plus claire, même pour les utilisateurs qui se fient sur une inspection visuelle ou sur un processus de lissage quelconque autre que la tendance X11. Cela offre dès lors de meilleures chances de détecter les points tournants de la série sous-jacente.

Pour ce qui est du mouvement de la tendance, l'estimateur B1 permet une réduction de 18 % de l'erreur- type pour les personnes ayant un emploi et une réduction de 8 % pour les personnes sans emploi. Quant au MR2, ces réductions sont de 35 % et de 7 % respectivement. Les estimateurs composites permettent également de réduire l'apport de l'erreur d'échantillonnage aux révisions de la série des tendances.

6.6 Résumé

Le présent exposé décrit une variante de l'estimateur MELSB, l'estimateur B1, qui permet d'appliquer la technique de régression généralisée à une estimation composite fondée sur une fenêtre de sept mois de données. Pour les données de l'Australie, le B1 comporte une erreur d'échantillonnage moins élevée que les estimateurs MELSB ou *AK* traditionnels pour un choix de mesures comprenant les estimations de tendance et désaisonnalisées. L'auteur a également évalué un estimateur MR2 composite « de régression modifiée » proposé par A.C. Singh et une variante de cet estimateur proposée par W. Fuller. Ces estimateurs ont donné des erreurs d'échantillonnage appréciablement moins élevées que l'estimateur B1 pour un certain nombre de mesures, en particulier celles qui se fondent sur des personnes ayant un emploi.

L'évaluation d'un estimateur composite dépend de nombreux facteurs autres que les erreurs d'échantillonnage. L'estimateur B1 comporte l'inconvénient d'exiger pour les totalisations des agrégations pondérées de sept mois de données, tandis que les estimateurs de régression modifiée fournissent des poids pour un seul mois de données. Par contre, les estimateurs de régression modifiée risquent d'être biaisés si les personnes déclarantes au cours de deux mois successifs (l'échantillon apparié) ne sont pas représentatives des autres personnes (par exemple les personnes qui changent de logement). L'introduction des estimateurs de régression modifiée entraînerait également un changement plus prononcé de l'estimation et du caractère saisonnier que l'introduction de l'estimateur B1.

REMERCIEMENTS

Les idées exprimées dans le présent exposé sont celles de l'auteur et ne reflètent pas nécessairement celles du Bureau of Statistics de l'Australie. Les personnes qui citent ou qui utilisent ces idées voudront bien les attribuer clairement à l'auteur.

BIBLIOGRAPHIE

BAILAR, B.A. (1975). The effect of rotation group bias on estimates from panel surveys. *Journal of American Statistical Association*, 70, 23-29.

BELL, P.A. (1998). Using state space models and composite estimation to measure the effects of telephone interviewing on labour force estimates. Document de travail *Econometrics and Applied Statistics*, numéro de 1351.0, no. 98/2, ABS, Canberra.

BELL, P.A. (1999). The impact of sample rotation patterns and composite estimation on survey outcomes. Document de travail *Econometrics and Applied Statistics*, numéro de 1351.0, no. 99/1, ABS, Canberra.

BELL, P.A., et CAROLAN, A. (1998). Trend estimation for small areas from a continuing survey with controlled sample overlap. Document de travail *Econometrics and Applied Statistics*, numéro de 1351.0, no. 98/1, ABS, Canberra.

FULLER, W.A. (1990). Analyse d'enquêtes à passages répétés. *Techniques d'enquête*, 16, 177-190.

FULLER, W.A. (1999). Canadian Regression Composite Estimation. Manuscript non publié.

GURNEY, M., et DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 247-257.

JESSEN, R.J. (1942). Statistical investigation of a farm survey for obtaining farm facts. *Iowa Agricultural station research Bulletin*, 304.

KOTT, P.S. (1998). Using the delete-a-group jackknife variance estimator in practice. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 483-486.

LENT, J., MILLER, S. et CANTWELL, P. (1994). Composite weights for the Current Population Surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 867-872.

LENT, J., MILLER, S. et CANTWELL, P. (1996). Effect of composite weights on some estimates from the Current Population Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 130-139.

RAO, C.R. (1973). *Linear Statistical Inference and its Applications*. Deuxième édition. New York : John Wiley and Sons.

SINGH, A.C., and MERKOURIS, P. (1995). Composite estimation by modified regression for repeated surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 420-425.

SINGH, A.C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 120-129.

SINGH, A.C., KENNEDY B., WU S. et BRISEBOIS F. (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 300-305.

SHISKIN, J., YOUNG, A., and MUSGRAVE, J. (1967). *The X-11 variant of Census Method II Seasonal Adjustment*, Bureau of the Census, U.S. Department of Commerce, Papier technique 15.

YANSANEH, I.S., et FULLER, W.A. 1998. Méthode optimale d'estimation récursive pour les enquêtes répétitives. *Techniques d'enquête*, 24, 33-42.

Techniques d'enquête, juin 2001
Vol. 27, n° 1, pp. 69-79
Statistique Canada

Estimation composite par régression pour l'Enquête sur la population active du Canada : Évaluation et application

JACK GAMBINO, BRIAN KENNEDY et MANGALA P. SINGH[1]

RÉSUMÉ

L'Enquête sur la population active (EPA) du Canada est une enquête mensuelle réalisée selon un plan de sondage complexe avec renouvellement de panel. Après des études approfondies, y compris l'examen de diverses méthodes tirant parti du chevauchement de l'échantillon pour améliorer la qualité des estimations, l'équipe de l'EPA a choisi une méthode d'estimation composite qui permet d'atteindre cet objectif tout en respectant les contraintes d'ordre pratique. En outre, pour les variables pour lesquelles le gain d'efficacité est important, la nouvelle série chronologique a tendance à être plus compréhensible pour les spécialistes du domaine. Il est donc plus facile d'expliquer les estimations produites d'après l'EPA aux utilisateurs des données et aux membres des médias. Comme l'estimation composite réduit la variance, il est maintenant possible de publier des estimations mensuelles dans bien des cas où, jusqu'à présent, seules les moyennes mobiles de trois ans pouvaient être publiées. En outre, la méthode permet de désaisonnaliser correctement un grand nombre de séries chronologiques.

MOTS CLÉS : Enquête avec renouvellement de panel; système d'estimation; pondération; estimation de la variation; estimation du niveau.

1. INTRODUCTION

1.1 Pourquoi recourir à l'estimation composite?

L'Enquête sur la population active (EPA) du Canada est une enquête mensuelle réalisée auprès de 54 000 ménages sélectionnés selon un plan d'échantillonnage stratifié à plusieurs degrés. Comme les nouveaux ménages sélectionnés font partie de l'échantillon pendant six mois consécutifs, les cinq sixièmes de l'échantillon sont communs d'un mois à l'autre. Chaque mois, on pose aux membres des ménages faisant partie de l'échantillon des questions sur la situation d'activité, les gains et d'autres aspects de l'activité. Dans le système d'estimation de l'EPA utilisé avant 2000, les poids de sondage initiaux étaient modifiés par régression pour produire des poids finals obéissant aux *totaux de population* selon la catégorie âge-sexe et la région géographique (niveau infraprovincial) utilisés comme contrôles. On obtenait ainsi pour chaque enregistrement un *poids final unique* que l'on utilisait pour toutes les totalisations.

Ce système d'estimation n'utilisait que les données du mois courant et ne tirait nullement parti du fait que l'on peut se servir de l'échantillon commun pour améliorer les estimations. Pourtant, certaines caractéristiques, comme l'emploi selon la branche d'activité, sont fortement corrélées dans le temps et d'autres, comme le chômage, le sont modérément, si bien que des gains d'efficacité seraient possibles. C'est pour cette raison que, pour des enquêtes comparables à l'EPA, comme la *Current Population Survey* aux États-Unis, on recourt depuis des années à l'estimation composite pour améliorer les estimations. Cependant, le programme de l'EPA a attendu jusqu'à janvier 2000 pour introduire cette méthode.

Au début des années 1980 (voir Kumar et Lee 1983), la stratégie d'estimation composite de la CPS a été étudiée en vue de son application éventuelle à l'EPA. Bien que cette étude ait montré que l'estimation composite augmentait l'efficacité de l'estimation de l'emploi et, dans une moindre mesure, de l'estimation du chômage, on a estimé à l'époque que les aspects négatifs de la méthode surpassaient les gains. Entre autres, le fait que les paramètres qui optimisent l'estimation de l'emploi et du chômage soient assez différents aurait obligé à trouver un équilibre entre, d'une part, l'utilisation d'un ensemble de paramètres intermédiaires, qui réduirait les gains d'efficacité et, d'autre part, l'obtention de valeurs des variables dont *la somme n'est pas égale aux totaux* (par exemple, la somme des personnes occupées et des chômeurs ne serait pas égale à la population active, à moins que l'une des trois valeurs ne soit obtenue par différence). Le fait que cette méthode d'estimation composite n'était pas *compatible avec les systèmes de pondération, d'estimation et de diffusion* utilisés à ce moment là pour l'EPA était un autre facteur en sa défaveur, puisque son adoption aurait nécessité un remaniement complet des systèmes.

Auparavant, les taux de chômage mensuels étaient les principales estimations produites d'après l'Enquête sur la population active. Toutefois, comme l'intérêt pour l'estimation du taux d'emploi et de sa variation s'est accru ces dernières années, il a fallu multiplier les efforts en vue de découvrir un moyen de profiter de l'échantillon commun, puisque la production de ces estimations en bénéficierait considérablement. Par conséquent, au milieu des années

[1] Jack Gambino, Brian Kennedy et Mangala P. Singh, Statistique Canada, Parc Tunney, Ottawa, (Ontario), Canada K1A 0T6.

1990, Statistique Canada a recommencé à s'intéresser à l'estimation composite et a mis au point une méthode par régression qui s'adapte bien au système existant de calcul des estimations de l'EPA. Cette méthode est décrite dans Singh, Kennedy, Wu et Brisebois (1997), et une version plus à jour est présentée dans Singh, Kennedy et Wu (2001). La nouvelle méthodologie permet de faire un choix entre diverses méthodes, selon l'objectif poursuivi. Si l'on s'intéresse principalement à l'estimation du niveau, on peut utiliser les prédicteurs axés sur le niveau. Par contre, si l'on accorde plus d'importance à la variation, on peut utiliser les prédicteurs axés sur la variation. On peut aussi aller une étape plus loin et inclure les deux types de prédicteurs dans la méthode. Cependant, dans ce dernier cas, le grand nombre de variables indépendantes dans l'équation de régression peut produire une distorsion des poids d'échantillonnage finals.

Les résultats provisoires obtenus d'après la nouvelle méthode en se servant des prédicteurs axés sur la variation et d'autres obtenus en se servant des prédicteurs axés sur le niveau ont été examinés à la réunion du Comité consultatif des méthodes statistiques de Statistique Canada. La méthode permet de résoudre les problèmes que posent les estimateurs composites classiques et de réaliser des gains d'efficacité considérables. Toutefois, l'estimateur qui utilise les prédicteurs axés sur la variation peut causer une dérive des estimations du niveau au fil du temps dans certaines situations extrêmes. Enfin, il a été décidé, à la suite des recommandations du Comité, de donner, tant pour les estimations du niveau que de la variation, un ordre d'importance pour le choix des prédicteurs. Après l'échange de notes techniques entre Wayne Fuller, J.N.K. Rao et les employés de Statistique Canada, le Bureau a adopté une méthode proposée par Fuller, qui combine les stratégies axées sur le niveau et sur la variation sans poser les contraintes dues à l'intégration des deux ensembles de prédicteurs dans la régression (voir Fuller et Rao 2001). La solution est étonnamment simple : prendre une combinaison linéaire des prédicteurs de niveau et de variation : $X = (1-\alpha)X_L + \alpha X_C$, et l'utiliser comme prédicteur. Les prédicteurs de variation et de niveau deviennent alors des cas spéciaux. De surcroît, on peut choisir α de façon à refléter l'importance relative que l'on veut accorder à l'estimation du niveau par opposition à l'estimation de la variation.

Le présent article décrit le nouvel estimateur composite à la section 2. On a procédé à l'évaluation approfondie de cet estimateur en se servant de données réelles de l'EPA pour plusieurs caractéristiques observées pendant une longue période. Les résultats de ces études sont résumés à la section 3. Contrairement aux estimateurs composites classiques, l'estimateur composite de régression exige que l'appariement de l'échantillon entre deux mois consécutifs se fasse au niveau de l'enregistrement individuel. Cette contrainte produit des situations intéressantes où il faut traiter les non-répondants, ainsi que les personnes dans le champ et hors du champ d'un mois à l'autre de façon à ce que la qualité des estimations de la variation ne soit pas compromise. À la section 4, on examine la méthode d'imputation mise au point pour résoudre les divers problèmes que posent des données qui manquent pendant deux mois consécutifs. Enfin, l'évaluation de la qualité de ce nouvel estimateur composite se fonde non seulement sur son efficacité statistique, mais aussi sur sa stabilité au cours du temps, sa rentabilité et la capacité d'atteindre les objectifs suivants : i) réduire au minimum les changements à apporter à l'ancien système d'estimation, ii) produire un poids unique pour chaque unité d'échantillonnage, iii) produire des estimations qui concordent avec les totaux de contrôle selon la catégorie âge-sexe et la région géographique et iv) produire des estimations cohérentes (au sens où, par exemple, personnes occupées + chômeurs = population active et population active + population inactive = personnes de 15 ans et plus). Ces objectifs sont examinés à divers points dans l'article, mais plus spécialement à la section 3.

2. L'ESTIMATEUR COMPOSITE DE RÉGRESSION

Certaines enquêtes, notamment la *Current Population Survey* (CPS) aux États-Unis, mettent à profit le chevauchement de l'échantillon grâce à l'utilisation des estimateurs composites K ou AK. À l'origine, dans le cas de la CPS, on utilisait l'estimateur composite K

$$y_t' = (1-K)y_t + K(y_{t-1}' + \text{variation}_{t-1,t})$$

où $K = 1/2$ à la période t, et où variation $_{t-1,t}$ représente une estimation de la variation fondée sur l'échantillon commun, ou apparié. Cet estimateur a été remplacé par la suite par l'estimateur composite AK

$$y_t' = (1-K)y_t + K(y_{t-1}' + \text{variation}_{t-1,t}) + A(\text{non apparié} - \text{apparié})$$

où $A = 0{,}2$ et $K = 0{,}4$ (voir Cantwell et Ernst 1992). Les valeurs optimales de A et K dépendent de la variable étudiée; or, si l'on applique ce dernier estimateur, l'utilisation de valeurs différentes pour des variables différentes pose des problèmes de cohérence (en ce sens que la somme de parties n'est pas égale au total). Ces problèmes nous ont poussé à rechercher d'autres méthodes pour répondre aux objectifs mentionnés à la fin de la section précédente.

Il convient de souligner que nous décrivons ici l'application de la nouvelle méthode au niveau des particuliers, mais qu'en pratique, les données au niveau des particuliers sont agrégées au niveau du ménage et que ce sont donc les enregistrements concernant les ménages qui sont utilisés pour produire les estimations.

Dans le cas de l'ancien système de calcul des estimations de l'EPA, on crée une matrice de régression X pour pouvoir

utiliser la régression pour la pondération. Chaque personne qui figure dans l'échantillon correspond à une ligne de la matrice X. Chaque colonne de X correspond à un total de contrôle; par exemple si la colonne c correspond aux hommes de 20 à 24 ans, la valeur qui figure à l'intersection de la ligne i et de la colonne c sera égale à 1 si la personne i est un homme de 20 à 24 ans et sera égale à 0 autrement (la situation est similaire pour les colonnes qui correspondent aux régions géographiques). Pour des renseignements plus détaillés sur les méthodes d'estimation utilisées pour l'Enquête sur la population active, consulter Gambino, Singh, Dufour, Kennedy et Lindeyer (1998).

Pour tirer parti de l'échantillon qui est commun d'un mois à l'autre, on ajoute à la matrice X des colonnes dont les éléments sont définis de façon que, lorsque l'on applique les poids finals du mois courant aux éléments de chaque nouvelle colonne, le total soit une estimation composite basée sur le mois précédant, c'est-à-dire que l'estimation composite du mois précédant soit utilisée comme total de contrôle (strictement parlant, le total de contrôle est fondé sur les poids qui reflètent la population du mois courant). Comme nous l'avons mentionné dans l'introduction, il existe plusieurs moyens de définir les nouvelles colonnes, selon les objectifs poursuivis. Nous ne présentons plus bas que les solutions dont la mise en œuvre a été proposée.

Une nouvelle colonne type correspond à l'emploi dans une branche d'activité donnée, disons l'agriculture. Si l'on s'intéresse principalement aux estimations de nouveau, la méthode qui suit proposée pour former les colonnes donne de bons résultats. Représentons par M et U l'échantillon apparié (commun) et non apparié (nouvelles unités d'échantillonnage), respectivement. Pour la personne i, aux périodes t-1 et t, représentons par $y_{i,t-1}$ et $y_{i,t}$ les variables indicateurs dont la valeur est égale à 1 chaque fois que la personne travaille en agriculture. Alors, posons

$$x_i^{(L)} = \begin{cases} \bar{y}'_{t-1} & \text{si } i \in U \\ y_{i,t-1} & \text{si } i \in M, \end{cases}$$

où $\bar{y}'_{t-1}$ représente l'estimation composite de la proportion de personnes qui travaillaient en agriculture le mois précédant; en pratique, nous utilisons $\bar{y}'_{t-1} = \hat{Y}'_{t-1}/P_{15+}$, où P_{15+} représente la population de 15 ans et plus. Le total de contrôle correspondant est l'estimation du nombre de personnes qui travaillaient en agriculture le mois précédant, c'est-à-dire $\hat{Y}'_{t-1}$. Donc, en dernière analyse, la somme pondérée finale des éléments de la nouvelle colonne sera égale à l'estimation du mois précédant. Cette situation équivaut presque à forcer les poids du mois courant, appliqués aux valeurs du mois précédant pour l'échantillon commun, à reproduire l'estimation de l'emploi en agriculture du mois précédant (après l'application d'un facteur de correction de 5/6). Nous utilisons l'exposant L pour rappeler que l'objectif est ici d'améliorer les estimations de niveau (*level*).

Si l'on s'intéresse principalement à l'estimation de la variation, la méthode de création de nouvelles colonnes dans X qui suit donne de bons résultats :

$$x_i^{(C)} = \begin{cases} y_{i,t} & \text{si } i \in U \\ y_{i,t} + R(y_{i,t-1} - y_{i,t}) & \text{si } i \in M, \end{cases}$$

où R est un rapport de correction qui permet de tenir compte du fait que les cinq sixièmes de l'échantillon sont communs de mois en mois. La valeur $R = \Sigma_{\text{tout}} w_i / \Sigma_{\text{apparié}} w_i$ est celle qui est utilisée dans le système de production. Par souci de commodité, nous utilisons ici $R = 6/5$, puisque, en pratique, la différence entre les deux est faible, car on applique des méthodes visant à équilibre les poids par groupe de renouvellement (par exemple, la correction pour la non-réponse est faite séparément, selon le groupe de renouvellement). Comme précédemment, le total de contrôle correspondant est l'estimation du nombre de personnes employées en agriculture le mois précédant. Par application des poids finals aux éléments de la colonne de la matrice X et totalisation, nous obtenons l'égalité

$$\hat{Y}'_{t-1} = \hat{Y}'_t - \hat{\Delta}^{M,f}_{t-1,t},$$

ce qui signifie, en toutes lettres, que l'estimation du mois précédant est égale à la différence entre l'estimation du mois courant et une estimation $\hat{\Delta}$ de $Y_t - Y_{t-1}$ fondée sur l'échantillon commun. Nous utilisons l'exposant f dans $\hat{\Delta}$ pour rappeler que l'estimation de la variation se fonde sur l'application des poids finals après l'estimation composite. Pour ce qui est des poids «avant l'estimation composite», il est facile de démontrer que, dans le cas d'une seule variable :

$$\hat{Y}'_t = (1-b)\hat{Y}_t + b(\hat{Y}'_{t-1} + \hat{\Delta}^{M}_{t-1,t}),$$

où b est le coefficient de régression et $\hat{\Delta}$ est l'estimation de la variation fondée sur les poids de départ. Le cas plus général où des variables auxiliaires sont incluses est décrit par Fuller et Rao (2001, équation 2.3).

Des études antérieures montrent que l'utilisation de contrôles axés sur le niveau (*level*), appelés ici contrôles L, produit de meilleures estimations du niveau pour les variables de contrôle ajoutées à la matrice X. Pareillement, l'ajout de contrôles axés sur la variation (*change*), appelés ici contrôles C, produit de bonnes estimations de la variation pour les variables ajoutées à la matrice. Singh, et coll. (1997, 2001) décrivent les gains d'efficacité observés pour les estimations du niveau et de la variation axées sur la variation (contrôles C), et renvoient à des résultats antérieurs sur les estimations axées sur le niveau (contrôles L).

Au départ, il avait été proposé d'élaborer un système d'estimation qui s'appuierait uniquement sur des contrôles axés sur la variation (C). Cependant, certains auteurs ont exprimé des réserves à l'idée d'utiliser un tel système, puisque les estimations du niveau sont tout aussi importantes (elles jouent, par exemple, un rôle essentiel

dans le programme d'assurance-emploi de l'administration fédérale). Fuller et Rao (2001) ont résumé ces réserves.

En principe, nous pourrions inclure à la fois les contrôles L et C dans la régression, mais le grand nombre de colonnes que nous obtiendrions dans la matrice X aurait notamment pour conséquence indésirable d'augmenter le nombre de poids finals extrêmes, y compris des poids négatifs. Pour éviter ceci, nous serions obligés de limiter le nombre de branches d'activité incluses dans l'estimateur. Wayne Fuller (voir Fuller et Rao 2001) a proposé une autre solution qui permet d'inclure les branches d'activité présentant le plus d'intérêt, tout en faisant un compromis entre l'amélioration des estimations du niveau, d'une part, et de la variation, d'autre part. Sa solution consiste à prendre une combinaison linéaire de la colonne L et de la colonne C pour une branche d'activité donnée et de s'en servir comme nouvelle colonne dans la matrice X, c'est-à-dire à utiliser

$$x_i = (1 - \alpha)\, x_i^{(L)} + \alpha\, x_i^{(C)}.$$

Les variables originales axées sur le niveau ou sur la variation représentent des cas spéciaux du compromis de Fuller.

Choix de α : Fuller et Rao (2001) ont montré, en s'appuyant sur certaines hypothèses raisonnables que des valeurs de α telles que 0,65 et 0,75 produisent des estimations satisfaisantes tant du niveau que de la variation. Le choix de la valeur de α dépend de la variable étudiée (plus précisément de sa corrélation dans le temps) et de l'importance relative du niveau par rapport à la variation.

Nos études (voir l'annexe 1) montrent que, pour les deux variables les plus importantes, c'est-à-dire les personnes occupées et les chômeurs, les meilleures valeurs de α pour les estimations du niveau sont 0,39 et 0,24, respectivement. Les valeurs correspondantes pour l'estimation de la variation sont 0,99 et 0,81, respectivement. Il faut donc manifestement faire un compromis entre l'amélioration des estimations du niveau et des estimations de la variation.

Pour déterminer quelles valeurs de α étudier, nous avons obtenu des valeurs intermédiaires de α en calculant pour chaque variable la moyenne des valeurs basées sur le niveau et des valeurs basées sur la variation; nous avons obtenu ainsi les valeurs approximatives de 0,7 et 0,52 pour les personnes occupées et les chômeurs, respectivement. Les résultats fondés sur les valeurs de $\alpha = 1$ et $\alpha = 0{,}75$ ayant déjà été obtenus, nous y avons ajouté les résultats pour $\alpha = 0{,}67$ et $\alpha = 0{,}6$. Compte tenu des résultats présentés plus loin, qui n'indiquent aucune différence appréciable que l'on utilise 0,6, 0,67 ou 0,75 comme valeur de α, nous avons décidé d'utiliser la valeur $\alpha = 2/3$ dans le système de production.

3. CARACTÉRISTIQUES, PROPRIÉTÉS ET RÉSULTATS

Nous résumons ici certaines caractéristiques et propriétés de l'estimateur composite de régression. Certains résultats graphiques et numériques sont présentés plus loin, à la section 3.1.

Application dans les systèmes. Un avantage important de l'estimateur tient au fait qu'il peut être appliqué dans l'ancien système d'estimation de l'EPA de façon simple, puisque, comme nous l'avons décrit plus haut, il suffit essentiellement d'agrandir la matrice de régression. Ce facteur a joué un rôle important dans notre décision d'étudier et, en bout de ligne, d'adopter l'estimation composite, car autrement, la transformation du système aurait coûté beaucoup plus cher.

Pondération. Dans le cas de l'estimateur *A-K*, la pondération pour satisfaire aux totaux de population de contrôle et la pondération pour tenir compte de l'estimation composite représentent deux étapes distinctes. Par contre, dans le cas de l'estimateur composite de régression, la pondération se fait en une seule étape, c'est-à-dire en même temps que la pondération pour satisfaire aux totaux de population selon le groupe âge-sexe et la région géographique de contrôle. À titre d'illustration, nous décrivons à l'annexe 3 la façon dont la matrice de régression serait agrandie si on ajoutait les éléments $x_i^{(C)}$ définis à la section 2. L'ajout des éléments $x_i = (1 - \alpha)x_i^{(L)} + \alpha\, x_i^{(C)}$ est similaire. Cette façon de procéder assure non seulement la cohérence des estimations décrites au paragraphe suivant, mais retient aussi les avantages qu'offrent les contrôles appliqués à l'estimateur de régression habituel, c'est-à-dire, ici, les contrôles selon le groupe âge-sexe et la région géographique.

Cohérence. Les estimations demeurent cohérentes, puisque la pondération en fonction des totaux de contrôle selon le groupe âge-sexe et la région géographique a lieu en même temps que la pondération en fonction des valeurs de contrôle pour l'estimation composite. En particulier, la somme des parties est égale au total; par exemple, personnes occupées + chômeurs = population active. Dans le cas d'autres méthodes d'estimation composite, la cohérence est assurée par d'autres moyens qui nécessitent une étape distincte ou une certaine forme de compromis.

Gains d'efficacité. Pour les variables ajoutées comme totaux de contrôle, la méthode produit des gains d'efficacité considérables aussi bien dans le cas de l'estimation du niveau que de la variation. Si $\alpha = 1$, pour les estimations de la variation, les gains d'efficacité peuvent être spectaculaires; en choisissant une valeur plus faible de α, nous obtenons un gain plus important pour les estimations du niveau, mais nous réduisons l'importance du gain pour les estimations de la variation. Certains résultats obtenus en choisissant $\alpha = 2/3$ sont présentés à la section 3.1.

Désaisonnalisation. Les séries chronologiques de données sur l'emploi pour diverses branches d'activité sont examinées de près par les utilisateurs tant internes qu'externes des données provenant de l'Enquête sur la population active. L'un des avantages importants du gain d'efficacité susmentionné est que plusieurs séries que l'on ne pouvait désaisonnaliser auparavant peuvent l'être

maintenant. Autrement dit, l'estimation composite augmente suffisamment le rapport signal-bruit pour rendre la désaisonnalisation efficace. Une conséquence connexe de l'estimation composite appréciée par les utilisateurs des données tient au fait que plusieurs estimations publiées antérieurement sous forme de moyenne mobile de trois mois peuvent désormais l'être sous forme d'estimations mensuelles.

Écarts systématiques entre les estimations composites et ordinaires du niveau. Théoriquement, les espérances, calculées sur tous les échantillons possibles, devraient être les mêmes pour les estimateurs ordinaire et composite, ce qui les rend tous deux non biaisés ou presque non biaisés. On s'attendrait donc à ce que les estimations du niveau obtenues en se servant des deux estimateurs se recoupent au cours du temps. Malheureusement, cela ne se concrétise pas en pratique car, si l'on tient compte des conditions réelles d'enquête, l'estimateur composite et l'estimateur ordinaire n'ont pas la même valeur prévue; consulter, par exemple, Bailar (1975) et Kumar et Lee (1983) pour des résultats sur les estimateurs composites K et AK, respectivement. Kumar et Lee démontrent ce fait en dérivant des expressions explicites pour la valeur prévue de l'estimateur ordinaire et de l'estimateur composite AK. Les échantillons apparié et non apparié diffèrent à cause de différences entre les taux de non-réponse et le mode de collecte de données (par exemple, interview sur place par opposition à téléphonique, interview centralisée par opposition à décentralisée). En pratique, les unités du «nouvel» échantillon ont un taux de non-réponse plus élevé et les ménages manquants ont tendance à être de plus petite taille et à avoir un taux d'emploi plus élevé que les ménages répondants. Comme l'estimateur ordinaire et l'estimateur composite produisent des poids différents pour les échantillons apparié et non apparié, leur valeur prévue sera différente. Par conséquent, les séries chronologiques obtenues au moyen des deux estimateurs peuvent présenter des différences systématiques. En pratique, ces différences sont souvent masquées par la variation d'échantillonnage, mais elles sont évidentes pour des séries plus précises, comme celles sur les personnes occupées dans les grandes provinces comme l'Ontario et au Canada. Nos résultats pour les personnes occupées concordent avec ceux décrits par Bailar (1975) pour la *Current Population Survey* aux États-Unis; autrement dit, pour les personnes occupées, les estimations composites ont tendance à être plus faibles que celles obtenues au moyen de l'estimateur ordinaire. Pour les chômeurs, en Ontario, l'écart entre les deux types d'estimation a tendance à être nettement plus faible.

Nous recherchons à l'heure actuelle un moyen de réduire les écarts systématiques entre les estimations pour des groupes de renouvellement différents. Plus précisément, nous étudions la possibilité de faire une correction de la pondération pour le nombre de ménages de tailles différentes selon le groupe de renouvellement pour tenir compte du fait que les ménages de petite taille sont sous-représentés dans le nouvel échantillon. Une telle méthode donnerait de bons résultats pour les estimateurs composites et l'estimateur de régression ordinaire, et atténuerait probablement l'écart entre ces estimateurs.

3.1 Résultats empiriques

Emploi et chômage au niveau provincial. Le graphique 1 montre l'emploi total au niveau provincial de 1987 à 1998 pour l'Ontario. La série chronologique d'estimations composites pour les quatre valeurs de α, c'est-à-dire pour 0,6, 0,67, 0,75 et 1, se comporte de façon similaire. Ces graphiques montrent clairement qu'il y a une variation du niveau pour cette série dans le cas de l'estimation composite — le nombre estimé de personnes occupées est plus faible. Les versions désaisonnalisées de la série de données sur l'emploi en Ontario fondées sur l'estimateur ordinaire GREG et sur l'estimateur composite pour $\alpha = 0{,}67$ sont illustrées au graphique 2.

Le graphique 3 donne une comparaison de l'estimation ordinaire GREG du chômage en Ontario à l'estimation composite par régression pour $\alpha = 0{,}67$. L'effet de l'estimation composite sur cette variable est manifestement moins prononcée que sur les variables liées à l'emploi.

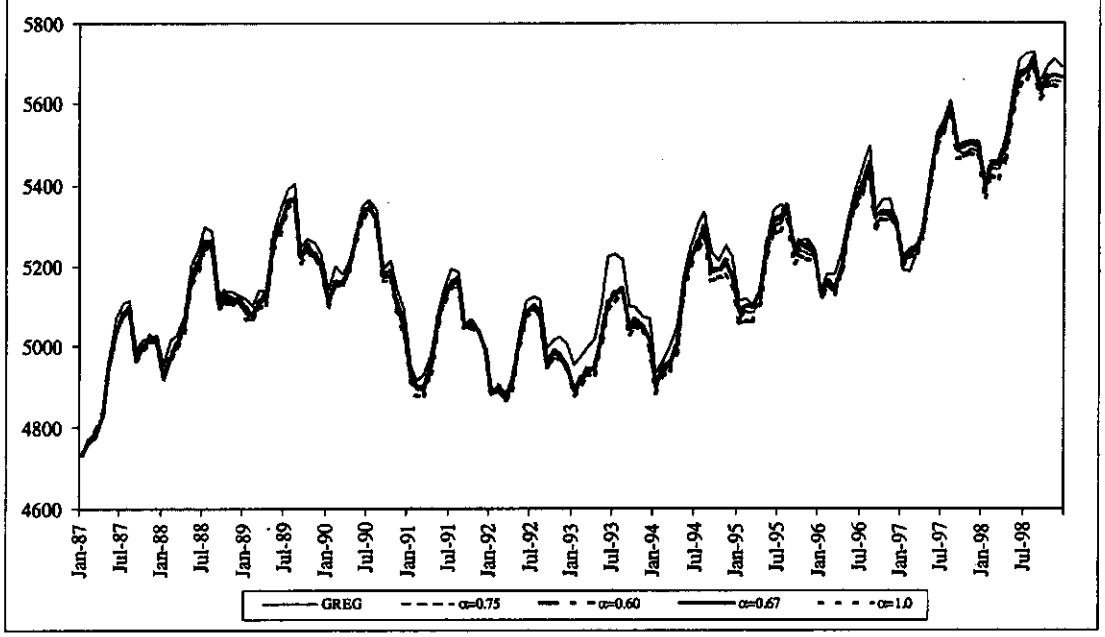
Le graphique 4 donne une comparaison de la variation annuelle de l'emploi en Ontario pour les deux estimateurs. Chaque point de la série correspond à l'écart entre l'emploi durant le mois m de l'année y et le mois m de l'année $y - 1$. Par exemple, le premier point est le chiffre d'emploi en janvier 1988 dont est soustrait l'emploi en janvier 1987. La série d'estimations composites est de toute évidence plus lisse, particulièrement pour la deuxième moitié de la période de 12 mois.

Emploi selon la région infraprovinciale. Le graphique 5 donne une comparaison de l'estimation ordinaire et de l'estimation composite de l'emploi pour $\alpha = 0{,}67$ pour une région économique de l'Ontario. Les résultats pour d'autres régions infraprovinciales sont comparables. Le comportement des séries d'estimations ordinaires et composites étant fort semblable, l'effet de l'estimation composite n'est ni bénéfique ni néfaste. Pour les totalisations spéciales, le système de calcul des estimations de l'EPA est suffisamment souple pour permettre à l'utilisateur d'ajouter des valeurs de contrôle au niveau de la région économique, au besoin. Il existe déjà un contrôle pour l'estimation de la population totale selon la région économique.

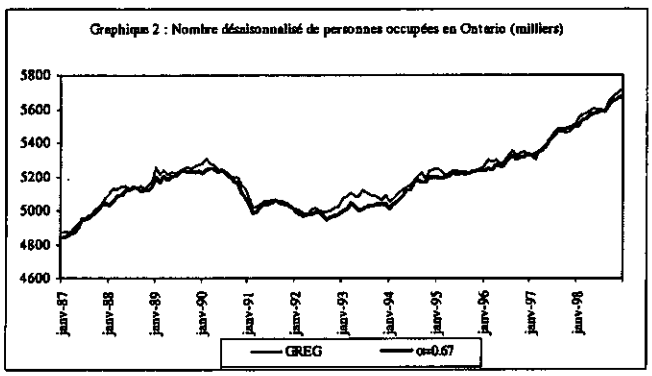
Emploi selon la branche d'activité et désaisonnalisation. Nous avons comparé les estimations composites aux estimations ordinaires par régression pour 16 branches d'activité. Les graphiques 6A à 6D montre les résultats pour deux d'entre elles en Ontario. Bien qu'elles ne soient pas incluses dans ces graphiques, de nouveau, les quatre valeurs de α produisent des séries d'estimations composites qui se comportent en général de façon comparable, même si la série obtenue pour $\alpha = 1$ s'écarte

parfois de façon significative des autres. La série d'estimations composites a tendance à être moins instable que celle obtenue par régression. Cette différence est particulièrement évidente pour la série désaisonnalisée sur le commerce que nous avons incluse ici parce qu'elle illustre le cas le plus extrême. Pour cette série, le comportement des estimations par régression obtenues au départ les premières années est difficile à expliquer d'un point de vue technique, qu'il s'agisse de la série désaisonnalisée ou de la série non corrigée. Le comportement de la série sur la fabrication est plus typique des 14 branches d'activité restantes.

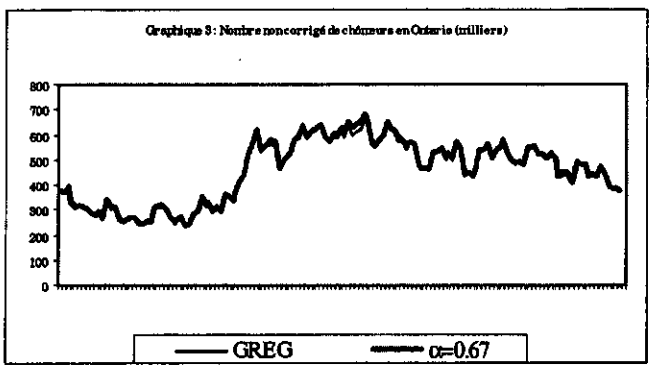
Graphique 1 : Nombre non corrigé de personnes occupées en Ontario (milliers)



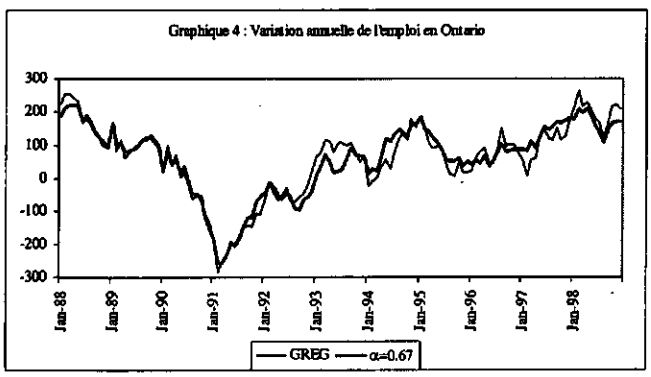
Graphique 2 : Nombre désaisonnalisé de personnes occupées en Ontario (milliers)



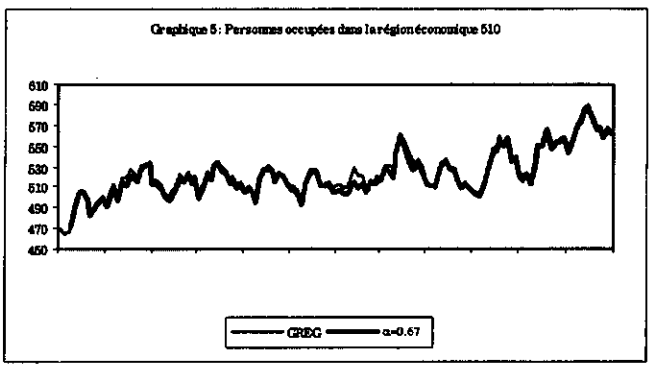
Graphique 3 : Nombre non corrigé de chômeurs en Ontario (milliers)



Graphique 4 : Variation annuelle de l'emploi en Ontario



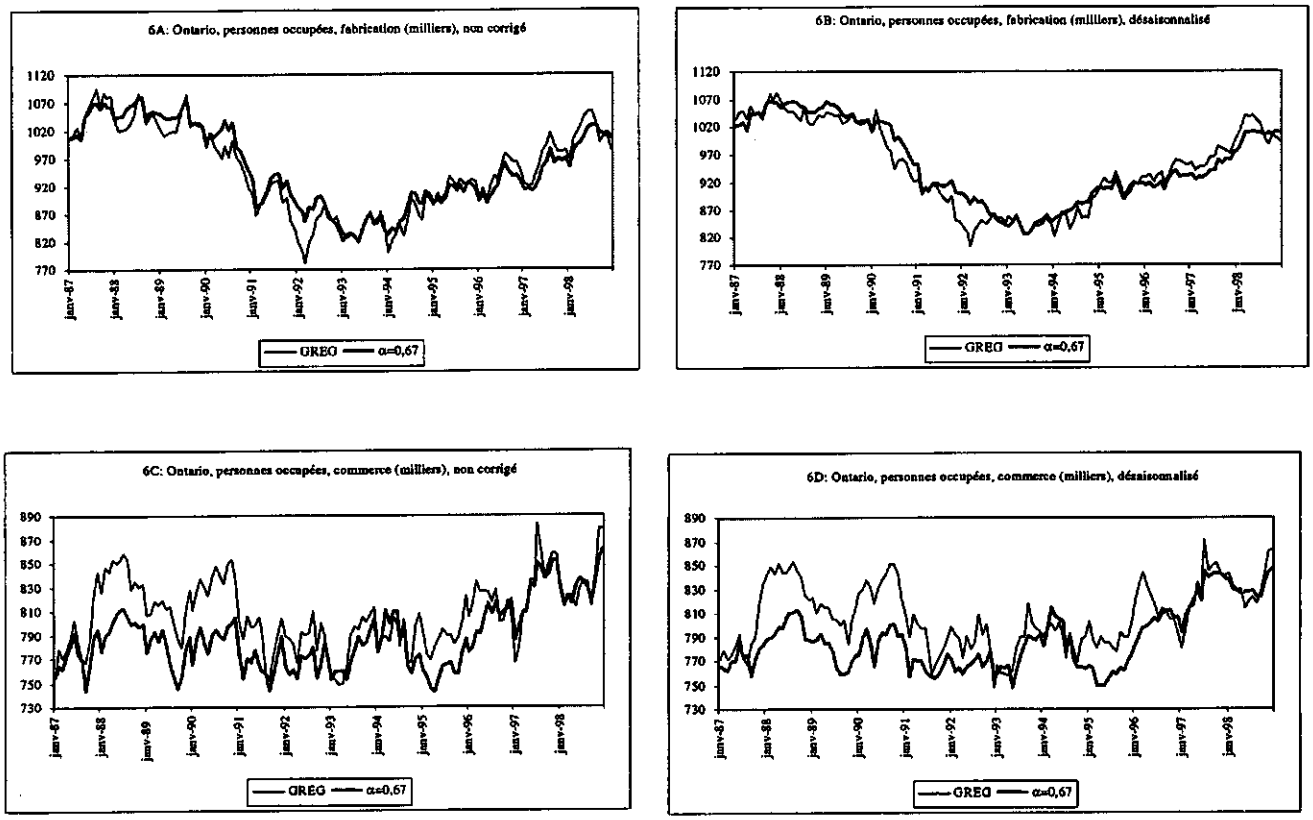
Graphique 5 : Personnes occupées dans la région économique 510



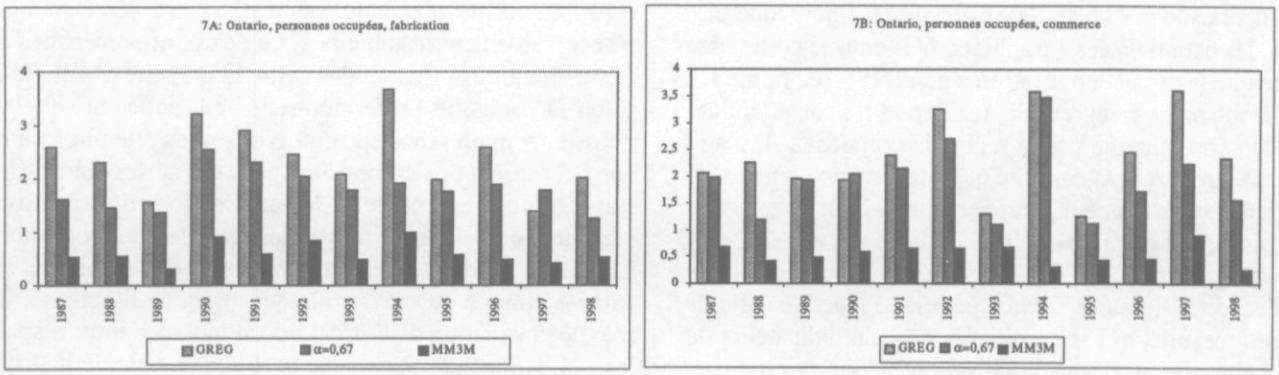
La comparaison des séries désaisonnalisées (graphique 6D) et non désaisonnalisées (graphique 6C) pour le commerce montre que la désaisonnalisation a un effet assez faible sur la série obtenue par régression, mais modifie considérablement la série obtenue par la méthode composite. Il s'agit d'une illustration de la capacité qu'a l'estimation composite d'accroître suffisamment le rapport signal-bruit pour rendre la désaisonnalisation efficace.

Le programme de désaisonnalisation appliqué aux données de l'Enquête sur la population active calcule diverses mesures qui sont utilisées comme indicateurs de l'efficacité de la désaisonnalisation. Certaines de ces mesures, présentées à l'annexe 2, montrent que, pour l'emploi en Ontario durant la période de deux ans allant de 1996 à 1998, le recours à l'estimation composite permet de désaisonnaliser convenablement les données pour un plus grand nombre de branches d'activité. Les résultats pour d'autres provinces et pour le Canada sont similaires.

Mesure de la stabilité. Pour plusieurs séries de données importantes, on publiait par le passé les moyennes mobiles de trois ans au lieu des estimations mensuelles, parce que la forte variabilité d'échantillonnage de ces séries rendait inacceptablement instable la série de données mensuelles. Les estimations provinciales selon la branche d'activité et selon la catégorie professionnelle sont celles auxquelles nous nous intéressons surtout. Nous nous attendions à ce que les estimations composites pour ces séries soient plus stables, donc permettent la publication d'estimations mensuelles au lieu de moyennes mobiles de trois mois. L'indice d'instabilité, qui est une mesure de la stabilité, est calculé comme suit. Pour chaque branche d'activité, on calcule la variation de l'emploi d'un mois à l'autre d'après les estimations désaisonnalisées. Puis, on calcule la différence entre les estimations consécutives de la variation. La valeur absolue de cette «variation de la variation» est exprimée en pourcentage de l'estimation mensuelle totale correspondante. Enfin, on calcule la moyenne de ce pourcentage sur l'année complète. Les valeurs de cette mesure sont grandes lorsqu'une série présente de nombreux mouvements consécutifs en directions opposées, ce qui est un signe d'instabilité.



Graphique 6. Emploi selon la branche d'activité



Graphique 7. L'indice d'instabilité

L'indice d'instabilité a été calculé pour seize branches d'activité. Nous présentons ici deux graphiques (7A et 7B), pour les branches de la fabrication et du commerce en Ontario, où sont comparés l'estimateur ordinaire et l'estimateur composite mensuel pour α= 0,67 (GREG) la moyenne mobile de trois mois MM3M pour l'estimateur ordinaire. Pour la branche de la fabrication, la valeur moyenne de l'indice pour les estimateurs ordinaire, composite et de la moyenne mobile est de 2,4, 1,8 et 0,60, respectivement. Pour le commerce, les valeurs correspondantes sont 2,4, 1,9 et 0,55. Pour toutes les branches d'activité, l'instabilité des estimations composites est habituellement comprise entre celles des estimations mensuelles ordinaires et des estimations de la moyenne mobile de trois mois. À l'occasion, pour certaines années de référence, les estimations composites sont moins instables que les moyennes mobiles de trois mois ou plus instables que les estimations mensuelles ordinaires, mais, en général, leur instabilité est comprise entre celles des estimations mensuelles ordinaires et des moyennes mobiles de trois mois. Nous constatons aussi que, si les estimations mensuelles sont extrêmement instables, la série d'estimations composites a tendance à être plus stable. Les estimations mensuelles par régression ne concurrencent les estimations composites que si l'indice d'instabilité est faible pour les deux méthodes.

Suite à l'introduction de l'estimation composite, le calcul des moyennes mobiles de trois mois a été abandonné et remplacé par celui des estimations mensuelles, plus souhaitables pour les séries sur les branches d'activité.

Estimations de la variance. Pour les variables ajoutées à titre de totaux de contrôle, comme l'emploi selon la branche d'activité, il est possible de réaliser des gains considérables d'efficacité au niveau provincial, l'efficacité étant définie ici par Var(greg)/Var(composite). Pour la plupart des branches d'activité, les gains sont habituellement de l'ordre de 10 % à 20 % mais ils peuvent aller jusqu'à 40 %. Un gain d'efficacité de 40 % équivaut, par exemple, à faire baisser un coefficient de variation de 15 % jusqu'à 12,7 % ou un coefficient de variation de 10 % jusqu'à 8,5 %. Pour les estimations des niveaux provinciaux d'emploi et de chômage, nous observons des gains d'efficacité plus modestes, habituellement de l'ordre de 5 % à 10 %. Pour les estimations de la variation mensuelle, les gains d'efficacité sont plus importants pour les variables contrôlées – habituellement deux fois plus élevés que ceux observés pour les estimations du niveau.

Pour les variables non contrôlées, l'estimation composite n'a que peu d'effet, voire aucun, sur l'efficacité, à moins que la variable ne soit fortement corrélée à une variable contrôlée. Par exemple, au niveau provincial, on constate une amélioration de l'efficacité des estimations du nombre d'hommes occupés, parce que cette variable est corrélée au nombre total de personnes occupées, qui est une variable contrôlée. Par contre, pour l'estimation de l'emploi selon la région économique infraprovinciale, on ne constate aucun gain ni perte d'efficacité.

4. TRAITEMENT DES DONNÉES MANQUANTES

Par définition, les variables x_i comprennent des données du mois courant et du mois précédent, ce qui pose des difficultés lorsque l'on ne dispose de données que pour l'un des deux mois pour une personne donnée de l'échantillon commun. Cette situation peut être due à la non-réponse lors de l'un des deux mois, ou à un déménagement ou une modification par rapport au champ d'observation dans l'intervalle de deux mois. Les différents cas possibles sont présentés dans le diagramme qui suit, où R représente une réponse, X représente une non-réponse et O représente une unité hors du champ d'observation.

	A	B	-	C	D
Mois *t*	X XX...	R RR...	RRR...	RRR...	OOO...
Mois *t*-1	R RR...	XXX...	RRR...	OOO...	RRR...

Dans tous ces cas, à savoir A, B, C et D, l'objectif est de trouver une solution telle que $\Sigma_{i \in S} w_{it} x_{it}$ soit encore un estimateur de Y_{t-1}. Nous énonçons les deux objectifs qui suivent pour le traitement des cas où des données manquent pour l'un ou l'autre mois de l'échantillon commun :

i) retenir autant de réponses valides que possible, c'est-à-dire rejeter l'option de supprimer une unité du processus d'estimation;

ii) élaborer une méthode d'imputation qui ne sous-estime pas l'estimation de la variation de façon significative.

Dans le cas de la non-réponse, on distingue deux situations : le cas A, où un ménage a répondu le mois précédent, mais non le mois courant et le cas B, où la situation est inverse. Dans ce qui suit, i représente une personne dans un ménage affecté.

Cas A : Remplacer y_{it} par $\hat{y}_{it}$. La substitution peut se faire de plusieurs façons. Une méthode simple consiste à remplacer y_{it} par la réponse correspondante donnée le mois précédent, c'est-à-dire $y_{i,t-1}$. Nous avons utilisé cette méthode aux premières étapes de l'étude, mais l'avons rejetée plus tard car elle peut biaiser (sous-estimer) l'estimation de la variation de façon significative. Pour le système d'estimation de l'EPA, nous avons décidé d'utiliser les caractéristiques démographiques et d'emploi connues de la personne pour le mois précédent pour créer des catégories d'imputation puis nous avons appliqué la méthode d'imputation hot deck (c'est-à-dire, données du mois courant) pour obtenir $\hat{y}_{it}$. Une autre solution consisterait à utiliser une certaine sorte de moyenne.

Cas B : La procédure est analogue; autrement dit, si la valeur du mois précédent manque, on forme des catégories d'imputation en se servant des données recueillies pour le mois t et on se sert de données provenant des unités répondantes au mois $t-1$ pour trouver l'enregistrement donneur.

Si l'unité i a déménagé ou que sa situation par rapport au champ d'observation a été modifiée, les cas qui suivent peuvent se présenter.

Cas C : Supposons que l'unité i était hors du champ d'observation à la période $t-1$, mais dans le champ à la période t (par exemple, une personne qui vient tout juste d'avoir 15 ans ou un immigrant qui vient d'arriver). Alors, la contribution de l'unité i devrait être 0 à la période $t-1$ et y_{it} à la période t. Donc, nous posons $x_{it} = 0$ puisque $\sum w_{it} x_{it}$ devrait estimer Y_{t-1}.

Cas D : Inversement, supposons que l'unité i était dans le champ d'observation le mois précédent et est maintenant hors du champ. Ce scénario inclut, par exemple, les personnes qui ont quitté le pays, sont entrées dans l'armée ou sont décédées. Ces unités devraient être abandonnées, puisque la population cible est la population dans le champ d'observation à la période t (et que l'objectif ultime est d'estimer Y_t). Puisque nous échantillonnons des logements, mais que nous recueillons des données sur les personnes qui habitent ces logements, deux autres situations peuvent se présenter à cause de l'emménagement de nouvelles personnes dans les logements échantillonnés ou du déménagement de personnes hors de ces logements.

Cas i) : Supposons que l'unité i faisait partie de la population aux deux périodes, mais ne résidait dans un logement échantillonné qu'à la période t (c'est-à-dire, une personne qui a déménagé d'un logement non échantillonné dans un logement échantillonné). Alors, sa situation à la période $t-1$ est inconnue, c'est-à-dire que $y_{i,t-1}$ est inconnu. Pour tous les cas de ce genre, comme dans celui de la non-réponse, nous pouvons imputer une valeur $\hat{y}_{i,t-1}$ pour $y_{i,t-1}$ à partir d'un donneur compris dans l'échantillon ou par calcul d'une moyenne d'échantillon. Dans le cas de l'EPA, nous utilisons l'imputation hot deck.

Cas ii) : Enfin, considérons le cas où l'unité i faisait partie de l'échantillon à la période $t-1$, mais avait déménagé dans un logement non échantillonné à la période t. Puisque l'échantillon de l'EPA est un échantillon de logements plutôt qu'un échantillon de personnes, cette unité devrait simplement être supprimée lors du calcul des estimations de Y_t.

5. CONCLUSION

L'estimateur composite permet d'atteindre tous les objectifs établis au début du projet et résumés dans l'introduction. Il produit des estimations du niveau et de la variation plus efficaces que celles obtenues au moyen de l'estimateur de régression ordinaire, tout en satisfaisant aux contraintes opérationnelles et de cohérence des données. Dans l'ensemble, l'utilisation de l'estimateur composite en choisissant la valeur $\alpha = 2/3$ a un effet modéré sur les nombreuses séries chronologiques produites d'après les données de l'Enquête sur la population active. Lorsque l'effet est important, comme dans le cas de la série sur le commerce en Ontario, la nouvelle série a tendance à être plus compréhensible pour les spécialistes du domaine. Ce genre d'amélioration permet en outre d'expliquer plus facilement les estimations fondées sur les données de l'EPA aux utilisateurs des données et aux membres des médias.

Les estimations composites présentent d'autres caractéristiques que les utilisateurs trouveront fort intéressantes. Comme l'estimation composite réduit la variance, il est possible de publier des estimations mensuelles dans de nombreux cas où il fallait se limiter à des moyennes mobiles de trois mois auparavant. De surcroît, l'estimation composite permet de désaisonnaliser convenablement un plus grand nombre de séries.

L'application de l'estimation composite aux données de l'EPA est une première étape importante. Des études en vue d'améliorer le traitement des erreurs non dues à l'échantillonnage sont en cours et les résultats pourront être intégrés dans le système de pondération et d'estimation n'importe quand. Le système offre le grand avantage d'être très souple. Par exemple, comme on peut modifier facilement la valeur de α, on prévoit comparer les résultats pour toute une gamme de valeurs de α pour un nombre important de variables. On pourra peut-être aboutir ainsi à un système

permettant d'utiliser des valeurs de α différentes pour des variables de contrôle différentes, en continuant d'utiliser un poids final unique pour chaque enregistrement.

REMERCIEMENTS

Nous aimerions remercier Avi Singh et le Comité consultatif des méthodes statistiques de Statistique Canada pour leurs contributions à ce projet. Nous sommes aussi redevables à plusieurs personnes dont les commentaires sur des versions antérieures, ont permis de grandement améliorer cet article.

ANNEXE 1

Relation entre α, ρ et *(A, K)*. Kumar et Lee (1983) ont déterminé les valeurs optimales de A et K de l'estimateur composite AK pour les estimations du niveau et de la variation en fonction du coefficient de corrélation ρ. Nous avons dérivé une relation approximative entre les valeurs de A et K, ρ et α. Puis, nous avons utilisé cette relation pour déterminer de bonnes valeurs de α pour plusieurs variables. Ces valeurs sont présentées dans les tableaux 1 et 2 pour les estimations du niveau et de la variation, respectivement. Les valeurs de A et K présentées dans les tableaux sont les valeurs optimales pour la valeur correspondante de ρ. Les valeurs de α présentées dans les tableaux concordent avec celles obtenues par Wayne Fuller en se servant d'un modèle AR(1) (communication personnelle). Dans le tableau 2, la valeur de α pour la population active est supérieure à l'unité à cause de l'approximation.

Tableau 1
Valeur de α pour plusieurs variables – niveau

Variable	ρ	A	K	α
Personnes occupées	0,852	0,49	0,8	0,385
Chômeurs	0,58	0,38	0,5	0,242
Population active	0,843	0,48	0,8	0,403
Emploi en agriculture	0,955	0,38	0,8	0,448

Tableau 2
Valeur de α pour plusieurs variables – variation

Variable	ρ	A	K	α
Personnes occupées	0,852	0,1	0,9	0,995
Chômeurs	0,58	0,2	0,6	0,806
Population active	0,843	0,1	0,9	1,009
Emploi en agriculture	0,955	0	0,9	0,959

ANNEXE 2 :
Mesures de la désaisonnalisation pour l'emploi selon la branche d'activité en Ontario

	Valeur F			M7			LISSAGE	
Branche d'activité	greg	α = 0,60	α = 0,75	greg	α = 0,60	α = 0,75	greg	α = 0,60
Agriculture	87,76	120,18	112,7	0,27	0,23	0,24	37,94	45,36
Foresterie	21,34	24,58	23,22	0,5	0,52	0,57	21,76	26,78
Services publics	4,29	3,48	6,8	1,1	1,25	0,82	15,39	15,52
Construction	128,3	275,06	246,93	0,26	0,16	0,17	41,68	57,5
Fabrication	38,22	55,6	69,21	0,37	0,3	0,3	29,02	31,94
Commerce	9,93	15,12	20,35	0,8	0,68	0,53	25,13	34,92
Transport	9,16	8,64	9,69	0,94	0,75	0,7	15,36	23,33
Finances	6,49	8,94	8,84	1,22	0,76	0,77	13,45	19,67
Serv. Professionnels	5,3	12,91	9,81	1,03	0,72	0,76	12,45	19,52
Gestion	14,72	24,98	20,35	0,67	0,52	0,52	16,2	22,17
Éducation	67,37	219,62	214,37	0,33	0,16	0,19	53,25	66,47
Services de santé	8,78	10,73	8,48	0,8	0,66	0,75	16,09	19,92
Information	21,13	52,31	62,94	0,66	0,36	0,35	24,29	33,46
Hébergement	44,85	75,37	78,03	0,36	0,34	0,3	31,89	44,29
Autres services	2,61	13,17	12	1,41	0,75	0,81	18,58	26,27

Description des mesures

Valeur F : Valeur F du test exécuté dans le cadre du programme X11-ARIMA pour déceler l'existence d'une saisonnalité stable. L'existence d'une saisonnalité stable est d'autant plus significative que la valeur F est élevée.

M7 : Mesure qui combine les tests pour les saisonnalités stable et mobile. En général, si M7 est supérieur à 1, la série ne présente aucune saisonnalité décelable; par conséquent, elle ne doit pas être rajustée.

LISSAGE : Écart en pourcentage entre l'écart-type des variations mensuelles dans la série originale et l'écart-type des variations mensuelles dans la série désaisonnalisée. Le lissage que produit la méthode de désaisonnalisation est d'autant plus important que la valeur de l'écart est grande.

ANNEXE 3 :
Application de l'estimation composite par régression aux données de l'EPA
Illustration au moyen de la méthode axée sur la variation

Matrice *X* originale

Indicateurs de groupe âge-sexe								Indicateurs de région						
0	0	1	0	.	.	.	0	0	1	0	.	.	.	0
0	1	0	0	.	.	.	0	0	1	0	.	.	.	0
.	.						.	.	.	.				
.	.						.	.	.	.				
.	.						.	.	.	.				
X_1	X_2	.	.	.	.	.	X_k	X_{k+1}	.	.	.	.	.	X_p

Totaux de population de contrôle

Matrice *X* modifiée pour l'estimation composite quand des $x_i^{(C)}$ sont ajoutés

Indicateurs de groupe âge-sexe								Indicateurs de région							E	C	Agr.	Expl. minière				services
0	0	1	0	.	.	.	0	0	1	0	.	.	.	0	a	0	0	b	.	.	.	0
0	1	0	0	.	.	.	0	0	1	0	.	.	.	0	c	0	d	0	.	.	.	0
.	.						.	.	.	.												
.	.						.	.	.	.												
.	.						.	.	.	.												
X_1	X_2	.	.	.	.	.	X_k	X_{k+1}	.	.	.	.	.	X_p	E'	C'	Agr.	.	.	.	.	S'

E' est l'estimation de l'emploi le mois précédent

Pour les *nouvelles* unités, fixer a, b, c, . . . de façon à indiquer la situation durant le mois courant (par exemple, a = 1 si la personne est occupée, 0 autrement). Pour les unités *appariées*, procéder comme suit :

$a = e_t + (e_{t-1} - e_t)\times 6/5$ où e=1 si la personne est occupée, e=0 autrement.

$d = ag_t + (ag_{t-1} - ag_t)\times 6/5$ où ag=1 si la personne est occupée en agriculture, ag=0 autrement.

Exemples :

i) Supposons que la personne 2 était employée en agriculture le mois précédent et le mois courant. Alors, $e_{t-1} = e_t = 1$ et $ag_{t-1} = ag_t = 1$; par conséquent, $c = 1 - 0 = 1$ et $d = 1 - 0 = 1$.

ii) Supposons que la personne 2 était employée en agriculture le mois précédent et qu'elle l'est dans le secteur minier le mois courant. Alors, $e_{t-1} = e_t = 1$, $ag_{t-1} = 1$ et $ag_t = 0$; donc, $c = 1 - 0 = 1$ et $d = 0 + (1 - 0)^* 6/5 = 1{,}2$.

iii) Supposons que la personne 2 était employée dans le secteur minier le mois précédent et qu'elle est employée en agriculture le mois courant. Alors, $e_{t-1} = e_t = 1$, $ag_{t-1} = 0$ et $ag_t = 1$; donc, $c = 1 - 0 = 1$ et $d = 1 + (0 - 1)^* 6/5 = -0{,}2$.

BIBLIOGRAPHIE

BAILAR, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

CANTWELL, P.J., et ERNST, L.R. (1992). Nouveaux développements dans l'estimation composite pour l'enquête « *Current Population Survey* ». *Recueil : Symposium 92, Conception et analyse des enquêtes longitudinales*, Statistique Canada, 139-149.

FULLER, W.A., et RAO, J.N.K. (2001). Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada. *Techniques d'enquête*, 27, 49-56.

GAMBINO, J.G., SINGH, M.P., DUFOUR, J., KENNEDY, B. et LINDEYER, J. (1998). *Méthodologie de l'Enquête sur la population active du Canada*. Statistique Canada, numéro de catalogue, 71-526.

KUMAR, S., et LEE, H. (1983). Évaluation de l'application d'estimateurs composites à l'Enquête sur la population active du Canada. *Techniques d'enquête*, 9, 196-221.

SINGH, A.C., KENNEDY, B. et WU, S. (2001). Estimation composite par régression pour l'Enquête sur la population active du Canada avec plan de sondage à renouvellement de panel. *Techniques d'enquête*, 27, 35-48.

SINGH, A.C., KENNEDY, B., WU, S. et BRISEBOIS, F. (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 300-305.

Techniques d'enquête, juin 2001
Vol. 27, n° 1, pp. 81-90
Statistique Canada

Estimation de la variance après imputation

JAE-KWANG KIM[1]

RÉSUMÉ

On recourt fréquemment à l'imputation pour compenser la non-réponse partielle. L'estimation de la variance après imputation a suscité de nombreux débats et plusieurs estimateurs ont été proposés. Nous proposons un estimateur de la variance fondé sur un ensemble de pseudo-données créé uniquement pour estimer la variance. L'application des estimateurs type de la variance de données complètes à l'ensemble de pseudo-données produit des estimateurs cohérents dans le cas des estimateurs linéaires pour diverses méthodes d'imputation, y compris l'imputation par la méthode hot deck sans remise et avec remise. Nous illustrons l'équivalence asymptotique de la méthode proposée et de la méthode corrigée du jackknife de Rao et Sitter (1995). La méthode proposée s'applique directement à l'estimation de la variance en cas d'échantillonnage à deux phases.

MOTS CLÉS : Échantillonnage à deux phases; non-réponse partielle; imputation déterministe; imputation aléatoire.

1. INTRODUCTION

L'imputation, c'est-à-dire l'insertion de valeurs pour remplacer les réponses manquantes, est utilisée fréquemment lorsque des données d'enquête manquent. L'un des avantages de l'imputation est sa commodité. Autrement dit, nous pouvons appliquer aux ensembles contenant des données imputées des programmes types conçus pour des données complètes en vue de calculer des estimations ponctuelles. Rubin (1996), Fay (1996) et Rao (1996) ont passé en revue divers problèmes posés par l'imputation.

Toute méthode d'imputation s'appuie sur un modèle. Après avoir précisé ce dernier, nous pouvons recourir à l'imputation déterministe ou à l'imputation aléatoire selon le modèle. Dans le cas de l'imputation aléatoire, on procède à une forme d'échantillonnage probabiliste en vue d'imputer des données pour remplacer les valeurs manquantes. Nous donnons à ce mécanisme aléatoire supplémentaire le nom de mécanisme d'imputation. L'imputation déterministe, quant à elle, ne nécessite pas l'ajout d'un mécanisme aléatoire supplémentaire. Si nous considérons l'ensemble de répondants comme un échantillon aléatoire de l'échantillon original, nous donnons au mécanisme de sélection des répondants le nom de mécanisme de réponse. Ce dernier est souvent considéré comme étant la deuxième phase d'échantillonnage. Pour plus de précisions, consulter Särndal et Swensson (1987).

Le choix d'un modèle et d'une méthode d'imputation appropriés permet de réduire considérablement le biais dû à la non-réponse qui entache les résultats lorsque l'on se sert uniquement des données observées. Cependant, il est reconnu qu'un estimateur de la variance qui traite les données imputées comme s'il s'agissait de données observées est incohérent.

Diverses méthodes ont été proposées pour estimer la variance après l'imputation. Rubin et Schenker (1986) et Rubin (1987) recommandent l'imputation multiple qui crée plusieurs ensembles de données et produit des statistiques complètes pour chaque ensemble contenant des données imputées. L'estimateur de la variance est calculé par combinaison de deux termes, à savoir la variance dans les ensembles de données et la variance entre les ensembles de données. La méthode d'imputation multiple consiste à appliquer un estimateur type de la variance à chaque ensemble de données pour calculer les termes de variance à l'intérieur de l'ensemble de données et des estimateurs ponctuels type pour calculer la variance entre ensembles contenant des données imputées. L'application de cette méthode dépend du choix d'une méthode d'imputation appropriée. Autrement dit, l'imputation devrait remplir les conditions 1 à 3 décrites par Rubin (1987, pages 118 et 119). Ces conditions ne sont pas toujours faciles à satisfaire. (Par exemple, voir Fay 1992.) Même dans Schafer (1997), il n'est pas montré que les méthodes d'imputation multiples décrites sont appropriées au sens de Rubin. Comme le fait remarquer Rao (1996), certaines méthodes d'imputation utilisées couramment, y compris l'imputation par la méthode hot deck et l'imputation par régression, ne conviennent pas.

Rao et Shao (1992) et Rao et Sitter (1995) ont proposé un estimateur jackknife corrigé de la variance. La méthode proposée est applicable à plusieurs méthodes d'imputation et à plusieurs plans d'échantillonnage. Le calcul réel, au moyen d'un logiciel standard applicable à des données complètes n'est pas facile, car des calculs spéciaux sont exécutés pour corriger les valeurs imputées pour chaque pseudo-répétition. En outre, Särndal (1992) a proposé une méthode d'estimation de la variance qui tient compte explicitement du modèle utilisé pour l'imputation.

Essentiellement, la méthode de Rubin produit plusieurs ensembles de pseudo-données pour l'estimation de la variance et consiste à appliquer un estimateur type de la

[1] Jae-Kwang Kim, Westat, 1650 Research Boulevard, Rockville, Maryland, 20850, U.S.A.

variance à chaque ensemble de données pour calculer le terme de variance à l'intérieur de l'ensemble de données, tandis que la méthode de Rao et celle de Särndal consiste à appliquer un estimateur particulier de la variance à l'ensemble contenant des données imputées. Ici, nous proposons une méthode pour créer un ensemble de pseudo-données unique pour l'estimation de la variance. À la section 2, nous présentons la nouvelle méthode dans le cadre de l'échantillonnage à deux phases. À la section 3, nous illustrons l'extension de la méthode proposée à la méthode d'imputation aléatoire. À la section 4, nous étendons la méthode proposée aux plans d'échantillonnage complexes. A la section 5, nous comparons les résultats à ceux de l'estimateur jackknife corrigé de la variance. À la section 6, nous présentons une étude en simulation limitée. Enfin, nous présentons certaines conclusions à la section 7. L'exposé de certaines preuves figure en annexe.

2. UNE MÉTHODE D'ESTIMATION DE LA VARIANCE

Nous décrivons une méthode d'estimation de la variance applicable aux échantillons à deux phases et aux échantillons contenant des données imputées. En plus de l'ensemble de données totalisées, la méthode nécessite un ensemble distinct de données pour l'estimation de la variance. Pour présenter la méthode et illustrer les concepts, considérons un échantillon à deux phases. Supposons que la deuxième phase consiste en un échantillon aléatoire simple de taille r sélectionné à partir de l'échantillon de première phase, qui est un échantillon aléatoire simple de taille n sélectionné à partir d'une population infinie. Supposons que l'estimateur par régression de la moyenne d'une caractéristique y est

$$\hat{\mu}_y = \bar{y}_2 + (\bar{x}_1 - \bar{x}_2)\hat{\beta}, \tag{1}$$

où

$$(\bar{y}_2, \bar{x}_2) = r^{-1}\sum_{i=1}^{r} (y_i, x_i),$$

$$\bar{x}_1 = n^{-1}\sum_{i=1}^{n} x_i,$$

$$\hat{\beta} = \left[\sum_{i=1}^{r} (x_i - \bar{x}_2)^2\right]^{-1} \sum_{i=1}^{r} (x_i - \bar{x}_2)\,(y_i - \bar{y}_2)$$

et que les unités de deuxième phase sont marquées d'un indice dont la valeur varie de un à r. Il est bien connu (par exemple, Cochran 1977, équation 12.51) que la variance de l'estimateur par régression est, approximativement,

$$V\{\hat{\mu}_y\} = \left[n^{-1}\rho^2 + r^{-1}(1-\rho^2)\right]\sigma_y^2, \tag{2}$$

où ρ est la corrélation de population entre y et x et où σ_y^2 est la variance de population de y. Selon la théorie classique de la régression, un estimateur de la variance prend la forme

$$\hat{V}\{\hat{\mu}_y\} = n^{-1}(n-1)^{-1}\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_I)^2 + r^{-1}(r-2)^{-1}\sum_{i=1}^{r} (y_i - \hat{y}_i)^2 \tag{3}$$

où $\hat{y}_i = \bar{y}_2 + (x_i - \bar{x}_2)\hat{\beta}$ pour $i = 1, 2, \ldots, n$, et $\bar{y}_I = n^{-1}\sum_{i=1}^{n}\hat{y}_i$. Notons que $\bar{y}_I$ est une autre façon d'écrire $\hat{\mu}_y$ dans (1).

Supposons que

$$c_r = \left[n(n-1)r^{-1}(r-2)^{-1}\right]^{1/2} \tag{4}$$

et

$$y_i^* = \begin{cases} \hat{y}_i, & i = r+1, r+2, \ldots, n \\ \hat{y}_i + c_r(y_i - \hat{y}_i), & i = 1, 2, \ldots, r. \end{cases} \tag{5}$$

Alors,

$$\hat{V}\{\hat{\mu}_y\} = n^{-1}(n-1)^{-1}\sum_{i=1}^{n} (y_i^* - \bar{y}_I)^2 \tag{6}$$

où $\bar{y}_I$ représente la moyenne des y_i^*, ainsi que la moyenne des $\hat{y}_i$, parce que la somme des termes $y_i - \hat{y}_i$ est nulle. L'équation (6) est la forme opérationnelle de l'estimateur proposé. L'ensemble de données pour l'estimation de la variance contient la pseudo-observation y_i^*.

Dans la mesure où le modèle utilisé pour l'imputation correspond au modèle d'échantillonnage à deux phases, l'équation (6) peut être appliquée à un ensemble contenant des données imputées. Par exemple, si nous supposons que les données manquantes le sont au hasard et que nous utilisons la régression pour imputer les valeurs manquantes de $\hat{y}_i$, alors l'équation (6) est applicable immédiatement. Naturellement, l'imputation par régression ou l'échantillonnage à deux phases peuvent comprendre l'utilisation d'un vecteur x.

3. EXTENSIONS À L'IMPUTATION ALÉATOIRE

Une extension modérée de la méthode décrite à la section 2 nous permet d'estimer la variance d'une moyenne d'échantillon par imputation aléatoire. En fait, d'autres méthodes sont possibles.

Par exemple, supposons que le modèle d'imputation est le modèle de régression

$$y_i = \mathbf{x}_i\beta + e_i \tag{7}$$

où le premier élément de chaque $\mathbf{x}_i$ est égal à 1 et où les e_i sont des variables aléatoires non corrélées $(0, \sigma_e^2)$.

Supposons que le modèle est estimé et que les valeurs imputées sont données par

$$\ddot{y}_i = \hat{y}_i + \ddot{e}_i, \quad i = r+1, r+2, \ldots, n \tag{8}$$

où $\hat{y}_i = \mathbf{x}_i \hat{\boldsymbol{\beta}}$ avec $\hat{\boldsymbol{\beta}} = (\Sigma_{i=1}^{r} \mathbf{x}_i' \mathbf{x}_i)^{-1} \Sigma_{i=1}^{r} \mathbf{x}_i' y_i$ et où $\ddot{e}_i$ est choisi au hasard dans l'ensemble $\hat{\mathbf{e}}_r = \{\hat{e}_i = y_i - \hat{y}_i; i = 1, 2, \ldots, r\}$. L'estimateur de la moyenne y est

$$\hat{\mu}_y = n^{-1} \sum_{i=1}^{n} \ddot{y}_i \tag{9}$$

où $\ddot{y}_i = y_i$ si $i = 1, 2, \ldots, r$.

Si nous choisissons les $\ddot{e}_i$ avec remise et avec probabilité égale à partir de l'ensemble $\hat{\mathbf{e}}_r$, alors la variance $\hat{\mu}_y$ est donnée, approximativement, par

$$V\{\hat{\mu}_y\} = \left[n^{-1}R^2 + (r^{-1} + n^{-2}m)(1 - R^2)\right]\sigma_y^2, \tag{10}$$

où $m = n - r$ et où R^2 représente le carré du coefficient de corrélation multiple entre y et $\mathbf{x}$. L'augmentation de la variance due à l'utilisation de l'imputation aléatoire avec $\ddot{e}_i$, plutôt que l'utilisation de $\ddot{e}_i \equiv 0$, est $n^{-2}m(1 - R^2)\sigma_y^2$.

Par conséquent, un estimateur de la variance de la moyenne d'échantillon imputée est donné par (6) où le c_r de (4) est donné par

$$c_I = \left[n(n-1)(r^{-1} + n^{-2}m)(r-p)^{-1}\right]^{1/2}, \tag{11}$$

et p représente la dimension de $\boldsymbol{\beta}$. Nous avons

$$\hat{V}\{\hat{\mu}_y\} = n^{-1}(n-1)^{-1} \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_I)^2 + (r^{-1} + n^{-2}m)(r-p)^{-1} \sum_{i=1}^{r} (y_i - \hat{y}_i)^2 \tag{12}$$

où $\bar{y}_I = \Sigma_{i=1}^{n} \hat{y}_i$. L'estimateur de la variance calculé en utilisant c_I de l'équation (11) est un estimateur de la variance non conditionnelle, c'est-à-dire la moyenne sur tous les échantillons possibles contenant des données imputées. Les calculs de (10) et (12) figurent à l'annexe A.

Pour considérer une autre méthode d'estimation de la variance, nous supposons que l'on applique une méthode de sélection aléatoire pour l'imputation, mais n'imposons aucune contrainte, si ce n'est que les probabilités de sélection sont inversement proportionnelles à la probabilité que la valeur y réponde. En outre, nous enregistrons le nombre de fois qu'une valeur $\hat{e}$ est utilisée comme donneur dans l'imputation.

Supposons que

$$y_i^* = \begin{cases} \hat{y}_i & i = r+1, r+2, \ldots, n \\ \hat{y}_i + c_r(y_i - \hat{y}_i) & i = 1, 2, \ldots, r \end{cases} \tag{13}$$

et que

$$c_r = \left[n^{-1}(n-1)r(r-p)^{-1}\right]^{1/2}(1 + d_i) \tag{14}$$

où d_i est le nombre de fois que $\hat{e}_i$ est utilisé comme donneur. Le terme $\left[n^{-1}(n-1)r(r-p)^{-1}\right]^{1/2}$ est utilisé pour tenir compte de l'effet de l'estimation de p paramètres de régression. Alors, l'estimateur de la variance (6) peut être représenté par

$$\hat{V}\{\hat{\mu}_y\} = n^{-1}(n-1)^{-1} \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_I)^2 + n^{-2}r(r-p)^{-1} \sum_{i=1}^{r} (1 + d_i)^2 (y_i - \hat{y}_i)^2. \tag{15}$$

Si la méthode d'imputation est une méthode d'échantillonnage aléatoire simple avec remise, alors, subordonnée à l'échantillon et aux répondants,

$$E_I\left\{(1 + d_i)^2\right\} = \left(\frac{n}{r}\right)^2 + \frac{m}{r}\left(1 - \frac{1}{r}\right) \tag{16}$$

où nous utilisons la notation I pour représenter l'espérance en rapport avec le mécanisme d'imputation produit par imputation aléatoire. L'égalité dans (16) établit l'équivalence de (12) et (15) en cas de sélection avec remise. Nous montrons à l'annexe B que $\hat{V}\{\hat{\mu}_y\}$ donné par l'équation (15) est aussi un estimateur valide si les donneurs sont sélectionnés sans remise. Puisque la méthode proposée d'estimation de la variance produit la variance conditionnelle, étant donné l'échantillon imputé réalisé, son applicabilité est très grande.

4. PLANS DE SONDAGE COMPLEXES

4.1 Imputation déterministe

La méthode proposée est applicable à des plans de sondage complexes, ainsi qu'à l'échantillonnage aléatoire simple. Supposons que l'estimateur pour l'échantillon complet de la moyenne de y peut être représenté par $\bar{y} = \Sigma_{i=1}^{n} w_i y_i$, où w_i est le poids d'échantillonnage de l'unité i dans l'échantillon. Supposons aussi que $\Sigma_{i=1}^{n} w_i = 1$.

Si les r premiers éléments sont observés et que les $n - r$ autres éléments manquent, alors l'estimateur de la moyenne de y dans des conditions d'imputation par régression est

$$\bar{y}_I = \sum_{i=1}^{r} w_i y_i + \sum_{i=r+1}^{n} w_i \hat{y}_i \tag{17}$$

où

$$\hat{y}_i = \mathbf{x}_i \hat{\boldsymbol{\beta}},$$

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{r} w_i^* \mathbf{x}_i' \mathbf{x}_i\right]^{-1} \sum_{i=1}^{r} w_i^* \mathbf{x}_i' y_i.$$

Ici, w_i^* représente le poids d'échantillonnage de l'unité i dans l'échantillon de deuxième phase et est défini par

$$w_i^* = [\Pr(i \text{ soit dans l'échantillon de deuxième phase} \mid i \text{ soit dans l'échantillon de première phase})]^{-1} w_i.$$

En outre, $\Sigma_{i=1}^r w_i^* = 1$. Si nous supposons que l'échantillon de deuxième phase est un échantillon aléatoire de taille r sélectionné à partir de l'échantillon de première phase n, alors $w_i^* = nr^{-1} w_i$. Dans certaines conditions, nous pouvons écrire l'estimateur donné par (17) sous la forme

$$\bar{y}_I = \sum_{i=1}^n w_i \hat{y}_i. \tag{18}$$

La représentation (18) est vérifiée si $(w_i^*)^{-1} w_i$ se trouve dans l'espace colonne de la matrice $\mathbf{X} = (\mathbf{x}_1', ..., \mathbf{x}_r')'$ parce qu'alors, nous avons $\Sigma_{i=1}^r w_i(y_i - \hat{y}_i) = 0$ provenant de $\Sigma_{i=1}^r w_i^* \mathbf{x}_i'(y_i - \hat{y}_i) = \mathbf{0}$.

Nous supposons une série d'échantillons et de populations finies telle que celle décrite dans Fuller (1998). Définissons $\bar{\mathbf{x}}_1 = \Sigma_{i=1}^n w_i \mathbf{x}_i$ et $(\bar{\mathbf{x}}_2, \bar{y}_2) = \Sigma_{i=1}^r w_i^*(\mathbf{x}_i, y_i)$. Nous adoptons aussi les mêmes hypothèses que dans Fuller (1998). C'est-à-dire

$$E(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \bar{y}_2) = (\mu_x, \mu_x, \mu_y), \tag{19}$$

et

$$V\{(\hat{\beta} - \beta)', \bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \bar{y}_2\} = O(n^{-1}), \tag{20}$$

où $(\mu_x, \mu_y) = N^{-1} \Sigma_{i=1}^N (\mathbf{x}_i, y_i)$ et $\beta = (\Sigma_{i=1}^N \mathbf{x}_i'\mathbf{x}_i)^{-1} \Sigma_{i=1}^N \mathbf{x}_i' y_i$.

Pour $i = 1, 2, ..., N$, définissons

$$a_i = \begin{cases} 1 & \text{si l'unité } i \text{ répond à l'échantillonnage} \\ 0 & \text{autrement,} \end{cases}$$

et $\mathbf{a} = (a_1, a_2, ..., a_N)$. La définition étendue de a_i est dictée par Fay (1991) et utilisée par Shao et Steel (1999). Maintenant, supposons que

$$\bar{y}_{II} = \sum_{i=1}^n w_i \tilde{y}_i^* \tag{21}$$

où

$$\tilde{y}_i^* = \tilde{y}_i + a_i w_i^{-1} w_i^* (y_i - \tilde{y}_i) \tag{22}$$

avec $\tilde{y}_i = \mathbf{x}_i \beta$. Alors, nous avons $\bar{y}_I = \bar{y}_{II} + (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\hat{\beta} - \beta)$. $\bar{y}_I = \bar{y}_{II} + O_p(n^{-1})$ et $V(\bar{y}_I - \bar{Y}_N) = V(\bar{y}_{II} - \bar{Y}_N) + o(n^{-1})$. Maintenant,

$$V(\bar{y}_{II} - \bar{Y}_N) = V[E(\bar{y}_{II} - \bar{Y}_N \mid \mathbf{a})] + E[V(\bar{y}_{II} - \bar{Y}_N) \mid \mathbf{a})]. \tag{23}$$

Le premier terme du deuxième membre de (23) est nul, car $E(\bar{y}_{II} - \bar{Y}_N \mid \mathbf{a}) = 0$ en vertu du modèle (7). Pour estimer le deuxième terme de (23), notons que, subordonné à $\mathbf{a}$, $\bar{y}_{II}$ est un estimateur linéaire. Donc, la méthode standard d'estimation de la variance appliquée à l'ensemble de pseudo-données $\tilde{\mathbf{Y}}^* \equiv \{\tilde{y}_i^*; i = 1, 2, ..., n\}$ produira une estimation non biaisée de la variance de $\bar{y}_{II} = \Sigma_{i=1}^n w_i \tilde{y}_i^*$. Puisque l'ensemble $\tilde{\mathbf{Y}}^*$ ne peut être observé, nous pouvons utiliser l'ensemble $\mathbf{Y}^* \equiv \{y_i^*; i = 1, 2, ..., n\}$, où

$$y_i^* = \hat{y}_i + a_i w_i^{-1} w_i^* (y_i - \hat{y}_i) \tag{24}$$

pour obtenir un estimateur de la variance cohérent.

Pour montrer que nous pouvons utiliser l'ensemble $\mathbf{Y}^*$ pour obtenir un estimateur approximatif de la variance, supposons que l'estimateur de la variance de l'échantillon complet de $\bar{y}$ peut être représenté par

$$\hat{V} = \sum_{i=1}^L c_i (\bar{y}^{(i)} - \bar{y})^2$$

où L est le nombre de répétitions, c_i est le i^e facteur de répétition et $\bar{y}^{(i)} = \Sigma_{j=1}^n w_j M_j^{(i)} y_j$ est la i^e répétition de $\bar{y}$. Le terme $M_j^{(i)}$ est le multiplicateur de répétition appliqué au poids de l'unité j lors de la i^e répétition. Par exemple, en cas d'échantillonnage aléatoire simple, le multiplicateur jackknife est

$$M_j^{(i)} = \begin{cases} (n-1)^{-1} n & \text{si } i \neq j \\ 0 & \text{si } i = j. \end{cases}$$

Supposons que nous appliquons l'estimateur répété de la variance $\hat{V}$ à l'ensemble $\mathbf{Y}^*$ pour obtenir

$$\hat{V}^* = \sum_{i=1}^L c_i \left(\bar{y}_I^{*(i)} - \bar{y}_I\right)^2$$

où $\bar{y}_I^{*(i)} = \Sigma_{j=1}^n w_j M_j^{(i)} y_j^*$ avec y_j^* définit dans (24). Alors, nous obtenons

$$\bar{y}_I^{*(i)} - \bar{y}_I = \bar{y}_{II}^{*(i)} - \bar{y}_{II} + (\bar{\mathbf{x}}_1^{(i)} - \bar{\mathbf{x}}_2^{(i)} - \bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)(\hat{\beta} - \beta) \tag{25}$$

où

$$(\bar{\mathbf{x}}_1^{(i)}, \bar{\mathbf{x}}_2^{(i)}) = \sum_{j=1}^n w_j M_j^{(i)} (\mathbf{x}_j, a_j w_j^{-1} w_j^* \mathbf{x}_j).$$

Nous montrons à l'annexe C que,

$$\hat{V}^* = \sum_{i=1}^L c_i \left(\bar{y}_{II}^{*(i)} - \bar{y}_{II}\right)^2 + o_p(n^{-1}). \tag{26}$$

Par conséquent, nous pouvons utiliser l'estimateur jackknife type de la variance appliqué à l'ensemble de pseudo-données $\mathbf{Y}^*$ comme approximation de l'estimateur jackknife type de la variance appliqué à l'ensemble de pseudo-données $\tilde{\mathbf{Y}}^*$.

4.2 Imputation aléatoire

Les arguments concernant l'estimation de la variance en cas d'imputation aléatoire sont fort semblables à ceux

décrits pour l'imputation déterministe à la sous-section qui précède. En premier lieu, définissons la fonction indicatrice d'imputation

$$d_{ij} = \begin{cases} 1 & \text{si l'unité } i \text{ est utilisée comme donneur pour l'unité } j \\ 0 & \text{autrement.} \end{cases} \tag{27}$$

Alors, l'estimateur de la moyenne de y en recourant à l'imputation aléatoire est

$$\bar{y}_I = \sum_{i=1}^{n} w_i y_i^* \tag{28}$$

où

$$\bar{y}_i^* = \hat{y}_i + a_i(1 + d_i)(y_i - \hat{y}_i) \tag{29}$$

et

$$d_i = \sum_{j=1}^{n} (1 - a_j) d_{ij} w_i^{-1} w_j. \tag{30}$$

Si les poids d'échantillonnage originaux sont les mêmes, alors d_i représente le nombre de fois que l'unité i est utilisée comme donneur. Nous supposons que

$$E[a_i(1 + d_i) \mid \mathcal{F}_1] = 1 \tag{31}$$

où $\mathcal{F}_1 = \{(i, \mathbf{x}_i, y_i); i = 1, 2, ..., n\}$. Dans (31), l'espérance a trait à la distribution combinée du mécanisme de réponse et du mécanisme d'imputation. Alors, nous avons

$$E(\bar{y}_I \mid \mathcal{F}_1) \doteq \bar{y}.$$

Si nous supposons que les probabilités de réponse sont égales, alors, en vertu de l'hypothèse (31), la probabilité de sélection des donneurs devrait être proportionnelle aux poids, ce qui correspond aux conditions établies par Rao et Shao (1992) pour l'imputation aléatoire.

Maintenant, supposons que

$$\bar{y}_{II} = \sum_{i=1}^{n} w_i [\tilde{y}_i + a_i(1 + d_i)(y_i - \tilde{y}_i)] \tag{32}$$

où $\tilde{y}_i = \mathbf{x}_i \boldsymbol{\beta}$. Alors, nous avons aussi $\bar{y}_I = \bar{y}_{II} + (\bar{\mathbf{x}}_d - \bar{\mathbf{x}}_1)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ où $\bar{\mathbf{x}}_d = \sum_{i=1}^{n} w_i a_i (1 + d_i) \mathbf{x}_i$. En vertu de l'hypothèse (31), nous avons $E(\bar{\mathbf{x}}_d - \bar{\mathbf{x}}_1 \mid \mathcal{F}_1) = 0$. Si les contraintes sont peu sévères, $\bar{\mathbf{x}}_d - \bar{\mathbf{x}}_1 = O_p(n^{-1/2})$ et $\bar{y}_I = \bar{y}_{II} + O_p(n^{-1})$. Maintenant,

$$V(\bar{y}_I - \bar{Y}_N) = V[E(\bar{y}_I - \bar{Y}_N \mid \mathbf{a}, \mathbf{d})] + E[V(\bar{y}_I - \bar{Y}_N \mid \mathbf{a}, \mathbf{d})]$$

où $\mathbf{d} = (d_1, d_2, ..., d_N)$. Subordonné à $\mathbf{a}$ et à $\mathbf{d}$, l'estimateur $\bar{y}_{II}$ est linéaire. Donc, nous pouvons utiliser les pseudo-données

$$y_i^* = \hat{y}_i + a_i(1 + d_i)(y_i - \hat{y}_i) \tag{33}$$

pour estimer la variance de $\bar{y}_I$.

5. COMPARAISON À LA MÉTHODE CORRIGÉE DU JACKKNIFE

Rao et Sitter (1995) ont proposé un estimateur jackknife corrigé de la variance pour résoudre le problème de l'imputation par la méthode du quotient. Dans les conditions décrites à la section 4, l'estimateur imputé par quotient de μ_y est

$$\hat{\mu}_I = \sum_{i=1}^{n} w_i [a_i y_i + (1 - a_i)\hat{y}_i]$$

où $\hat{y}_i = x_i \hat{R}$ et $\hat{R} = (\sum_{i=1}^{n} w_i a_i x_i)^{-1} \sum_{i=1}^{n} w_i a_i y_i$. L'estimateur de la variance de Rao et Sitter (1995) est donné par

$$V_a = \sum_{i=1}^{L} c_i \left(\hat{\mu}_I^{(i)} - \hat{\mu}_I\right)^2, \tag{34}$$

où la répétition jackknife corrigée obtenue à la i^e itération est donnée par

$$\hat{\mu}_I^{(i)} = \sum_{j=1}^{n} w_j M_j^{(i)} y_j^{*(i)} \tag{35}$$

où

$$y_j^{*(i)} = \begin{cases} x_j \hat{R}^{(i)} & \text{si } a_i = 1 \\ x_j \hat{R} & \text{si } a_i = 0 \end{cases} \tag{36}$$

avec $\hat{R}^{(i)} = (\sum_{j=1}^{n} w_j M_j^{(i)} a_j x_j)^{-1} \sum_{j=1}^{n} w_j M_j^{(i)} a_j y_j$. Les valeurs corrigées (36) de la méthode de Rao et Sitter (1995) peuvent aussi être considérées comme des pseudo-données pour l'estimation de la variance. Notons que le calcul des pseudo-données (36) oblige à recalculer $\hat{R}^{(i)}$ pour chaque i, avec $a_i = 1$.

Nous modifions le calcul des pseudo-valeurs y_i^* dans (5) pour obtenir

$$y_i^* = \begin{cases} \hat{y}_i & \text{si } a_i = 0 \\ \hat{y}_i + c_r \left(\dfrac{\bar{x}_1}{\bar{x}_2}\right)(y_i - \hat{y}_i) & \text{si } a_i = 1, \end{cases} \tag{37}$$

où $\bar{x}_2 = \sum_{i=1}^{n} w_i r^{-1} n a_i x_i$, $\bar{x}_1 = n^{-1} \sum_{i=1}^{n} w_i x_i$ et $c_r \doteq r^{-1} n$. Nous insérons le terme $(\bar{x}_1 / \bar{x}_2)$ pour améliorer les propriétés conditionnelles de V_J, étant donné l'échantillon de première phase. L'estimateur résultant de la variance est approximativement équivalent à l'estimateur jackknife corrigé de la variance donné par (34). Pour le confirmer, notons que les valeurs corrigées (35) peuvent s'écrire sous la forme

$$\hat{\mu}_I^{(i)} = \left(\sum_{j=1}^{n} w_j M_j^{(i)} x_j\right) \frac{\sum_{j=1}^{n} w_j M_j^{(i)} a_j y_j}{\sum_{j=1}^{n} w_j M_j^{(i)} a_j x_j} =: \hat{Z}^{(i)} \frac{\hat{S}^{(i)}}{\hat{T}^{(i)}},$$

où $A =: B$ indique que nous définissons B pour qu'il soit équivalent à A. En outre, définissons $\hat{Z} = \sum_{j=1}^{n} w_j x_j$, $\hat{S} = \sum_{j=1}^{n} w_j a_j y_j$, et $\hat{T} = \sum_{j=1}^{n} w_j a_j x_j$. Alors, en vertu du développement par série de Taylor de premier ordre,

$$\begin{aligned}\hat{Z}^{(i)}\frac{\hat{S}^{(i)}}{\hat{T}^{(i)}} &\doteq \hat{Z}\frac{\hat{S}}{\hat{T}} + \left(\hat{Z}^{(i)} - \hat{Z}\right)\frac{\hat{S}}{\hat{T}} \\ &\quad + \left(\hat{S}^{(i)} - \hat{S}\right)\frac{\hat{Z}}{\hat{T}} - \left(\hat{T}^{(i)} - \hat{T}\right)\frac{\hat{Z}\hat{S}}{\hat{T}^2} \\ &= \left[\hat{Z}^{(i)}\frac{\hat{S}}{\hat{T}} + \frac{\hat{Z}}{\hat{T}}\left(\hat{S}^{(i)} - \hat{T}^{(i)}\frac{\hat{S}}{\hat{T}}\right)\right]. \end{aligned} \tag{38}$$

Notons que le deuxième membre de l'équation (38) est exactement égal à

$$\sum_{j=1}^{n} w_j M^{(i)}\left[\frac{\hat{S}}{\hat{T}} + \frac{\hat{Z}}{\hat{T}} a_j\left(y_j - \frac{\hat{S}}{\hat{T}}\right)\right].$$

Donc, les pseudo-données produites pour estimer la variance peuvent s'écrire sous la forme

$$y_i^* = \frac{\hat{S}}{\hat{T}} + \frac{\hat{Z}}{\hat{T}} a_i\left(y_i - \frac{\hat{S}}{\hat{T}}\right),$$

que l'on peut réduire à (37). Par conséquent, la méthode proposée représente exactement une linéarisation par série de Taylor de premier ordre de la méthode de Rao et Sitter dans le cas de l'imputation par la méthode du quotient. Par conséquent, en principe, la méthode que nous proposons doit avoir les mêmes propriétés asymptotiques que la méthode de Rao et Sitter jusqu'à l'ordre n^{-1}.

La méthode d'estimation de la variance basée sur l'ensemble de pseudo-données calculées au moyen de l'équation (37) est facile à appliquer parce que nous pouvons nous servir directement d'un logiciel existant, ce qui est plus difficile dans le cas de la méthode de Rao et Shao (1992) ou de Rao et Sitter (1995). En outre, si nous calculons les pseudo-données au moyen de l'équation (13), alors l'ensemble de données convient pour l'imputation par la méthode hot deck sans remise ainsi que pour l'imputation par la méthode hot deck avec remise.

6. ÉTUDE PAR SIMULATION

Pour éprouver la théorie qui précède, nous avons réalisé une étude en simulation portant sur une population artificielle, finie, à partir de laquelle nous avons sélectionné des échantillons répétés. La population compte $L = 32$ strates, N_h grappes dans la strate h et 20 unités finales dans chaque grappe. Les valeurs des paramètres de population ont été choisies de façon à ce qu'elles correspondent aux valeurs de population réelles observées dans le cadre de la *U.S. National Assessment of Educational Progress Study* (Hansen et Tepping 1985) et sont énumérées au tableau 1.

Les unités de la population finie sont représentées par

$$y_{hij} = y_{hi} + e_{hij},$$

où

$$y_{hi} \overset{\text{iid}}{\sim} N(\mu_h, \sigma_h^2), \quad h = 1, 2, ..., L, i = 1, 2, ..., N_h,$$

et

$$e_{hij} \overset{\text{iid}}{\sim} N\left(0, \frac{1-\rho}{\rho}\sigma_h^2\right), \quad j = 1, 2, ..., 20.$$

Shao, Chen et Chen (1998) se sont aussi servi de la même population pour réaliser leur étude en simulation. La valeur de la corrélation intra-grappe ρ considérée dans la simulation est $\rho = 0{,}3$. Les simulations réalisées en prenant d'autres valeurs de ρ ont donné des résultats similaires et ne sont donc pas présentés ici par souci de concision.

Tableau 1
Paramètres de la population finie utilisés pour la simulation

h	N_h	μ_h	σ_h	h	N_h	μ_h	σ_h
1	13	100,0	20,0	2	16	95,0	19,0
3	20	90,0	18,0	4	25	98,0	19,6
5	25	93,0	18,6	6	25	98,0	19,6
7	25	96,0	19,2	8	28	94,0	18,8
9	28	92,0	18,4	10	28	96,0	19,2
11	31	94,0	18,8	12	31	92,0	18,4
13	31	90,0	18,0	14	31	96,0	19,2
15	31	94,0	18,8	16	31	92,0	18,4
17	31	90,0	18,0	18	31	88,0	17,6
19	31	86,0	17,2	20	34	84,0	16,8
21	34	82,0	16,4	22	34	80,0	16,0
23	34	90,0	18,0	24	37	85,0	17,0
25	37	80,0	16,0	26	37	90,0	18,0
27	37	85,0	17,0	28	39	80,0	16,0
29	39	75,0	15,0	30	42	75,0	15,0
31	42	75,0	15,0	32	42	75,0	15,0

Nous considérons un plan d'échantillonnage par grappe stratifié où on sélectionne $n_h = 2$ grappes avec remise dans la strate h avec probabilité égale et où toutes les unités finales qui figurent dans les grappes sélectionnées sont dans l'échantillon. La fraction d'échantillonnage est de 6,4 %. Pour chaque unité échantillonnée y_{hij}, nous produisons une variable indicatrice de réponse a_{hij} à partir de

$$a_{hij} \overset{\text{iid}}{\sim} \text{Bernoulli } (p),$$

et cet a_{hij} est indépendant de y_{hij}. Les valeurs de p considérées pour la simulation sont $p = 0{,}9, 0{,}8, 0{,}7, 0{,}6$ et $0{,}5$.

Nous sélectionnons un ensemble de 5 000 échantillons selon le même plan d'échantillonnage. Pour chaque échantillon sélectionné, nous considérons trois méthodes d'imputation.

[M1] Imputation par la méthode hot deck pondérée avec remise considérée par Rao et Shao (1992), où une valeur manquante est remplacée par imputation d'une valeur sélectionnée au hasard parmi les répondants avec remise et probabilité proportionnelle au poids de sondage.

[M2] Imputation par la méthode hot deck pondérée sans remise, qui est identique à la méthode [M1], hormis le fait que la sélection est effectuée par échantillonnage sans remise. La sélection de donneurs sans remise est exécutée systématiquement selon la méthode décrite par Hansen, Hurwitz et Madow (1953, page 343) pour les répondants triés par ordre aléatoire.

[M3] Imputation globale de la moyenne, où l'on impute la moyenne pondérée calculée pour les répondants qui figurent dans l'échantillon.

Par conséquent, toutes les méthodes d'imputation comportent une cellule d'imputation unique qui regroupe toutes les strates.

Pour chaque ensemble contenant de données imputées, nous calculons trois estimateurs de la variance, à savoir $\hat{V}_n$, qui est l'estimateur naïf de la variance obtenu lorsque l'on traite les données imputées comme s'il s'agissait de données observées, $\hat{V}_a$, qui est l'estimateur jackknife corrigé de la variance proposé par Rao et Shao (1992) pour [M1] et [M2] et par Rao et Sitter (1995) pour [M3], et $\hat{V}^*$, qui est l'estimateur jackknife de la variance fondé sur les pseudo-données. Nous produisons l'ensemble de pseudo-données au moyen de l'équation (29) pour [M1] et [M2] et au moyen de l'équation (24) pour [M3]. L'estimateur complet de la variance d'échantillon est un estimateur jackknife type pour l'échantillonnage par grappe stratifié, selon lequel une grappe est supprimée à chaque itération. Notons que l'estimateur jackknife type est un estimateur cohérent de la variance dans les conditions du modèle avec corrélation intra-grappe non nulle. Donc, nous pouvons appliquer la méthode type du jackknife fondée sur les pseudo-données à l'ensemble de données considéré. Les estimateurs ponctuels de la moyenne de population sont non biaisés dans le cas des trois scénarios distincts d'imputation et nous ne les présentons pas ici.

Le tableau 2 donne le biais relatif des trois estimateurs de la variance, l'écart-type du biais relatif des estimateurs de la variance et le coefficient de corrélation d'échantillon entre l'estimateur jackknife corrigé de la variance de Rao et le nouvel estimateur de la variance fondés sur les 5 000 échantillons. Nous calculons le biais relatif de $\hat{V}$, en tant qu'estimateur de la variance de $\bar{y}_I$, au moyen de $[\mathrm{Var}_B(\bar{y}_I)]^{-1}[E_B(\hat{V}) - \mathrm{Var}_B(\bar{y}_I)]$, où l'indice B représente la distribution produite par la simulation de Monte Carlo. Enfin, nous calculons les coefficients de corrélation des deux estimateurs de la variance pour donner une mesure de la linéarité relative de ces estimateurs.

Tableau 2
Biais relatif de l'estimateur de la variance, écart-type du biais relatif et coefficient de corrélation de l'échantillon entre l'estimateur de la variance de Rao et le nouvel estimateur de la variance fondés sur 5 000 échantillons

Taux de réponse (p)	Méthode d'imputation	Biais rel. × 100 (É. -T. × 100)			Coeff. de corr. r
		Naïf	Rao	Nouveau	
0,9	M1	-17,40 (2,02)	1,61 (2,03)	1,70 (2,04)	0,967
	M2	-17,50 (2,00)	1,41 (2,01)	0,81 (2,03)	0,974
	M3	-18,03 (2,03)	1,16 (2,05)	1,15 (2,04)	1,000
0,8	M1	-34,45 (2,01)	0,65 (2,03)	0,49 (2,05)	0,939
	M2	-32,89 (2,01)	2,49 (2,04)	0,19 (2,03)	0,947
	M3	-34,96 (2,01)	1,59 (2,03)	1,59 (2,03)	1,000
0,7	M1	-48,96 (2,01)	0,21 (1,99)	0,41 (2,04)	0,912
	M2	-44,76 (2,02)	5,31 (2,05)	0,76 (2,05)	0,920
	M3	-50,21 (2,02)	1,53 (2,05)	1,52 (2,04)	1,000
0,6	M1	-59,80 (2,02)	1,58 (2,05)	1,27 (2,06)	0,892
	M2	-54,86 (2,03)	7,10 (2,07)	-0,75 (2,07)	0,899
	M3	-64,11 (2,00)	-0,35 (2,04)	-0,35 (2,01)	1,000
0,5	M1	-69,75 (1,99)	0,84 (2,03)	1,12 (2,03)	0,873
	M2	-59,90 (2,01)	15,07 (2,07)	2,27 (2,06)	0,872
	M3	-74,44 (1,97)	1,99 (2,00)	1,98 (2,00)	1,000

Le tableau 2 confirme notre théorie de la façon suivante.

1. Il est bien connu que l'estimateur naïf de la variance sous-estime fortement la variance réelle. L'estimateur jackknife corrigé de la variance donne de bons résultats pour [M1] et [M3], mais pas pour [M2]. La théorie qui sous-tend la méthode corrigée du jackknife suppose que les imputations par la méthode hot deck s'appuient sur l'échantillonnage avec remise, lequel n'est pas utilisé dans [M2]. À mesure que le taux de réponse diminue dans le tableau 2, le biais relatif de l'estimateur jackknife corrigé augmente.

2. La nouvelle méthode basée sur les pseudo-données donne de bons résultats même pour l'imputation avec échantillonnage sans remise [M2]. Comme nous en avons discuté à la fin de la section 3, la formule (29) suffit, à elle seule, à produire les pseudo-données pour une grande classe de méthodes d'imputation.

3. Comme le montrent les coefficients de corrélation, les comportements de l'estimateur jackknife corrigé et de l'estimateur proposé de la variance sont très semblables pour l'imputation de la moyenne [M3], parce que les deux estimateurs sont asymptotiquement équivalents, comme nous l'avons exposé à la section 5.

7. CONCLUSIONS

Nous avons décrit certaines méthodes qui permettent de produire des pseudo-données pour estimer la variance.

Généralement parlant, les pseudo-données peuvent être décrites par la formule

$$y_i^* = \begin{cases} \hat{y}_i & i = r+1, r+2, ..., n \\ \hat{y}_i + c_i g_i (y_i - \hat{y}_i) & i = 1, 2, ..., r, \end{cases} \quad (39)$$

où $\hat{y}_i$ est la valeur prévue de y_i dans les conditions du modèle utilisé pour l'imputation. Si $c_i g_i = 1$, alors l'estimateur de la variance traite les valeurs imputées comme des observations. En choisissant convenablement $c_i g_i > 1$, on obtient un estimateur de la variance cohérent. Si la méthode d'imputation est déterministe et que les répondants utilisés sont considérés comme étant un échantillon aléatoire de l'échantillon original, alors $c_i \doteq r^{-1}n > 1$. Pour un échantillonnage à deux phases avec plan de sondage complexe, $c_i = w_i^{-1} w_i^*$, où w_i représente le poids d'échantillonnage de l'unité i pour l'échantillon de première phase et w_i^* représente le poids d'échantillonnage de l'unité i pour l'échantillon de seconde phase.

Le g_i de l'équation (39) représente la correction faite pour améliorer les propriétés conditionnelles, étant donné la variable auxiliaire x. Pour l'imputation par la méthode du quotient,

$$g_i = (\bar{x}_2)^{-1} \bar{x}_1$$

où $\bar{x}_2 = \Sigma_{i=1}^r w_i^* x_i$ et $\bar{x}_1 = \Sigma_{i=1}^n w_i x_i$. Pour l'imputation par régression avec la grandeur scalaire x,

$$g_i = 1 + (\bar{x}_1 - \bar{x}_2) \left\{ \sum_{k=1}^r w_k^* (x_k - \bar{x}_2)^2 \right\}^{-1} (x_i - \bar{x}_2).$$

Dans l'un et l'autre cas, nous avons

$$\sum_{i=1}^r w_i^* g_i x_i = \bar{x}_1.$$

Pendant l'évaluation du présent article, Shao et Steel (1999) ont proposé des méthodes similaires dans le cas de l'imputation déterministe. Notre méthode est plus générale en ce sens que nous considérons aussi l'imputation aléatoire et que nous introduisons le terme c_i pour améliorer les propriétés de l'échantillon fini.

REMERCIEMENTS

L'auteur remercie son conseiller de thèse, Wayne A. Fuller, pour les discussions fructueuses qu'ils ont eues. Il remercie aussi Pamela Abbitt, F. Jay Breidt, Lou Rizzo, Richard Valliant et les examinateurs pour leurs commentaires utiles, qui lui ont permis d'améliorer considérablement l'article. La plupart des travaux présentés ici ont été réalisés pendant que l'auteur poursuivait des études de cycles supérieures à la Iowa State University et financés en partie par l'entente de coopération 68-3A75-43 conclue par le USDA Natural Resources Conservation Service et la Iowa State University et par l'entente de coopération 43-3AEU-3-80088 conclue par la Iowa State University, le National Agricultural Statistics Service et le U.S. Bureau of Census.

ANNEXE

A. Preuve des équations (10) et (12)

L'estimateur $\hat{\mu}_y$ qui figure dans l'équation (9) peut être écrit sous la forme

$$\hat{\mu}_y = n^{-1} \sum_{i=1}^n \hat{y}_i + n^{-1} \sum_{i=1}^r (1 + d_i) \hat{e}_i \quad (A.1)$$

où d_i est le nombre de fois que l'unité i est utilisée comme un donneur. En vertu du mécanisme d'imputation avec échantillonnage à probabilité égale et avec remise, nous avons

$$E_I(d_i) = r^{-1} m$$

et

$$\mathrm{Cov}_I(d_i, d_j) = \begin{cases} r^{-1} m (1 - r^{-1}) & \text{si } i = j \\ -r^{-2} m & \text{si } i \neq j \end{cases}$$

où l'indice I représente la variation due au mécanisme d'imputation. Il s'ensuit que $E_I(\hat{\mu}_y) = n^{-1} \Sigma_{i=1}^n \hat{y}_i$ et $V_I(\hat{\mu}_y) = n^{-2} r^{-1} m \Sigma_{i=1}^r \hat{e}_i^2$. Par conséquent,

$$V(\hat{\mu}_y) \doteq V\left(n^{-1} \sum_{i=1}^n \hat{y}_i \right) + E\left(n^{-2} r^{-1} m \sum_{i=1}^r \hat{e}_i^2 \right). \quad (A.2)$$

Maintenant, par un argument similaire à celui qui nous a mené à (2), nous obtenons

$$\mathrm{Var}\left(n^{-1} \sum_{i=1}^n \hat{y}_i \right) = \left[n^{-1} R^2 + r^{-1} (1 - R^2) \right] \sigma_y^2. \quad (A.3)$$

Puisque $\hat{y}_i - \bar{y}_I = (\mathbf{x}_i - \bar{\mathbf{x}}_i) \beta + o_p(1)$, nous appliquons la théorie classique de régression pour obtenir

$$E\left[(r - p)^{-1} \sum_{i=1}^r \hat{e}_i^2 \right] = (1 - R^2) \sigma_y^2, \quad (A.4)$$

et

$$E\left[(n - 1)^{-1} \sum_{i=1}^n (\hat{y}_i - \bar{y}_I)^2 \right] = R^2 \sigma_y^2. \quad (A.5)$$

Par conséquent, l'équation (10) est prouvée et l'estimateur donné par (12) est cohérent pour la variance donnée par l'équation (10).

B. Validité de l'équation (15) dans les conditions du mécanisme d'imputation sans remise

Nous supposons que $m = kr + t$ où k et t sont des nombres entiers non négatifs et que $t < r$. Supposons que l'estimateur de la moyenne de y a la forme (A.1). Supposons que l'imputation est exécutée de façon telle que t répondants sont utilisés $k + 1$ fois pour l'imputation et que $r - t$ unités sont utilisés k fois pour l'imputation. Les t répondants qui sont utilisés $k + 1$ fois sont choisis par échantillonnage aléatoire simple sans remise. Alors,

$$E_I(d_i) = k + r^{-1}t = r^{-1}m$$

et

$$\mathrm{Cov}_I(d_i, d_j) = \begin{cases} r^{-1}t(1 - r^{-1}t) & \text{si } i = j \\ -r^{-2}t & \text{si } i \neq j. \end{cases}$$

Donc, par des arguments similaires à ceux utilisés dans la preuve de (A.2), nous obtenons

$$V(\hat{\mu}_y) \doteq V(\bar{y}_I) + E\left(n^{-2}r^{-1}t\sum_{i=1}^{r}\hat{e}_i^2\right). \qquad \text{(B.1)}$$

Par conséquent, en utilisant (A.3) et (A.4), nous obtenons

$$V\{\hat{\mu}_y\} = \left[n^{-1}R^2 + (r^{-1} + n^{-2}t)(1 - R^2)\right]\sigma_y^2. \qquad \text{(B.2)}$$

Maintenant, subordonnée à l'échantillon réalisé et aux répondants, nous obtenons

$$E_I\left\{(1 + d_i)^2\right\} = \left(\frac{n}{r}\right)^2 + \frac{t}{r}\left(1 - \frac{t}{r}\right)$$

de sorte que $\hat{V}\{\mu_y\}$ dans l'équation (15) satisfait

$$E_I\left(\hat{V}\{\mu_y\}\right) \doteq n^{-1}(n-1)^{-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_I)^2$$
$$+ [r^{-1} + n^{-2}t(1 - r^{-1}t)]$$
$$(r - p)^{-1}\sum_{i=1}^{r}(y_i - \hat{y}_i)^2.$$

Par conséquent, en utilisant (A.4) et (A.5), nous obtenons l'absence de biais approximative de $\hat{V}\{\mu_y\}$ dans le cas du mécanisme d'imputation sans remise.

C. Preuve de l'équation (26)

Premièrement, définissons $R_n^{(i)} = (\mathbf{x}_1^{(i)} - \bar{\mathbf{x}}_2^{(i)})(\hat{\beta} - \beta)$ et $R_n = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\hat{\beta} - \beta)$. D'après l'égalité (25),

$$\hat{V}^* = \sum_{i=1}^{L} c_i\left(\bar{y}_I^{*(i)} - \bar{y}_I\right)^2 = A_n + B_n + 2C_n$$

où $A_n = \sum_{i=1}^{L} c_i(\bar{y}_{II}^{*(i)} - \bar{y}_{II})^2$, $B_n = \sum_{i=1}^{L} c_i(R_n^{(i)} - R_n)^2$, et $C_n = \sum_{i=1}^{L} c_i(\bar{y}_{II}^{*(i)} - \bar{y}_{II})(R_n^{(i)} - R_n)$. Par conséquent, en vertu de l'hypothèse (20), il s'ensuit l'équation (26), car $A_n = O_p(n^{-1})$, $B_n = o_p(n^{-1})$, et $C_n = o_p(n^{-1})$. La dernière propriété provient de l'inégalité de Cauchy-Schwartz, $C_n^2 \leq A_n B_n$.

BIBLIOGRAPHIE

COCHRAN, W.G. (1977). *Sampling Techniques.* New York : John Wiley and Sons.

FAY, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the Bureau of the Census Annual Research conference*, 429-440.

FAY, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 227-232.

FAY, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.

FULLER, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica*, 8, 1153-1164.

HANSEN, M., HURWITZ, W.N. et MADOWS, W.G. (1953). *Sample Survey Methods and Theory.* Vol. I, New York : John Wiley and Sons.

HANSEN, M., et TEPPING, B.J. (1985). Estimation for Variance in NAEP. Note de service inédite, Westat, Washington, D.C.

RAO, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.

RAO, J.N.K., et SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

RAO, J.N.K., et SITTER, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York : John Wiley and Sons.

RUBIN, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

RUBIN, D.B., et SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.

SÄRNDAL, C.-E. (1992). Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation. *Techniques d'enquête*, 18, 257-268.

SÄRNDAL, C.-E., et SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *Revue Internationale de la Statistique*, 55, 279-294.

SCHAFER, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

SHAO, J., CHEN, Y. et CHEN, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

SHAO, J., et STEEL, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fraction. *Journal of the American Statistical Association*, 94, 254-265.

Techniques d'enquête, juin 2001
Vol. 27, n° 1, pp. 91-103
Statistique Canada

Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression

TRIVELLORE E. RAGHUNATHAN, JAMES M. LEPKOWSKI, JOHN VAN HOEWYK
et PETER SOLENBERGER[1]

RÉSUMÉ

Le présent article décrit et évalue une procédure d'imputation des valeurs manquantes pour une structure relativement complexe des données lorsque celles-ci manquent au hasard. On obtient les imputations en ajustant une séquence de modèles de régression et en tirant les valeurs des distributions prédictives correspondantes. Les types de modèle de régression utilisés sont les suivants : linéaire, logistique, de Poisson, logit généralisé, ou encore un mélange qui dépend du type de variable imputé. Deux autres caractéristiques communes du processus d'imputation sont intégrées : la restriction à une sous-population pertinente pour certaines variables et des limites ou contraintes logiques pour les valeurs imputées. Les restrictions comportent la création de sous-ensembles d'unités d'échantillon répondant à certains critères au moment de l'ajustement des modèles de régression. Les limites supposent que l'on tire des valeurs d'une distribution prédictive tronquée. L'élaboration de cette méthode s'est inspirée en partie de l'analyse de deux fichiers de données utilisés à titre d'illustration. On applique la procédure de régression séquentielle à l'analyse d'imputations multiples pour les deux problèmes appliqués. Les propriétés d'échantillonnage des inférences tirées de fichiers de données polyimputées créés à l'aide de la méthode de régression séquentielle sont évaluées en fonction de fichiers de données simulées.

MOTS CLÉS : Non-réponse partielle; manquant au hasard; imputation multiple; mécanisme de données manquantes non-ignorable; régression; simulations et propriétés d'échantillonnage.

1. INTRODUCTION

Les données incomplètes sont un problème fréquent dans la plupart des recherches appliquées. On a élaboré plusieurs méthodes permettant de tirer des inférences de fichiers de données comportant des valeurs manquantes (Little et Rubin 1987), et ce travail se poursuit. Le schéma d'imputations multiples proposé par Rubin (1978, 1987a, 1996) est une possibilité intéressante si un fichier de données est destiné à plusieurs chercheurs ayant différentes compétences en statistique. Cette façon de procéder suppose l'imputation de plusieurs ensembles plausibles de valeurs manquantes dans le fichier de données incomplètes de façon à fournir plusieurs fichiers de données complètes. Chaque fichier de données complètes est analysé séparément, par exemple en ajustant un modèle de régression particulier. Les inférences qui en résultent (estimations ponctuelles et matrices de covariances) sont alors combinées à l'aide de la formule de Rubin (1987a, chapitre 3) et de perfectionnements de cette formule (Li, Raghunathan et Rubin 1991; Li, Meng, Raghunathan et Rubin 1991; Meng et Rubin 1992; et Barnard 1995).

De façon générale, les stratégies de traitement des données manquantes par imputation sont fort utiles dans la pratique, car une fois les valeurs manquantes imputées, on peut avoir recours à des logiciels de données complètes existants pour analyser les données. Puisque l'élaboration de logiciels pour l'analyse des données complètes évolue en fonction de l'introduction de nouvelles méthodes statistiques, les personnes qui s'adonnent à des recherches appliquées, sans connaître les ressources ou techniques particulières leur permettant de créer leur propre code de mise en oeuvre de nouvelles procédures pour les données manquantes, seront en mesure d'ajuster des modèles fonctionnels peaufinés en présence d'un problèmes précis. Un autre avantage de la stratégie d'imputations multiples, c'est que l'application répétée d'un logiciel pour les données complètes permet d'obtenir des estimations ponctuelles et des estimations d'intervalles valides pour une série de conditions assez générales (Rubin 1987a). Plusieurs chercheurs (voir par exemple la bibliographie de Rubin 1996) ont appliqué cette technique dans différentes situations et ont montré, en analysant des fichiers de données simulées et réelles, le bien-fondé de cette stratégie. D'autres possibilités comme l'imputation unique en fonction d'une procédure d'estimation de la variance appropriée, par exemple la technique de répétition répétée de type jackknife modifiée (Rao et Shao 1992), offrent également cet avantage. La stratégie d'imputation décrite dans le présent exposé peut également servir à créer une imputation unique avec une autre procédure d'estimation de la variance.

L'élaboration de méthodes d'imputation en fonction de divers points de vue a un long passé (Madow, Nisselson, Olkin et Rubin 1983). Un schéma théoriquement élégant pour l'élaboration de méthodes d'imputation est la stratégie bayésienne. Celle-ci exige un modèle explicite pour des

[1] Trivellore E. Raghunathan, James M. Lepkowski, John van Hoewyk et Peter Solenberger, University of Michigan, Institute for Social Research, Survey Methodology Program, P.O. Box 1248, Ann Arbor, MI 48106-1248, U.S.A.

variables comportant des valeurs manquantes, lié aux variables observées intégralement et à certains paramètres inconnus, une distribution a priori pour les paramètres inconnus et un modèle du mécanisme des données manquantes, qu'il n'est pas nécessaire de préciser dans le cadre d'un schéma de données manquantesignorables (Rubin 1976). Ce modèle explicite donne alors lieu à une distribution prédictive a posteriori des valeurs manquantes, liée aux valeurs observées. Les imputations sont tirées de cette distribution prédictive a posteriori. Il existe plusieurs programmes et algorithmes informatiques pour l'imputation des valeurs manquantes en présence d'une normalité multidimensionnelle (Rubin et Schafer 1990), de la distribution *t* multidimensionnelle (Liu 1995) et de diverses variations du modèle d'emplacement général (Schafer 1997, Raghunathan et Grizzle 1995, Raghunathan et Siscovick 1996). Ce dernier modèle permet de traiter la distribution composée de variables catégoriques et continues; il a d'abord été proposé par Olkin et Tate (1961) et a été utilisé par Little et Schluchter (1985) en fonction précisément de problèmes de données manquantes. Une propriété importante de ces stratégies, c'est qu'elles dépendent entièrement de toute l'information observée. Plusieurs études de simulation (Raghunathan et Grizzle 1995, par exemple) indiquent que les inférences tirées de ce genre de données imputées offrent des propriétés d'échantillonnage souhaitables.

Les fichiers de données d'enquête comportent souvent de très nombreuses variables ayant différentes distributions. Typiquement, ces fichiers de données ont des centaines de variables, les unes continues, les autres de type comptage, souvent dichotomiques ou polytomiques, et même parfois dépendantes et semi-continues ou limitées. De plus, les variables continues peuvent comporter une distribution normale, normale logarithmique ou autre. Il peut être très difficile dans une telle situation de postuler un modèle bayésien intégral. De plus, les données d'enquête ont souvent deux autres caractéristiques qui rendent la modélisation encore plus complexe. Tout d'abord, certaines restrictions sont impératives. Ainsi, la variable « nombre d'années depuis la cessation de l'usage du tabac » est définie uniquement pour des personnes qui ont déjà fumé; par conséquent, le processus d'imputation pour cette variable devrait se limiter aux personnes qui ont déjà fumé. Certaines restrictions relèvent d'instructions « passez à » dans un questionnaire. Ainsi, certaines questions sur le revenu d'un deuxième emploi sont posées uniquement lorsque le répondant indique qu'il ou elle a un deuxième emploi. L'imputation de ce genre de variable exige un traitement hiérarchique.

Deuxièmement, il existe des limites logiques ou des limites de cohérence pour les valeurs manquantes qu'il faut intégrer au processus d'imputation. Une telle interdépendance des variables rend la création du modèle difficile. Ainsi, « années d'usage du tabac » se limite aux personnes qui fument ou qui ont fumé, et les valeurs imputées doivent être inférieures à Âge – x années, où x peut relever d'autres caractéristiques comme le fait d'avoir fumé au cours de l'adolescence. Dans le cas d'une personne qui a déjà fumé, x englobe également les années écoulées depuis la cessation de l'usage du tabac. Un autre exemple de limites est abordé dans Heeringa, Little et Raghunathan (1997). Ces auteurs examinent l'imputation de questions comportant des réponses entre crochets lorsque le répondant ne peut pas ou ne veut pas donner une réponse exacte (au sujet du revenu ou de l'avoir, par exemple), tout en définissant les limites à l'intérieur desquelles les valeurs imputées doivent se situer.

Le présent exposé a comme objectif de proposer et d'évaluer une procédure d'imputation multidimensionnelle générale permettant de traiter une structure de données relativement complexe lorsque des modèles multidimensionnels complets explicites ne se laissent pas facilement formuler, les valeurs imputées pour chaque unité étant toutefois entièrement liées à toutes les valeurs observées pour l'unité en question. La stratégie consiste à considérer l'imputation, une variable à la fois, mais en fonction de toutes les variables observées. La stratégie de base permet de créer des imputations en vertu d'une série de régressions multiples, le type de modèle de régression variant selon le type de variable imputée. Les covariables englobent toutes les autres variables observées ou imputées pour l'unité en question. Les imputations sont définies comme des tirages de la distribution prédictive a posteriori exigée par le modèle de régression avec une distribution a priori plate ou non informative pour les paramètres du modèle de régression. La séquence d'imputation des valeurs manquantes peut se poursuivre d'une façon cyclique, se superposant chaque fois aux valeurs tirées antérieurement, entraînant une interdépendance des valeurs imputées et misant sur la structure corrélative des covariables. Pour la création d'imputations multiples, on peut appliquer la même procédure avec différents points de départ aléatoires ou en prenant chaque P^e série de valeurs imputées des cycles mentionnés ci-dessus.

Les variables du fichier de données sont considérées comme relevant de l'un ou l'autre de cinq types : 1) continu; 2) binaire, 3) catégorique (polytomique avec plus de deux catégories), 4) de type comptage, 5) mixte (variable continue à masse de probabilité non nulle de 0). Du point de vue des calculs, les variables binaires et catégoriques se laissent traiter de façon identique, mais le fait de les distinguer facilite la conceptualisation et la description de l'algorithme de base. De plus, la population est considérée comme essentiellement infinie, l'échantillon étant simple et aléatoire et le mécanisme de données manquantes étant ignorable (Rubin 1976). Le recours à une imputation multiple en présence d'un plan complexe n'a toujours pas été étudié à fond et dépasse le cadre du présent exposé.

Nous décrivons ci-dessous la stratégie d'imputation multidimensionnelle par régression séquentielle (IMRS) à la section 2 et, aux sections 3 et 4, nous en évaluons deux applications. Il est difficile, pour la première application, de postuler une distribution multidimensionnelle composée à

cause des liens systématiques complexes entre les variables et les restrictions. Pour la deuxième application, on peut utiliser un modèle d'emplacement général pour créer des imputations multiples (Olkin et Tate 1961 et Little et Schluchter 1985). Nous comparons donc des inférences d'imputations multiples résultant de la stratégie IMRS à des inférences résultant d'un modèle multidimensionnel composé. La section 5 contient les résultats d'une étude de simulation portant sur les propriétés d'échantillonnage d'inférences tirées de données imputées. Pour terminer, nous discutons de l'orientation des recherches à venir à la section 6.

2. MÉTHODE D'IMPUTATION

Pour un échantillon de taille n, nous notons X un plan ou une matrice explicative $n \times p$ contenant toutes les variables n'ayant pas de valeurs manquantes. X comporte des variables continues, binaires, de type comptage ou mixtes, ainsi que des variables fictives appropriées représentant des variables catégoriques. De plus, X peut comporter une colonne de uns afin de modéliser un paramètre de coordonnée à l'origine, des variables de décalage et certaines variables de plan. Soit $Y_1, Y_2, ..., Y_k$, des variables k comportant des valeurs manquantes, ordonnées, en toute généralité, selon le nombre de valeurs manquantes, des moins nombreuses aux plus nombreuses. Le schéma n'est pas nécessairement monotonique. (Dans un schéma monotonique de valeurs manquantes, Y_2 est observée uniquement pour un sous-ensemble de sujets pour lesquels Y_1 est observée, tandis que Y_3 est observée uniquement pour un sous-ensemble de ceux pour lesquels Y_2 est observée, et ainsi de suite.)

Pour des imputations modélisées, la densité conditionnelle composée de $Y_1, Y_2, ..., Y_k$ compte tenu de X peut être factorisée comme suit

$$f(Y_1, Y_2, ..., Y_k, | X, \theta_1, \theta_2, ..., \theta_k) = f_1(Y_1 | X, \theta_1) f_2(Y_2 | X, Y_1, \theta_2) ... f_k(Y_k | X, Y_1, Y_2, ..., Y_{k-1}, \theta_k) \quad (1)$$

où $f_j, j = 1, 2, ..., k$ sont les fonctions de densité conditionnelle et θ_j est un vecteur de paramètres de la distribution conditionnelle (par exemple, coefficients de régression et paramètres de dispersion). Dans le contexte d'une enquête sur échantillon, on peut considérer cela comme un modèle de superpopulation. Nous modélisons chaque densité conditionnelle à l'aide d'un modèle de régression approprié comportant des paramètres inconnus, θ_j, et nous puisons dans la distribution prédictive correspondante des valeurs manquantes compte tenu des valeurs observées. Nous supposons que la distribution a priori pour les paramètres $\theta = (\theta_1, \theta_2, ..., \theta_k)$ est $\pi(\theta) \propto 1$ (diffuse relativement à la vraisemblance). Toutefois, la méthode se laisse facilement modifier pour certaines distributions a priori convenables.

Chaque régression conditionnelle se fonde sur un des modèles ci-dessous :

1. un modèle de régression linéaire normale à une échelle appropriée (par exemple, une transformée exponentielle de Box-Cox peut servir à atteindre la normalité) si Y_j est continue;
2. un modèle de régression logistique si Y_j est binaire;
3. un modèle de régression logit polytomique ou généralisée si Y_j est catégorique;
4. un modèle linéaire logarithmique de Poisson si Y_j est une variable de comptage;
5. un modèle à deux degrés dont la valeur nulle-non nulle est imputée par régression logistique; pour une valeur non nulle, un modèle de régression linéaire normale sert à imputer des valeurs non nulles, si Y_j est composée.

Chaque imputation est constituée de c « cycles ». On commence le cycle 1 par régression de la variable comportant le plus petit nombre de valeurs manquantes, Y_1 sur X, par imputation des valeurs manquantes en fonction du modèle de régression approprié. Si l'on suppose une distribution a priori plate pour les coefficients de régression, les imputations, pour les valeurs manquantes en Y_1 sont les tirages de la distribution prédictive a posteriori correspondante (on trouvera à l'annexe A des détails sur le tirage de valeurs pour divers modèles de régression). Il s'agit alors de mettre X à jour en annexant Y_1 de façon appropriée (par exemple des variables fictives, si elle est catégorique) et de passer à la prochaine variable, Y_2, occupant le rang suivant parmi les valeurs manquantes les moins nombreuses. Le processus d'imputation est répété à l'aide de X mises à jour à titre de variables explicatives jusqu'à ce que toutes les variables aient été imputées. Autrement dit, il y a régression de Y_1 sur $U = X$; de Y_2 sur $U = (X, Y_1)$ où Y_1 comporte des valeurs imputées; de Y_3 sur $U = (X, Y_1, Y_2)$ où Y_1 et Y_2 comportent des valeurs imputées, et ainsi de suite.

Le processus d'imputation est alors répété pour les cycles 2 à c, la série de variables explicatives étant modifiée pour inclure toutes les variables Y sauf celle qui sert de variable dépendante. Ainsi, il y a régression de Y_1 sur X et $Y_2, Y_3, ... Y_k$; il y a régression de Y_2 sur X et $Y_1, Y_3, ..., Y_k$; et ainsi de suite. Les cycles répétés se succèdent un nombre de fois déterminé d'avance, ou jusqu'à ce que l'on ait des valeurs imputées stables.

La procédure décrite ci-dessus doit être modifiée si l'on veut incorporer des restrictions et des limites. Les restrictions sont traitées par ajustement des modèles à un sous-ensemble approprié d'unités. Par exemple, un modèle de régression de Poisson peut être appliqué à l'imputation de valeurs manquantes pour la variable « nombre de grossesses ». L'imputation se limitera aux femmes de

l'échantillon. À titre de covariable, toutefois, cette variable pourra être traitée différemment lors de l'imputation de variables subséquentes. Ainsi, certaines variables fictives pourront être créées en fonction de cette variable, puis annexées à la matrice U avant que l'imputation de la variable suivante ne se poursuive.

Considérons un autre exemple, « années d'usage de la cigarette », l'échantillon se limitant à des personnes qui fument ou qui ont fumé. En l'absence d'indication que ces personnes ont fumé au cours de leur adolescence, « années d'usage de la cigarette » devra pour une personne qui fume actuellement satisfaire la limite (0, Âge - 18). S'il y a lieu de croire que la personne a fumé au cours de son adolescence, on pourra restreindre l'étendue, par exemple : (0, Âge - 12). Pour une personne qui a déjà fumé, ces étendues seront (0, Âge - 18 – YRSQUIT) et (0, Âge - 12 – YRSQUIT), respectivement, où YRSQUIT représente le nombre d'années qui se sont écoulées depuis que la personne a cessé de fumer. Le modèle de régression approprié pour cette variable est une version tronquée du modèle de régression linéaire normale (possiblement en fonction d'une échelle transformée). Les paramètres, les coefficients de régression et la variance résiduelle doivent être tirés des distributions a posteriori correspondantes. Les imputations sont alors tirées de la distribution normale tronquée correspondante en fonction de la valeur tirée des paramètres.

Il est difficile de tirer des valeurs de paramètres directement de leur distribution a posteriori selon des vraisemblances normales tronquées. Toutefois, le calcul est facile pour une valeur de paramètre donnée. L'algorithme SIR (échantillonnage-importance-rééchantillonnage) (Rubin 1987b; Raghunathan et Rubin 1988) permet de puiser dans la distribution a posteriori elle-même. Tout d'abord, on tire plusieurs valeurs de paramètre d'essai de la distribution a posteriori sans appliquer les limites (modèle de régression linéaire normale non tronquée). Deuxièmement, on rattache un coefficient d'importance à chaque valeur d'essai, défini comme le rapport entre la densité a posteriori réelle avec limites et la densité d'essai (la densité a posteriori sans limites), les deux étant évaluées à la valeur tirée. Enfin, on échantillonne de nouveau une même valeur de paramètre avec probabilité proportionnelle aux coefficients d'importance. Cette méthode exige une surveillance soignée de la distribution des coefficients d'importance (Gelman, Carlin, Stern et Rubin 1995).

Les limites s'appliquent également à des variables polytomiques. Ainsi, supposons qu'une variable Y puisse avoir une valeur k quelconque, mais que les données observées indiquent que la valeur manquante pour un sujet particulier peut être soit j ou l. Le rôle de ce sujet dans la vraisemblance correspond à la distribution binomiale conditionnelle. Les tirages de l'étape multinomiale (voir l'annexe A) se font à partir de la distribution conditionnelle pour ces deux catégories. Autrement dit, la valeur imputée est j avec probabilités $s_{j\cdot} = P_{j\cdot}/(P_{j\cdot} + P_{l\cdot})$ et l avec la probabilité $1 - s_{j\cdot}$.

À la fin du premier cycle d'imputations, on a le premier fichier de données complètes sans valeurs manquantes. La factorisation à l'équation (1) définit une distribution conditionnelle composée de $Y_1, Y_2, ..., Y_k$; étant donné X. Si le profil des données manquantes est monotone, les imputations du premier cycle sont des tirages approximatifs de la densité prédictive a posteriori composée des valeurs manquantes compte tenu des valeurs observées. À noter que les tirages des variables logistiques, polytomiques et de comptage proviennent d'approximations (pour de grands échantillons) de la densité a posteriori des coefficients de régression. Il est possible d'améliorer ces approximations en ayant recours, par exemple, à l'algorithme SIR ou à un autre algorithme de rejet à chaque cycle subséquent.

Lorsque le profil des données manquantes n'est pas monotone, il est possible d'élaborer un algorithme d'échantillonnage de Gibbs (Geman et Geman 1984, Gelfand et Smith 1990) qui correspond au modèle (1). Ainsi, moyennant les valeurs tirées des paramètres $\theta_2, \theta_3, ..., \theta_k$ et les valeurs manquantes tirées du premier cycle, le deuxième cycle tirerait des valeurs de θ_1 de la densité a posteriori conditionnelle appropriée qui est proportionnelle au premier terme de l'équation (1). Il s'agit ensuite de tirer les valeurs manquantes en Y_1 moyennant cette valeur tirée du paramètre θ_1, toutes les autres valeurs observées ou imputées pour ce sujet et d'autres paramètres $\theta_2, \theta_3, ..., \theta_k$ du modèle. Autrement dit, les valeurs manquantes en Y_j au cycle $(t+1)$ doivent être tirées de la densité conditionnelle

$$f_j^*(Y_j | \theta_1^{(t+1)}, Y_1^{(t+1)}, ..., \theta_j^{(t+1)}, \theta_{j+1}^{(t)}, Y_{j+1}^{(t)}, ..., \theta_k^{(t)}, Y_k^{(t)}, X), \quad (2)$$

calculée en fonction de la distribution composée en (1), où $Y_l^{(t)}$ représente les valeurs imputées ou observées pour la variable Y_l au cycle t. Bien que cela soit conceptuellement possible, il est difficile de calculer même cette densité dans la plupart des situations concrètes vu les restrictions, les limites et le type de variables à l'étude.

Nous proposons un tirage des valeurs manquantes en Y_j au cycle $(t+1)$ à partir d'une distribution prédictive correspondant à la densité conditionnelle,

$$g_j(Y_j | Y_1^{(t+1)}, Y_2^{(t+1)}, ..., Y_{j-1}^{(t+1)}, Y_{j+1}^{(t)}, ..., Y_k^{(t)}, X, \phi_j), \quad (3)$$

où la densité conditionnelle g_j est définie par l'un des modèles de régression décrits antérieurement qui dépend du type de variable pour Y_j, et ϕ_j représente les paramètres de régression inconnus ayant une distribution a priori diffuse. Autrement dit, les nouvelles valeurs imputées pour une variable dépendent des valeurs imputées antérieurement pour d'autres variables, et des valeurs nouvellement imputées de variables qui ont précédé la variable imputée présentement. Cette proposition peut être considérée comme une approximation d'un échantillonnage réel de Gibbs où la densité conditionnelle (3) fournit une approximation de la densité conditionnelle (2). De plus, on peut améliorer cette approximation en considérant l'algorithme SIR ou un autre algorithme de type rejet si la densité conditionnelle en (2) peut être calculée jusqu'à une constante.

Il existe d'autres cas particuliers dans lesquels cette approximation est l'équivalent du tirage de valeurs d'une distribution prédictive a posteriori relevant d'un modèle complètement paramétrique. Si donc toutes les variables sont continues et si chaque modèle de régression conditionnelle est un modèle de régression linéaire normale à variance constante, il y a convergence de l'algorithme vers une distribution prédictive composée relevant d'une distribution normale multidimensionnelle comportant une distribution a priori irrégulière pour la moyenne et la matrice des covariances.

Il est théoriquement possible qu'une série de tirages fondés sur les densités en (3) ne converge pas vers une distribution stationnaire, car ces densités conditionnelles ne sont peut-être compatibles avec aucune distribution conditionnelle composée multidimensionnelle de $Y_1, Y_2, \ldots, Y_k$ étant donné X (Gelman et Speed 1993). Nos études empiriques fondées sur plusieurs fichiers de données concrètes n'ont toujours pas permis d'identifier ce genre d'anomalie. Dans plusieurs grands fichiers de données, les densités conditionnelles (2) et (3) paraissent assez semblables. Comme il a été mentionné aux sections 4 et 5, les tirages axés sur cette stratégie sont comparables à ceux qui se fondent sur un modèle bayésien explicite.

3. EFFET DE L'USAGE DU TABAC SUR L'ARRÊT CARDIAQUE PRIMAIRE

Dans notre première illustration, la stratégie IMRS est appliquée à une étude de cas-témoins portant sur la relation entre l'usage de la cigarette et l'incidence de l'arrêt cardiaque primaire (Siscovick, Raghunathan, King, Weinmann, Wicklund, Albright, Bovbjerg, Arbogast, Kushi, Cobb, Copass, Psatsy, Retzlaff, Childs and Knopp 1995). Dans cette étude, il est difficile de formuler un modèle explicite qui englobe toute la complexité des données. Les sujets de cas étaient tous des résidents de King County, Washington, ayant subi un arrêt cardiaque primaire à l'extérieur de l'hôpital entre 1988 et 1994. Les sujets de cas ont été identifiés à l'aide d'un examen des rapports d'incidents paramédicaux. Les sujets témoins ont été sélectionnés à l'aide d'un sondage téléphonique au hasard (King County) et appariés à des sujets de cas en fonction du sexe et de l'âge (à sept ans près). Pour être admissibles, les sujets (cas et témoins) devaient être âgés de 25 à 74 ans, mariés et libres de toute maladie du coeur (diagnostic clinique) ou de tout autre péril comme un cancer, une maladie du foie, une maladie des poumons, ou encore une insuffisance rénale terminale.

Puisque l'arrêt cardiaque primaire comporte un taux de létalité supérieur à 80 %, le fait d'être marié a été ajouté comme critère d'admissibilité afin que l'information sur l'exposition au facteur de risque (état de fumeur, années d'usage) puisse être confirmée par les répondants substituts (conjoints). Parmi les sujets témoins et les sujets de cas de survie, on a interviewé des sujets et des substituts afin de recueillir des données sur l'exposition. On a interviewé les sujets témoins et les sujets de cas de survie principalement afin d'étudier la fiabilité des mesures obtenues de leurs substituts. Parmi les variables considérées dans le présent exposé, il n'y avait pratiquement aucune différence entre les mesures obtenues des sujets et de leurs substituts pour les témoins et les sujets de cas.

Le tableau 1 contient les moyennes, les écarts types et les valeurs manquantes en pourcentage pour des variables clés selon l'état de cas-témoins. Les variables d'exposition sont des variables indicatrices pour les personnes qui ont déjà fumé (X_1), les personnes qui fument actuellement (X_2) et les années d'usage du tabac (X_3). Les variables explicatives considérées sont l'âge, l'indice de masse corporelle (BMI=Poids [en kg]/Taille^2[en mètres]) et les variables binaires Femme (Female) et Éducation (diplôme d'études secondaires). Le modèle de fond qui nous intéresse est le modèle de régression logistique,

$$\log[\Pr(C=1)/\Pr(C=0)] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1 X_3 + \alpha_4 X_2 X_3 + \alpha_5 \text{Âge} + \alpha_6 \text{BMI} + \alpha_7 \text{Femme} + \alpha_8 \text{Éducation},$$

où C est un indicateur de l'arrêt cardiaque. Les résultats préliminaires indiquent que des termes linéaires pour l'âge et l'indice de masse corporelle sont appropriés.

Tableau 1
Moyennes et proportions (en %) des variables clés et pourcentage manquant

Variable	Témoin (n = 551)		Cas (n = 347)	
	% manquant	Moyenne (ET)	% manquant	Moyenne (ET)
Âge	0,0	58,4 (10,4)	0,0	59,4 (9,9)
Indice de masse corporelle	8,2	25,8 (4,1)	2,6	26,4 (4,6)
Années d'usage du tabac	16,8	24,8 (14,7)	5,4	31,7 (13,8)
		Proportion		Proportion
Femme	0,0	23,2	0,0	19,9
≥ École sec.	0,0	76,8	0,0	61,9
État de fumeur				
N'a jamais fumé	0,0	47,2	0,0	27,3
A déjà fumé	0,0	42,1	0,0	38,2
Fume actuellement	0,0	10,7	0,0	34,5

Il n'y a pas de valeurs manquantes pour les variables Âge, Femme, Éducation, État de fumeur (X_1, X_2) et C. Ainsi, pour ce qui est de l'imputation, définissons X = (1, Âge, Femme, Éducation, X_1, X_2, C). Log (BMI), avec le moins de valeurs manquantes, a d'abord subi une régression sur X en fonction d'un modèle de régression linéaire normale. Des diagnostics résiduels ont indiqué qu'une transformation logarithmique améliorait la normalité des résidus.

Ensuite, les années d'usage du tabac ont subi une régression sur $U = (X, \log(\text{BMI}))$. Pour cette variable, l'échantillon se limitait à des personnes qui fumaient ou qui avaient déjà fumé. De plus, les valeurs imputées pour les années d'usage du tabac étaient limitées par AGE-18, sauf lorsque le répondant avait fumé à l'école (SCHSMK), et elles étaient alors limitées par AGE-12. Pour les personnes qui avaient déjà fumé, les valeurs imputées étaient également limitées par le nombre d'années écoulées depuis la cessation (YRSQUIT). Ainsi, les valeurs imputées pour les personnes qui avaient déjà fumé, mais non à l'école, étaient limitées par AGE-18–RSQUIT, tandis que les valeurs imputées pour les personnes qui avaient déjà fumé, y compris à l'école, étaient limitées par AGE-12-YRSQUIT. Certains sujets (5 %) avaient des valeurs manquantes pour les deux variables auxiliaires (SCHSMK, YRSQUIT) imputées avant que l'on ne définisse les limites supérieures des années d'usage du tabac. À cause de la structure inhérente de ce fichier de données, il est difficile d'élaborer explicitement une distribution composée des variables ayant des valeurs manquantes dépendant des variables observées complètement. La stratégies IMRS est donc une façon attrayante de traiter ce genre de données.

Pour l'imputation des valeurs manquantes, 1 000 cycles ont été exécutés pour chacun des 25 points de départ aléatoires, ce qui a donné $M = 25$ imputations. Le modèle de régression logistique a été ajusté à chaque fichier de données imputées de façon à donner des estimations correspondant à un maximum de vraisemblance pour les coefficients de régression et les matrices de covariances asymptotiques.

Nous avons utilisé la formule de variance par imputation multiple standard (Rubin 1987a, chapitre 3) afin de calculer l'estimation polyimputée des coefficients de régression et de la matrice des covariances. En résumé, supposons que $\hat{\alpha}^{(l)}$ est l'estimation du vecteur de coefficients de régression α du modèle logistique, et $V^{(l)}$ sa matrice des covariances, en fonction du fichier de données imputées l. L'estimation polyimputée de α est

$$\hat{\alpha}_{\text{MI}} = \sum_{l=1}^{M} \hat{\alpha}^{(l)} / M$$

et sa matrice des covariances est

$$V_{\text{MI}} = \sum_{l=1}^{M} V^{(l)} / M + \frac{M+1}{M} B_M$$

où

$$B_M = \sum_{l=1}^{M} (\hat{\alpha}^{(l)} - \hat{\alpha}_{\text{MI}})(\hat{\alpha}^{(l)} - \hat{\alpha}_{\text{MI}})' / (M-1)$$

Le nombre d'imputations dépasse la valeur recommandée. Nous avons exécuté 25 imputations avec différents points de départ aléatoires afin de déterminer si les cycles de type Gibbs mènent à une région des valeurs imputées qui est très différente des valeurs observées. Des affichages graphiques des valeurs imputées et observées ont indiqué qu'aucune des imputations des 25 000 cycles n'était incompatible avec la distribution des données observées.

Le tableau 2, l'analyse de cas complets, contient les estimations ponctuelles et leurs erreurs types fondées sur des sujets dont toutes les valeurs sont observées. Au total, 103 sujets (11,5 %) avaient des valeurs manquantes pour une ou plusieurs variables explicatives. Une analyse de cas complets, qui n'est généralement valide que lorsque les données sont manquantes tout à fait au hasard, a été exécutée après l'élimination de ces 103 sujets (voir la colonne 2, tableau 2). Des analyses de régression logistique comportant un indicateur de données manquantes à titre de

Tableau 2

Estimations ponctuelles (erreurs types) des coefficients de régression logistique pour le modèle de l'arrêt cardiaque primaire pour des cas complets, méthodes IMRS 1* et 2**

Variables explicatives	Cas complets		IMRS			
	(n = 795)		Méthode 1 (n = 898)		Méthode 2 (n = 898)	
	Estimation	(ET)	Estimation	(ET)	Estimation	(ET)
Ordonnée à l'origine	-2,922	(0,791)	-2,610	(0,757)	-2,348	(0,627)
Âge	0,015	(0,009)	0,015	(0,009)	0,014	(0,008)
Femme	-0,007	(0,203)	-0,115	(0,189)	-0,119	(0,177)
Éducation	-0,448	(0,173)	-0,467	(0,166)	-0,444	(0,133)
Indice de masse corporelle	0,056	(0,018)	0,049	(0,013)	0,055	(0,009)
Personne qui fume	1,693	(0,569)	2,001	(0,543)	1,998	(0,448)
Personne qui a fumé	0,003	(0,284)	-0,029	(0,262)	-0,011	(0,223)
Personne qui fume × Années d'usage	-0,003	(0,015)	-0,008	(0,013)	-0,005	(0,011)
Personne qui a fumé × Années d'usage	0,019	(0,009)	0,014	(0,009)	0,014	(0,009)

* Méthode 1 – Imputation limitée à des variables de modèle

** Méthode 2 – Imputation comprenant des variables modélisées et auxiliaires

variable dépendante et un certain nombre de variables observées complètement à titre de variables explicatives ont indiqué que les données ne manquent pas complètement au hasard. On peut donc s'attendre à ce que les estimations de cas complets et les erreurs types soient biaisées.

Le tableau 2 (IMRS, méthode 1) contient des estimations et leurs erreurs types pour l'IMRS d'après les variables du modèle de fond seulement. Ces estimations sont assez semblables à celles de l'analyse de cas complets. Les erreurs types de l'imputation multiple sont plus petites à cause des sujets additionnels ayant des données imputées. Il y a de faibles changements du rapport entre l'usage du tabac et l'arrêt cardiaque primaire. L'analyse de cas complets indique un rapport statistiquement significatif entre les années d'usage du tabac et l'arrêt cardiaque primaire pour des personnes qui ont déjà fumé, tandis qu'une telle association n'est pas révélée par l'analyse des données polyimputées.

Un des avantages de la stratégie d'imputation multiple est que le processus d'imputation peut faire appel à des variables additionnelles qui ne se trouvent pas dans l'analyse de fond. De telles situations se présentent lorsqu'une base de données de recherche commune comportant plusieurs variables est utilisée par différents chercheurs, ayant chacun recours à un sous-ensemble des variables. L'imputation peut se faire pour la base de données entière, la prédiction des valeurs manquantes pour chaque variable étant renforcée par toutes les autres variables du fichier de données. On a pu montrer que de telles imputations améliorent l'efficacité comparativement à celles qui se fondent uniquement sur les variables d'un modèle de fond particulier (Raghunathan et Siscovick 1996).

Le tableau 2 (IMRS, méthode 2) contient des estimations de l'imputation multiple et leurs erreurs types obtenues lorsque le fichier de données entier a été imputé en fonction de 50 variables additionnelles. Celles-ci comprenaient des indicateurs diététiques, des mesures physiologiques, des situations socioéconomiques et des variables de comportement. Les estimations ponctuelles sont légèrement différentes pour toutes les variables. Les erreurs types, par contre, sont appréciablement plus petites comparativement à la stratégie d'imputation multiple fondée uniquement sur des variables du modèle de fond (IMRS, méthode 1). Il n'y a là rien de surprenant car plusieurs des variables additionnelles, par exemple la tension artérielle, le compte de cholestérol, la consommation d'alcool et l'activité physique, étaient fortement prédictives de l'indice de masse corporelle et des variables liées à l'usage du tabac.

4. TROUBLES PSYCHOLOGIQUES PARENTAUX ET DÉVELOPPEMENT CHEZ L'ENFANT

Une deuxième illustration examine les effets des troubles psychologiques parentaux sur plusieurs mesures du développement chez l'enfant. Little et Schuchter (1985) ont analysé les données à l'aide d'un modèle d'emplacement général afin d'obtenir des estimations correspondant à un maximum de vraisemblance des paramètres de la distribution composée. On a utilisé ce modèle d'emplacement général pour créer des imputations multiples en employant des méthodes de Monte Carlo à chaînes markoviennes (Schafer 1997), ce qui a donné des fichiers de données polyimputées entièrement à base de modèle bayésien. Nous avons également créé des imputations multiples à l'aide de la procédure IMRS.

Les données de l'étude se rapportent à 69 familles ayant chacune deux enfants. Chaque famille a été classée dans une des trois catégories de risque suivantes : 1) risque normal : aucun trouble psychiatrique parental; 2) risque modéré : diagnostic chez un parent d'un trouble psychiatrique ou d'une maladie physique chronique; 3) risque élevé : diagnostic chez un parent de schizophrénie ou de trouble mental affectif. Il y avait trois variables dépendantes primaires d'intérêt : Y_{1c}, nombre de symptômes psychiatriques (dichotomisation : élevé/faible) chez l'enfant c; Y_{2c}, test de lecture normalisé chez l'enfant c; et Y_{3c}, test normalisé de compréhension verbale chez l'enfant c.

Nous considérons trois modèles pour l'étude de l'effet des troubles psychologiques parentaux sur le développement des enfants. Le premier est un modèle de régression logistique à effets mixtes :

$$\text{logit}[\Pr(Y_{1ic} = 1)] = \beta_0 + \beta_1 U_{1i} + \beta_2 U_{2i} + \gamma_i,$$

où $Y_{1ic} = 1$ si un enfant c d'une famille i est considéré comme ayant un nombre élevé de symptômes et 0 autrement; $U_{1i} = 1$ si une famille i est considérée comme étant exposée à un risque modéré et 0 autrement; $U_{2i} = 1$ si une famille i est considérée comme étant exposée à un risque élevé et 0 autrement; et γ_i sont des effets aléatoires considérés comme étant des variables aléatoires normales distribuées de façon identique et indépendante avec une moyenne 0 et une variance ϕ_γ^2. Cet effet aléatoire tient compte d'une corrélation intraclasse entre deux enfants au sein d'une même famille. Pour des données complètes, on peut ajuster ce modèle en maximisant la fonction de vraisemblance numériquement intégrée de $(\beta_0, \beta_1, \beta_2, \phi_\gamma^2)$ à l'aide de l'algorithme de Newton-Raphson et de la méthode de quadrature gaussienne pour l'intégration numérique de la fonction de vraisemblance. Il est facile d'ajuster ce type de modèle avec des données complètes, mais difficile de le faire avec des valeurs manquantes.

Les deuxième et troisième modèles de régression établissent le rapport entre les notes de l'enfant en lecture et en compréhension verbale, respectivement, et le groupe de risque après ajustement selon le nombre de symptômes (Y_1). Une étude des valeurs résiduelles après quelques cycles préliminaires ou imputations de notes en lecture et en compréhension verbale a indiqué qu'une échelle logarithmique était appropriée. Ainsi, en notant Y_{2ic} et Y_{3ic} comme logarithme des notes en lecture et en compréhension verbale, respectivement, pour un enfant c d'une famille i,

nous avons énoncé le modèle de régression à effets mixtes ci-dessous,

$$Y_{2ic} = \alpha_0 + \alpha_1 U_{1i} + \alpha_2 U_{2i} + \alpha_3 Y_{1ic} + \delta_i + \epsilon_{ic}.$$

où δ_i et ϵ_{ic} sont des variables aléatoires normales indépendantes l'une de l'autre avec une moyenne 0 et des variances σ_δ^2 et σ_ϵ^2 respectivement. Encore une fois, en l'absence de données manquantes dans les covariables, il est facile d'obtenir les estimations correspondant à un maximum de vraisemblance des paramètres inconnus en utilisant, par exemple, la procédure PROC MIXED de SAS.

Il n'y avait pas de valeurs manquantes dans la classification des groupes de risque, et nous avons donc défini $X = (1, U_1, U_2)$. Les variables comportant des valeurs manquantes, Y_{21}, Y_{22}, Y_{31} et Y_{32} ont été imputées par régression linéaire normale, et les valeurs manquantes en Y_{11} et Y_{12} ont été imputées par régression logistique. Nous avons créé M=25 IMRS, en répétant le processus pour 1 000 cycles et 25 points de départ différents. Les fichiers de données polyimputées IMRS ont été analysés et combinés à l'aide des méthodes décrites antérieurement. Afin de comparer ces résultats aux inférences polyimputées lorsque les imputations sont tirées de la distribution prédictive a posteriori relevant du modèle d'emplacement général, nous avons créé 25 imputations relevant d'un modèle entièrement bayésien à l'aide d'un logiciel préparé par Schafer (1997). Le tableau 3 contient les estimations ponctuelles et les erreurs types des trois modèles fondés sur stratégies d'imputation multiple IMRS et de Bayes. Il n'existe pas de différences réelles significatives entre les estimations et les erreurs types IMRS d'une part et celles qui résultent de l'imputation bayésienne d'autre part. Les enfants de parents du groupe à risque élevé ont environ 7,8 [exp (2,048)] fois plus de chances d'avoir un nombre élevé de symptômes que les enfants de parents du groupe normal dans le cadre de l'IMRS. L'intervalle de confiance de 95 % pour ce risque relatif est de (3,8, 16,0). Pour le groupe à risque modéré, les estimations ponctuelle et d'intervalles correspondantes sont de 3,7 et de (1,8, 7,8). Ces estimations peuvent être comparées à celles que fournit l'analyse de cas complets (non indiquées); 7,4 (2,3, 24,2) pour le groupe à risque élevé, et 3,5 (1,0, 11,9) pour le groupe à risque modéré (données non indiquées). Même si les estimations ponctuelles des risques relatifs sont semblables, les intervalles de confiance de cas complets sont plus grands parce qu'ils se fondent uniquement sur 60 % des observations.

D'après les coefficients de régression estimés du tableau 3, on peut inférer, après ajustement, pour le nombre de symptômes, que les enfants des groupes à risque modéré et élevé ont des notes plus faibles en lecture, de 11 points environ [exp (4,654)-exp(4,654-0,110)], comparativement au groupe normal. D'autre part, l'analyse de cas complets donne comme estimation pour les enfants du groupe à risque modéré une note inférieure de 16 points à celle de leurs homologues du groupe normal, et pour les enfants du groupe à risque élevé une note inférieure de 19 points environ à celle du groupe normal.

L'analyse IMRS des notes en compréhension verbale indique que les enfants des groupes à risque modéré et élevé ont des notes inférieures de 20 et de 24 points environ, respectivement, à celles de leurs homologues du groupe normal. Toutefois, l'analyse de cas complets indique que les notes du groupe à risque modéré sont inférieures de 36 points et celles du groupe à risque élevé de 39 points environ comparativement au groupe normal. Ainsi, les estimations pour les cas complets des effets des troubles psychologiques parentaux sur les notes de l'enfant en lecture et en compréhension verbale sont assez différentes de celles que l'on obtient par l'analyse des données polyimputées. Cela n'est pas surprenant car les données sur les notes en lecture et en compréhension verbale ne manquent pas tout à fait au hasard; elles sont liées au groupe à risque aussi bien qu'au nombre de symptômes chez l'enfant.

Tableau 3

Estimations ponctuelles (erreurs types) des coefficients de régression pour trois modèles de développement chez l'enfant dans le cadre d'une imputation IMRS et de Bayes

Variables explicatives	Méthode d'imp.	Variable dépendante					
		Symptômes		Note en lecture		Note en compréhension verbale	
Ordonnée à l'origine	IMRS	-0,678	(0,256)	4,654	(0,013)	4,873	(0,020)
	Bayes	-0,688	(0,257)	4,556	(0,013)	4,991	(0,021)
Groupe à risque élevé	IMRS	2,048	(0,356)	-0,109	(0,022)	-0,191	(0,032)
	Bayes	2,033	(0,350)	-0,108	(0,021)	-0,180	(0,033)
Groupe à risque modéré	IMRS	1,289	(0,366)	-0,110	(0,022)	-0,162	(0,033)
	Bayes	1,300	(0,360)	-0,109	(0,023)	-0,167	(0,035)
Symptômes	IMRS	-		0,032	(0,022)	-0,083	(0,032)
	Bayes	-		0,031	(0,019)	-0,080	(0,030)

5. ÉTUDE PAR SIMULATION

Les analyses décrites aux sections 3 et 4 indiquent que l'on peut obtenir des résultats raisonnables en appliquant la stratégie IMRS au traitement des données manquantes. Néanmoins, il est difficile de déterminer, d'après de telles études de cas, si la stratégie donnera des inférences valides dans des applications courantes. Une étude de simulation a été conçue afin d'examiner les propriétés d'échantillonnage répété des inférences tirées de fichiers de données imputées créés à l'aide de la stratégie IMRS. On a tiré de populations hypothétiques des fichiers de données complètes et on a supprimé des éléments en vertu d'un mécanisme de données manquantes ignorables. On a imputé les valeurs supprimées et on a évalué les différences de statistiques sommaires d'après les fichiers de données imputées et les fichiers de données antérieures à la suppression ou complètes.

Plus précisément, la stratégie a servi :

(1) à préparer un fichier de données complètes qui ne correspondait pas tout à fait à notre stratégie d'imputation multiple,

(2) à estimer des paramètres de régression choisis,

(3) à supprimer certaines valeurs à l'aide d'un mécanismes de données manquantes ignorables,

(4) à utiliser l'IMRS en vue de la polyimputation des valeurs manquantes, et

(5) à obtenir des estimations polyimputées pour les paramètres de régression estimés à l'étape 2.

Les différences du paramètre sont examinées pour plusieurs répétitions indépendantes de cette stratégie.

Au total, on a préparé 2 500 fichiers de données complètes comportant trois variables (U, Y_1, Y_2) et une taille d'échantillon de 100 à l'aide des modèles ci-dessous :

1. U ~ Normal (0, 1);
2. Y_1 ~ Gamma avec une moyenne $\mu_1 = \exp(U\text{-}1)$ et une variance $\mu_1^2/5$; et
3. Y_2 ~ Gamma avec une moyenne $\mu_2 = \exp(-1 + 0{,}5U + 0{,}5\ Y_1)$ et une variance $\mu_2^2/2$.

Le modèle pour Y_2 à l'étape 3 est le modèle de régression primaire d'intérêt avec de vrais coefficients de régression $\beta_0 = -1, \beta_1 = \beta_2 = 0{,}5$, et un paramètre de dispersion $\varphi^2 = 0{,}5$. Pour les données complètes, on peut ajuster ce modèle à l'aide de progiciels statistiques comme GLIM ou Splus.

Les mécanismes de suppression ou de données manquantes étaient les suivants :

(1) aucune valeur manquante en U;

(2) des valeurs manquantes en Y_1 qui dépendent de U suivant une fonction logistique logit [Pr(Y_1 manquant)] = 1,5 + U; et

(3) des valeurs manquantes en Y_2 qui dépendent de U et de Y_1 suivant une fonction logistique logit [Pr(Y_2 manquant)]= 1,5 - 0,5 Y_1 - 0,5 U.

Ces mécanismes de données manquantes ou ont généré 22 % de données manquantes en Y_1 et 29 % de données manquantes en Y_2. L'analyse de cas complets n'aurait utilisé que 48 % des données.

Puisque l'IMRS nous permet seulement d'ajuster un modèle de régression linéaire normale, les imputations ont été exécutées comme suit. Supposons que Y_1 a moins de valeurs manquantes, et notons $Z_1 = (Y_1^{\lambda_1} - 1)\ /\lambda_1$, la transformée de Box-Cox de la variable continue. Dans le premier cycle d'imputations, supposons que Z_1 a une distribution normale avec une moyenne $a_0 + a_1 U$ et une variance σ_1^2, où l'on estime λ_1 à l'aide de la stratégie du maximum de vraisemblance, et que $Z_2 = ((Y_2^{\lambda_2} - 1)/\lambda_2$ a une distribution normale avec une moyenne $b_0 + b_1 U + b_2 Z_1$ et une variance σ_2^2, où l'on estime λ_2 à l'aide du maximum de vraisemblance. Pour les cycles subséquents, U et Z_2 sont des variables explicatives pour Z_1, et U et Z_1 sont des variables explicatives pour Z_2. L'estimation d'une transformée exponentielle à l'aide du maximum de vraisemblance a été automatisée au moment d'ajuster chaque modèle de régression.

Pour chacun des 2 500 fichiers de données simulées comportant des valeurs manquantes, on a créé au total 250 cycles ayant M=5 différents points de départ aléatoires à l'aide de l'IMRS. Pour chaque répétition, on a analysé les M=5 fichiers de données imputées résultants et le fichier de données complètes (avant la suppression) en ajustant le modèle Gamma pour Y_2 à l'aide du maximum de vraisemblance. L'estimation polyimputée a été construite comme la moyenne des cinq estimations des données imputées. Afin d'évaluer les différences des estimations ponctuelles, nous avons calculé les différences normalisées entre l'IMRS et des estimations de données complètes,

$$\Delta(\beta) = 100 \times \frac{\text{abs(estimation IMRS} - \text{estimation de données complètes)}}{\text{ET(estimation IMRS)}}.$$

Le tableau 4 indique la moyenne et l'écart type de $\Delta(\beta)$ pour trois coefficients de régression β_0, β_1 et β_2 dans le modèle. Les estimations IMRS se situent typiquement à 8 % près des unités standard complètes. Pour les coefficients de régression, on a calculé la couverture réelle et la grandeur moyenne des intervalles de confiance de 95 % de l'IMRS à l'aide de la distribution de référence t décrite dans Rubin (1987b). Pour chaque fichier de données simulées et chaque paramètre, on a déterminé si la vraie valeur (β_1 = 0,5 par exemple) se trouve à l'intérieur de l'intervalle correspondant. On a calculé la proportion d'intervalles contenant les vraies valeurs pour les 2 500 répétitions (voir le tableau 4). Pour ce qui est des fichiers de données complètes, la couverture réelle pour β_1, par exemple, était de 94,9 % et pour l'IMRS de 95,4. De plus, on a calculé la grandeur moyenne des intervalles de confiance. La largeur moyenne

de l'intervalle de confiance des données complètes pour β_1 était de 0,91 et pour l'IMRS la grandeur moyenne était de 1,22. Autrement dit, les données de l'IMRS ont donné des estimations d'intervalles bien calées.

On a utilisé la même étude de simulation afin de comparer les propriétés distributives des imputations de l'IMRS et d'une méthode entièrement bayésienne. Pour les hypothèses modélisées servant à préparer des données complètes, nous avons élaboré un algorithme de Monte Carlo à chaînes markoviennes afin de tirer des valeurs de la distribution prédictive a posteriori réelle des valeurs manquantes compte tenu des valeurs observées. Chaque étape du tirage faisait appel à l'algorithme de Metropolis-Hastings et exigeait appréciablement plus de temps d'ordinateur que la méthode IMRS. Par conséquent, seuls les 500 premiers fichiers de données simulées ont été utilisés pour cette comparaison. Nous avons calculé deux statistiques de Kolmogrove-Smimoff (KS) à partir de chaque fichier de données simulées : une pour comparer les imputations de la méthode IMRS et les valeurs cachées réelles, l'autre pour comparer les imputations bayésiennes et les valeurs cachées réelles. Il n'y avait aucune différence discernable entre ces deux statistiques pour les 500 fichiers de données simulées. Un nuage de ces 500 paires de statistiques KS a indiqué un étroit nuage de points autour d'une pente de 45 degrés.

Tableau 4

Moyennes et écarts types de la différence normalisée entre les estimations IMRS et les estimations de données complètes d'une part et la couverture réelle d'intervalles de confiance de 95 % nominaux d'autre part

Coefficient de régression	Écart type		Confiance	Couverture
	Moyenne	ET	IMRS	Données complètes
β_0	8,2	2,0	96,1	95,4
β_1	8,8	1,7	95,4	94,9
β_2	8,0	2,2	95,3	94,7

6. DISCUSSION

Nous avons décrit et évalué une procédure d'imputation multidimensionnelle par régression séquentielle pouvant servir à imputer les valeurs manquantes d'un choix de structures de données complexes comportant de nombreux types de variables, de restrictions et de limites. Cette procédure devrait être utile lorsqu'il est difficile de définir une distribution composée de toutes les variables ayant des valeurs manquantes. Un réel avantage de la stratégie est sa souplesse lorsqu'il s'agit de traiter chaque variable individuellement. Ainsi, afin de conserver toutes les corrélations entre deux variables, il faut inclure tous les termes à effet majeur à titre de variables explicatives, et pour conserver, par exemple, trois interactions factorielles, il faut inclure toutes les interactions à deux facteurs à titre de variables explicatives dans le modèle d'imputation. La mise en oeuvre de cette procédure exige uniquement un bon générateur de nombres aléatoires et des programmes d'ajustage pour un choix de programmes de régression multiple. Une application à base de SAS permettant d'exécuter cette stratégie peut être téléchargée d'un site Web (www.isr.umich.edu/ src/smp/ive).

Dans certains cas, il est possible de modifier l'algorithme de façon à en faire un échantillonnage de Gibbs à partir de la distribution prédictive composée des valeurs manquantes compte tenu des valeurs observées. Toutefois, la procédure IMRS sera plus utile lorsqu'il est difficile de formuler un modèle explicite. Tant pour les illustrations que pour la simulation, différents points de départ aléatoires ont servi à surveiller les valeurs imputées, aspect important dans de nombreuses applications concrètes. Il s'agit d'une bonne pratique lorsqu'un échantillonnage de Gibbs est utilisé dans le cadre d'un modèle bayésien explicite (Gelman et Rubin 1992), et elle devait être utilisée lorsqu'on a recours à la méthode de régression séquentielle décrite dans le présent exposé.

L'étude de simulation décrite à la section 5, bien que limitée, est favorable pour ce qui est des inférences fondées sur l'IMRS. Les imputations relevant des modèles IMRS et de Bayes étaient comparables. Il s'agissait ici, toutefois, d'élaborer une stratégie d'imputation qui soit peaufinée une variable à la fois et complètement en fonction de toutes les informations observées, plutôt qu'une distribution multidimensionnelles composée explicite de toutes les variables. De plus, on peut réduire la sensibilité du modèle en ayant recours à un modèle de régression semi-paramétrique pour chaque régression conditionnelle. L'interprétation bayésienne des modèles de lissage de type spline (Silverman 1985) peut servir à tirer des valeurs imputées de la distribution prédictive. De telles modifications méritent également une recherche plus poussée.

Pour certains grands fichiers de données comportant de nombreuses variables, l'IMRS peut exiger beaucoup de temps d'ordinateur. On peut modifier l'algorithme de façon à appliquer une méthode de sélection des variables pour chaque régression de chaque cycle. Nous avons comparé les inférences avec et sans la sélection de variables pour plusieurs grands fichiers de données, par exemple la National Health Interview Survey et la National Medical Expenditure Survey, à l'aide de plusieurs centaines de variables. Les inférences descriptives aussi bien que les inférences fondées sur des modèles de régression linéaire et logistique étaient très semblables, mais il subsiste un besoin de recherches plus détaillées.

Il est également possible d'utiliser la stratégie d'imputation décrite dans la présent exposé en même temps que, par exemple, la technique JRR (répétition répétée de type jackknife) pour l'estimation de la variance. Plus précisément, il s'agit 1) d'imputer de nouveau, individuellement, les valeurs manquantes de chaque IMRS répétée de type jackknife, 2) d'analyser le fichier de données répétées imputées et, enfin, 3) de combiner les estimations répétées

de façon à obtenir l'estimation ponctuelle et sa matrice de covariances. Cette stratégie exige plus de calculs que la stratégie d'imputation multiple. Cette stratégie d'imputation JRR intégrée et plusieurs de ses variantes sont actuellement à l'étude.

Enfin, il a été supposé que le fichier de données provient d'un plan de sondage aléatoire simple. Toutefois, la plupart des enquêtes ont recours à des plans de sondage complexes mettant en jeu la stratification, le groupement et la pondération. Il y a lieu de poursuivre les travaux afin de modifier la méthode de régression séquentielle de façon à incorporer des fonctions de plan complexe qui ne sont pas reflétées dans les variables X de l'expression(1). Toutefois, même si le processus d'imputation ne tient pas compte des fonctions de plan complexe, l'analyse des données complètes devrait se fonder sur le plan. Même si cela ne donne pas des inférences fondées sur le plan qui sont valides, la robustesse qui sous-tend l'analyse fondée sur le plan est conservée dans une certain mesure. La stratégie d'imputation JRR intégrée dont il est question ci-dessus pourrait offrir des propriétés fondées sur le plan plus attrayantes dans le cadre d'un plan complexe.

REMERCIEMENTS

Les auteurs aimeraient remercier les trois arbitres pour l'attention particulière qu'ils ont apportée à la lecture de cet article ainsi que pour leurs remarques pertinentes. Cette recherche a été partiellement soutenue par une subvention NSF DMS-0803720.

ANNEXE : MODÈLES DE RÉGRESSION ET IMPUTATIONS

Si l'on abandonne, par souci de brièveté, les indices inférieurs des variables, les étapes nécessaires à l'imputation de chaque type de variable sont les suivantes :

Variable continue : Pour Y (possiblement transformée à partir de l'échelle originale pour la normalité), une variable continue, construire un modèle de régression linéaire normale, $Y = U\beta + e$, où U est la matrice des variables explicatives mises à jour le plus récemment, e comporte une distribution normale multidimensionnelle avec une moyenne zéro et une variance $\sigma^2 I$, et I est une matrice d'identité. Supposer que $\theta = (\beta, \log\sigma)$ comporte une distribution a priori uniforme sur l'espace réel dimensionnel approprié. Ajuster ce modèle en fonction des unités pour lesquelles Y est observée.

Soit $B = (U'U)^{-1}U'Y$, le coefficient de régression estimé, SSE $= (Y - UB)'(Y - UB)$, la somme des carrés des résidus et df = lignes(Y) - cols(U), les degrés de liberté des résidus, et T la décomposition de Cholesky telle que $TT' = (U'U)^{-1}$. Il est facile de calculer les distributions a posteriori pertinentes (voir, par exemple, Gelman, Carlin, Stern et Rubin 1995, chapitre 7); les étapes ci-dessous fournissent ensuite des tirages de la distribution prédictive a posteriori des valeurs Y manquantes :

1. Produire un écart aléatoire chi carré u avec df degrés de liberté, et définir $\sigma_*^2 = \text{SSE}/u$.
2. Produire un vecteur $z = (z_1, z_2, ..., z_p)$ de dimension p = lignes (B) d'écarts aléatoires normaux, et définir $\beta_* = B + \sigma_* Tz$.
3. Soit U_{miss}, la matrice U pour celles qui ont des valeur Y manquantes. Les valeurs imputées sont $Y_* = U_{\text{miss}}\beta_* + \sigma_* v$, où v est un vecteur indépendant de lignes de dimension (U_{miss}) d'écarts aléatoires normaux.

Variable binaire : Lorsque Y est une variable binaire, on ajuste un modèle de régression logistique établissant le lien entre Y et U (mises à jour le plus récemment), $\text{logit}[\Pr(Y = 1 \mid U)] = U\beta$, à l'aide d'unités pour lesquelles Y est observée. Les valeurs imputées pour Y sont créées suivant les étapes ci-dessous :

2. Soit B, les estimations correspondant à un maximum de vraisemblance de β, et V sa matrice de covariances asymptotiques (inverse négatif de la matrice d'informations de Fisher observées). Soit T, la décomposition de Cholesky de V (c'est-à-dire, $TT' = V$). Produire un vecteur z d'écarts aléatoires normaux des lignes de dimension (B). Définir $\beta_* = B + Tz$.
2. Soit U_{miss}, la portion de U pour laquelle Y manque. Définir $P_* = [1 + \exp(-U_{\text{miss}}\beta_*)]^{-1}$. Produire un vecteur u, de lignes de dimension (U_{miss}) de nombres aléatoires uniformes entre 0 et 1. Imputer 1 si une composante particulière de u est inférieure ou égale à la composante correspondante de P_*, et imputer 0 autrement.

Cette stratégie ne donne que des tirages approximatifs de la distribution prédictive a posteriori des valeurs manquantes car les tirages du paramètre β proviennent de l'approximation asymptotique de sa distribution a posteriori réelle. Il est possible de puiser dans la distribution réelle en modifiant l'étape 1 à l'aide, par exemple, de l'échantillonnage-importance-rééchantillonnage (Rubin 1987b).

Variable mixte : Pour Y, une variable mixte (c'est-à-dire que Y a une valeur soit nulle, soit continue), modéliser les valeurs nulles à l'aide d'un indicateur 0-1 afin de distinguer entre 0 et des valeurs autres que 0, puis modéliser une variable à distribution normale pour le volet continu de la distribution à la condition que la variable indicatrice soit égale à 1. Autrement dit, utiliser une stratégie à deux degrés: imputer 1 ou 0 à l'aide de la stratégie logistique décrite ci-dessus, puis, en limitant l'échantillon aux unités ayant une valeur non nulle, utiliser la stratégie de la variable continue décrite ci-dessus pour imputer une valeur continue remplaçant la valeur 1 qui vient d'être imputée.

Variable de comptage : Pour Y, une variable de comptage, ajuster un modèle de régression de Poisson $Y \sim$ Poisson (λ), où $\log \lambda = U\beta$. Les imputations pour des valeurs manquantes en Y sont créées à l'aide des étapes ci-dessous :

1. Soit B, l'estimation correspondant à un maximum de vraisemblance de β, V sa matrice de covariances et T la décomposition de Cholesky de V. Produire un vecteur z d'écarts aléatoires normaux de lignes de dimension (B) et définir définit $\beta_* = B + Tz$.
2. Soit U_{miss}, la portion de U pour laquelle Y manque. Définir $\lambda_* = \exp(U_{\text{miss}}\beta_*)$. Produire des variables aléatoires de Poisson indépendantes avec des moyennes comme éléments de λ_*.

Variable polytomique : Pour Y pouvant avoir valeurs $j = 1, 2, ..., k$, noter $\pi_j = \Pr(Y = j \mid U)$. Ajuster un modèle de régression polytomique établissant le lien entre Y et U, où $\log = (\pi_j / \pi_k) = U\beta_j$ pour $j = 1, 2, ..., k-1$. Compte tenu de la restriction $\Sigma_j^k \pi_j = 1$, il s'ensuit que $\pi_k = (1 + \Sigma_j^{k-1} \exp(U\beta_j))^{-1}$.

Soit B, l'estimation correspondant à un maximum de vraisemblance des coefficients de régression $(\beta_1', \beta_2', ..., \beta_{k-1}')$, V la matrice des covariances asymptotique et T sa décomposition de Cholesky.

Les étapes ci-dessous permettent de créer des imputations :

1. Définir $\beta_* = B + Tz$, où z est un vecteur d'écarts aléatoires normaux de lignes de dimension (B).
2. Soit U_{miss}, la ligne de U comportant des Y manquants; soit $P_i^* = \exp\{U_{\text{misss}}\beta_{i\cdot}\} / \{1 + \Sigma_i \exp(U_{\text{miss}}\beta_{i\cdot})\}$, où $\beta_{i\cdot}$ représente les éléments appropriés de β_* où $i = 1, 2, ..., k-1$ et $P_k^* = 1 - \Sigma_i P_i^*$.
3. Soit $R_0 = 0, R_j = \Sigma_i^j P_i^*$ et $R_k = 1$, les sommes cumulatives des probabilités. Pour imputer des valeurs, produire un nombre aléatoire uniforme u et considérer j comme la catégorie imputée si $R_{j-1} \le u \le R_j$.

Encore une fois, l'imputation de variables mixtes, de comptage et catégoriques provient de distributions prédictives a posteriori approximatives puisque les paramètres correspondants sont tirés de leurs distributions a posteriori approximatives normales asymptotiques.

BIBLIOGRAPHIE

BARNARD, J. 1995. Cross-procedures for Multiple Imputation Inference: Bayesian Theory and Frequentist Evaluation. Thèse Doctorat, non publiée, University of Chicago, Department of Statistics.

GELFAND, A.E., et SMITH, A.M.F. (1990). Sampling based approaches to calculating marginal densities. *Journal of American Statistical Association*, 85, 398-409.

GELMAN, A., CARLIN, J., STERN, H. et RUBIN, D.B. (1995). *Bayesian Data Analysis*. London. Chapman and Hall.

GELMAN, A., et RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequences (avec discussion). *Statistical Science*, 7, 457-472.

GELMAN, A., et SPEED T.P. (1993). Characterizing a joint probability distribution by conditionals. *Journal of Royal Statistical Society*, B, 55, 185-188.

GEMAN, S., et GEMAN, D. (1984). Stochatic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

HEERINGA, S.G., LITTLE, R.J.A. et RAGHUNATHAN, T.E. (1997). Imputation of Multivariate Data on Household Net Worth. University of Michigan, Ann Arbor, Michigan.

LI, K.H., MENG, X.L., RAGHUNATHAN, T.E. et RUBIN, D.B. (1991). Significance levels from repeated p values from multiply-imputed data. *Statistical Sinica*, 1, 65-92.

LI, K.H., RAGHUNATHAN, T.E. et RUBIN, D.B. (1991). Large sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of American Statistical Association*, 86, 1065-1073.

LITTLE, R.J.A., et RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York : Wiley.

LITTLE, R.J.A., et SCHLUCHTER, M.D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, 72, 497-512.

LIU, C. (1995). Missing data imputation using the multivariate t distribution. *Journal of multivariate analysis*, 53, 139-158.

MADOW, W.G., NISSELSON, H., OLKIN, I. et RUBIN, D.B. (1983). *Incomplete Data in Sample Surveys*. 1,2, et 3, New York, Academic Press.

MENG, X.L., et RUBIN, D.B. (1992). Performing likelihood ratio tests with multiply imputed data sets. *Biometrika*, 79, 103-111.

OLKIN, I., et TATE, R.F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, 32, 448-465.

RAGHUNATHAN, T.E., et GRIZZLE, J.E. (1995). A split questionnaire survey design. *Journal of American Statistical Association*, 90, 54-63.

RAGHUNATHAN, T.E., et RUBIN, D.B. (1988). An application of Bayesian statistics using sampling/importance resampling to a deceptively simple problem in quality control. *Data Quality Control : Theory and Pragmatics*, (G.E. Liepins et V.R.R. Uppuluri, éds). New York: Marcel Dekker.

RAGHUNATHAN, T.E., et SISCOVICK, D.S. (1996). A multiple imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Applied Statistics*, 45, 335-352.

RAO, J.N.K., et SHAO, J. (1992). Jackknife variance estimation with survey data under hot-deck imputation. *Biometrika*, 79, 811-822.

RUBIN, D.B. (1976). Inference and missing data (avec discussion). *Biometrika*, 63, 581-592.

RUBIN, D.B. (1978). Multiple imputation in sample surveys – A phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 20-34.

RUBIN, D.B. (1987a). *Multiple Imputation for Nonresponse in Surveys*. New York : John Wiley and Sons.

RUBIN, D.B. (1987b). The SIR-algorithm – A discussion of Tanner and Wong's. The calculation of posterior distributions by data augmentation. *Journal of American Statistical Association*.

RUBIN, D.B. (1996). Multiple imputation after 18+ years. *Journal of American Statistical Association*, 91, 473-489.

RUBIN, D.B., et SCHAFER, J.L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. *Proceeding of the Statistical Computing Section of the American Statistical Association*, 83-88.

SCHAFER, J.L. (1997). *Analysis of Incomplete Multivariate Data by Simulation*. New York : Chapman and Hall.

SILVERMAN, B.W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of Royal Statistical Society*, B, 47, 1-52.

SISCOVICK, D.S., RAGHUNATHAN, T.E., KING, I., WEINMANN, S., WICKLUND, K.G., ALBRIGHT, J., BOVBJERG, V., ARBOGAST, P., KUSHI, L., COBB, L., COPASS, M.K., PSATY, B.M., RETZLAFF, B., CHILDS, M. et KNOPP, R.H. (1995). Dietary intake and cell-membrane levels of long-chain n-3 polyunsaturated fatty acids and the risk of primary cardiac arrest. *Journal of American Medical Association*, 274, 1363-1367.

Techniques d'enquête, juin 2001
Vol. 27, n° 1, pp. 105-116
Statistique Canada

Mieux comprendre la transformation des poids à l'aide d'une mesure de changement

JOHANE DUFOUR, FRANÇOIS GAGNON, YVES MORIN, MARTIN RENAUD et CARL-ERIK SÄRNDAL[1]

RÉSUMÉ

La littérature concernant les enquêtes longitudinales auprès des ménages propose plusieurs approches pour créer un ensemble de poids finaux à utiliser lors des analyses de données. La plupart de ces approches font appel à plusieurs procédures pour modifier les poids. Dans les faits, l'ensemble de poids initiaux est souvent transformé en un ensemble de poids intermédiaires afin de compenser pour la non-réponse, et par la suite en un ensemble de poids finaux via la poststratification pour redresser l'échantillon. La littérature dédie un grand intérêt à cette démarche mais aucune étude ne s'est vraiment penchée sur une approche pour mesurer l'importance relative de ces deux étapes ou pour mesurer l'efficacité des nombreuses alternatives qui existent pour créer les poids intermédiaires. L'objectif de cet article consiste à étudier et à mesurer le changement (du poids initial au poids final) qui est produit par la procédure de modification des poids. Une décomposition des poids finaux est proposée pour évaluer l'incidence relative de l'ajustement de non-réponse, la correction pour la poststratification et l'interaction entre ces deux ajustements. On utilise cette mesure de changement comme outil pour comparer l'efficacité de diverses méthodes d'ajustement pour la non-réponse, notamment les méthodes reposant sur la formation de groupes de réponse homogènes. La mesure de changement est étudiée par l'entremise d'une étude de simulation utilisant les données d'une enquête longitudinale de Statistique Canada, soit l'Enquête sur la dynamique du travail et du revenu. La mesure de changement est également appliquée aux données d'une deuxième enquête longitudinale, c'est-à-dire à l'Enquête nationale longitudinale sur les enfants et les jeunes.

MOTS CLÉS : Non-réponse; pondération; calage; enquête longitudinale; mesure de changement.

1. INTRODUCTION

Il est courant de retrouver dans la littérature une démarche de transformation de poids en deux étapes pour des enquêtes auprès des ménages. Comme première étape, un ajustement est appliqué aux *poids initiaux* afin de compenser pour la non-réponse; les poids résultants sont appelés les *poids intermédiaires*. La deuxième étape produit les *poids finaux* en faisant appel à la poststratification ou plus généralement à la technique de calage (voir Deville et Särndal 1992) afin de s'assurer que les poids finaux respectent certains totaux de contrôle de population connus. Toutes ces modifications de poids sont effectuées avec comme objectif la production du « meilleur ensemble possible de poids finaux ».

À Statistique Canada, les enquêtes longitudinales auprès des ménages suivent également cette démarche en deux étapes lors de la pondération et les travaux de recherche entrepris par l'Agence penchent dans cette direction. Le « Survey of Income and Program Participation (SIPP) » du U.S. Bureau of the Census (voir Rizzo, Kalton et Brick 1996) suit également une telle approche.

Dans la littérature, plusieurs méthodes d'ajustement des poids sont proposées pour pallier la non-réponse. Rizzo et coll. (1996) ont comparé les estimations obtenues de plusieurs de ces méthodes à des estimations de sources indépendantes. Cependant, peu d'auteurs ont effectué des études de simulation ou proposé des outils pour comparer l'efficacité relative des méthodes quant à leur capacité à réduire le biais de non-réponse.

L'objectif principal de ce document consiste à étudier et à mesurer le changement (entre les poids initiaux et les poids finaux) qui est produit suite à l'adoption d'une procédure de modification de poids en deux étapes. Pour ce faire, une mesure de changement à quatre composantes est proposée pour quantifier l'incidence relative de l'ajustement de non-réponse, la correction pour la poststratification et l'interaction entre ces deux ajustements. Le second objectif est d'utiliser la mesure de changement afin de comparer l'efficacité de diverses méthodes d'ajustement de non-réponse à l'aide d'une étude de simulation basée sur les données de l'Enquête longitudinale sur la dynamique du travail et du revenu (EDTR) et avec les données de l'Enquête nationale longitudinale sur les enfants et les jeunes (ENLEJ). Les enquêtes longitudinales se distinguent puisque beaucoup d'information à propos des répondants et des non-répondants de la vague courante est disponible à partir des répondants aux vagues précédentes. Ceci permet l'utilisation de méthodes plus complexes pour l'ajustement de non-réponse.

Un cadre général pour la pondération des enquêtes longitudinales auprès des ménages est tout d'abord présenté à la section 2. Suit ensuite, à la section 3, une présentation de la mesure de changement qui sera utilisée pour quantifier les étapes de transformation entre le poids initial et le poids final. La section 4 traite de stratégies d'ajustement pour la

[1] Johane Dufour, François Gagnon, Yves Morin, Martin Renaud et Carl-Erik Särndal, Statistique Canada, Parc Tunney, Ottawa, (Ontario) K1A 0T6.

non-réponse rencontrées dans la littérature. Viennent, aux sections 5 et 6, les résultats des études basées sur l'EDTR et l'ENLEJ. La dernière section présente les conclusions de cette étude.

2. CADRE GÉNÉRAL POUR LA PONDÉRATION LONGITUDINALE

Dans une enquête longitudinale auprès des ménages, les personnes composant l'échantillon initial sont suivies à travers le temps et sont communément appelées *personnes longitudinales*. C'est cet ensemble de personnes qui sera utilisé dans les études présentées dans ce document. On réfèrera à celles-ci en utilisant le terme « *unité de référence* ». La présente section donne un aperçu des étapes suivies afin de modifier le poids initial des personnes longitudinales en un poids final.

2.1 Poids initiaux

Soit $U = \{1, ..., k, ..., N\}$ une population finie. On s'intéresse à une variable y dont la valeur pour la k-ième unité est notée y_k. L'objectif est d'estimer le total $Y = \Sigma_U y_k$. Soit w_{0k}, le poids initial pour toute unité $k \in s$, où s désigne l'échantillon longitudinal. En l'absence de non-réponse, le système de poids initiaux $\{w_{0k}: k \in s\}$ donnerait un estimateur $\hat{Y} = \Sigma_s w_{0k} y_k$ pour Y. On suppose ici que les w_{0k} sont normés de façon à ce que $\Sigma_s w_{0k} = N$. Quoique sans biais pour Y, $\hat{Y}$ a l'inconvénient de ne pas incorporer d'information auxiliaire sous la forme de totaux de contrôle connus pour des poststrates.

2.2 Ajustement de non-réponse et poids intermédiaires

La plupart des enquêtes font face à de la non-réponse. Deux approches sont souvent utilisées afin de compenser pour la non-réponse : l'imputation et la correction des poids initiaux des répondants par un facteur d'ajustement. Cette dernière est la plus couramment utilisée dans les enquêtes auprès des ménages pour pallier la non-réponse totale, tandis que l'imputation est souvent privilégiée dans le traitement de la non-réponse partielle. La non-réponse totale réduit la taille d'échantillon étant donné que la valeur y_k n'est disponible que pour $k \in r$, où $r \subset s$ est l'ensemble des m unités répondantes. Pour cet ensemble réduit de données, les poids initiaux w_{0k} sont en moyenne trop petits et on a $\Sigma_r w_{0k} < N$. L'estimateur $\hat{Y}' = \Sigma_r w_{0k} y_k$ n'est pas admissible car il sous-estime systématiquement Y.

L'ajustement des poids est souvent choisi afin de compenser pour la non-réponse totale dans les enquêtes auprès des ménages. Une méthode courante pour ajuster les poids consiste à construire des groupes de réponse homogènes (GRH). Ces derniers sont formés de sorte que chacun d'entre eux soit constitué d'unités de référence ayant une probabilité de réponse semblable. Ensuite, à l'intérieur de chaque GRH, un facteur d'ajustement, égal à l'inverse du taux de réponse (pondéré ou non-pondéré) du GRH, est calculé. Pour chaque unité répondante k, l'ajustement de non-réponse consiste à multiplier w_{0k} par le facteur d'ajustement du GRH. Cette opération produit un système de poids intermédiaires $\{w_{1k}: k \in r\}$ tel que $\Sigma_r w_{1k} = N$. Avec ces poids, on peut construire l'estimateur $\hat{Y}'' = \Sigma_r w_{1k} y_k$, lequel élimine la sous-estimation qui caractérise $\hat{Y}' = \Sigma_r w_{0k} y_k$. Tout comme pour le système de poids initiaux, le principal inconvénient de ce système de poids intermédiaires est de ne pas incorporer d'information auxiliaire disponible pour des poststrates.

2.3 Poststratification et poids finaux

Une pratique courante dans les enquêtes auprès des ménages consiste à modifier les poids intermédiaires à l'aide d'une poststratification, ou de façon plus générale à l'aide d'un calage aux marges, de sorte que la somme des poids finaux sur l'ensemble des répondants corresponde aux comptes de population connus. La poststratification produit donc un système de poids finaux $\{w_{2k}: k \in r\}$, qui incorpore l'information auxiliaire et qui est à la fois cohérent avec les totaux de contrôle des poststrates. Dans ce cas, les poids finaux dans chaque poststrate p vérifient $\Sigma_{r_p} w_{2k} = N_p$ où N_p est l'effectif connu et r_p est l'ensemble des unités répondantes de la p-ième poststrate. Il s'ensuit que $\Sigma_r w_{2k} = N$. Des variables démographiques ou géographiques sont fréquemment utilisées pour définir des poststrates. Le choix des poststrates, qui doivent être suffisamment grandes, est limité par la disponibilité des totaux de contrôle. Plusieurs méthodes peuvent être utilisées afin de caler les poids intermédiaires aux totaux de contrôle choisis.

3. MESURE DE CHANGEMENT DES POIDS INITIAUX AUX POIDS FINAUX

Dans cette section, une mesure du changement entre les poids initiaux et les poids finaux est proposée afin de mieux comprendre l'effet de la procédure de modification de poids. La décomposition de cette mesure en quatre composantes permet de quantifier l'effet de chacune des étapes de pondération décrites à la section 2. Ces composantes seront utilisées aux sections 5 et 6 dans la comparaison de diverses méthodes d'ajustement des poids pour pallier la non-réponse.

Si les poids initiaux sont normés tels que $\Sigma_s w_{0k} = N$, et si $r \subset s$, alors les trois systèmes de poids décrits à la section 2 vérifient les relations suivantes :

$$\sum_r w_{0k} < N, \sum_r w_{1k} = N, \sum_r w_{2k} = N.$$

Soit

$$\bar{w}_{01} = \frac{\sum_r w_{1k}}{\sum_r w_{0k}} \text{ et } \bar{w}_{02} = \frac{\sum_r w_{2k}}{\sum_r w_{0k}}.$$

Le ratio $\bar{w}_{01}$ mesure le changement moyen du système de poids intermédiaires par rapport au système de poids initiaux. Plus la non-réponse totale est prononcée, plus $\bar{w}_{01}$ s'éloigne de la valeur 1 qui est obtenue seulement en l'absence de non-réponse. Le ratio $\bar{w}_{02}$ représente le changement moyen du système de poids finaux par rapport au système de poids initiaux.

Les ratios $\bar{w}_{01}$ et $\bar{w}_{02}$ mesurent un changement de poids moyen. Pour mesurer un changement de poids individuel, on définit, pour tout $k \in r$, $r_{01k} = w_{1k}/(w_{0k}\bar{w}_{01})$, et $r_{02k} = w_{2k}/(w_{0k}\bar{w}_{02})$. Ces quantités varient autour de 1. Plus précisément, leurs moyennes pondérées sont égales à 1 :

$$\frac{\sum_r w_{0k} r_{01k}}{\sum_r w_{0k}} = \frac{\sum_r w_{0k} r_{02k}}{\sum_r w_{0k}} = 1.$$

Les quantités r_{01k} et r_{02k} seront utiles pour mesurer les changements de poids individuels.

Le changement total que subissent les poids, du système de poids initiaux au système de poids finaux en passant par le système de poids intermédiaires, peut être calculé par une mesure de changement, aussi appelée *distance*. Soit D la mesure de changement suivante :

$$D = \frac{\sum_r w_{0k} \left(\frac{w_{2k}}{w_{0k}} - 1 \right)^2}{\sum_r w_{0k}}.$$

En fait, D est une moyenne pondérée des facteurs individuels de changement de poids suivants :

$$\left(\frac{w_{2k}}{w_{0k}} - 1 \right)^2 = \left(\frac{w_{2k}}{w_{1k}} \frac{w_{1k}}{w_{0k}} - 1 \right)^2.$$

La mesure de changement D se décompose en quatre composantes selon l'équation suivante :

$$D = R_{01} + R_{12} + R_{\text{int}} + G$$

où :

$$R_{01} = \bar{w}_{02}^2 \frac{\sum_r w_{0k}(r_{01k} - 1)^2}{\sum_r w_{0k}},$$

$$R_{12} = \bar{w}_{02}^2 \frac{\sum_r w_{0k}(r_{02k} - r_{01k})^2}{\sum_r w_{0k}},$$

$$R_{\text{int}} = 2\bar{w}_{02}^2 \frac{\sum_r w_{0k}(r_{01k} - 1)(r_{02k} - r_{01k})}{\sum_r w_{0k}} \quad \text{et}$$

$$G = (\bar{w}_{02} - 1)^2.$$

Il est à noter que la mesure de changement D prend uniquement des valeurs positives, avec l'égalité à zéro quand les deux conditions suivantes sont satisfaites :

(i) absence de non-réponse ($r = s$ et $w_{1k} = w_{0k}$ pour tout k),

(ii) absence d'effet de la poststratification sur les poids intermédiaires ($w_{2k} = w_{1k}$ pour tout k).

Un taux de non-réponse élevé aura tendance à augmenter la valeur de la mesure de changement D étant donné que, dans un tel cas, w_{1k} est en moyenne considérablement plus grand que w_{0k}.

Le terme R_{01} mesure les changements individuels que subissent les poids en passant du système initial au système intermédiaire. On verra un peu plus loin que la composante R_{01} est en quelque sorte associée à la qualité du modèle pour la non-réponse et qu'une grande valeur de R_{01} est préférable. Le terme R_{12} mesure les changements individuels que subissent les poids en passant du système intermédiaire au système final. Le terme R_{int} mesure l'interaction entre les deux types de changements et le terme G mesure le changement de poids moyen entre le système initial et le système final.

En plus de son interprétation comme une distance, la mesure de changement D peut aussi être interprétée comme un écart quadratique moyen des changements w_{2k}/w_{0k} par rapport à 1, et ceci relativement à la distribution définie par les w_{0k}. Dans cette optique, la composante G correspond au biais au carré (soit le carré de l'écart entre la moyenne $\bar{w}_{02}$ des w_{2k}/w_{0k} et 1) tandis que la somme des trois autres composantes correspond à la variance. Dans le plus simple des cas, où un ajustement de non-réponse est calculé en utilisant un seul GRH et où on n'applique aucune poststratification, on a $w_{0k} = N/n$ pour tout $k \in s$ (dans le cas d'un tirage aléatoire simple de taille n) et $w_{1k} = w_{2k} = N/m$ pour tout $k \in r$, (où le facteur d'ajustement de non-réponse est n/m, soit l'inverse du taux de réponse). On a alors $D = G = \{(n/m) - 1)\}^2$ et $R_{01} = R_{12} = R_{\text{int}} = 0$.

D'importantes conclusions peuvent être tirées en regardant l'importance relative des termes R_{01}, R_{12} et R_{int}. Si R_{01} est élevé, et qu'en même temps R_{12} est peu élevé, l'enquête en est une où l'ajustement de non-réponse cause d'importants changements individuels dans les poids, tandis que la poststratification ne modifie les poids individuels que très peu. Par contre, lorsque R_{12} est élevé, la poststratification engendre des changements individuels considérables. Les résultats présentés aux sections 5 et 6 démontreront que le terme R_{01} peut être utilisé dans la comparaison de l'efficacité de diverses méthodes d'ajustement de non- réponse. De plus, le signe de R_{int} indique si les deux types de changements individuels opèrent dans la même direction ($R_{\text{int}} > 0$) ou dans des directions opposées ($R_{\text{int}} < 0$). En pratique, on s'attend à ce que R_{int} soit peu important ou même négligeable.

4. STRATÉGIES D'AJUSTEMENT DE LA NON-RÉPONSE

Dans la littérature, il existe plusieurs méthodes d'ajustement de poids (dont la méthode décrite à la section 2.2) afin de compenser pour la non-réponse. Une autre méthode fréquemment utilisée dans les enquêtes longitudinales consiste à ajuster les poids selon l'inverse de la probabilité prédite de réponse obtenue à l'aide de la régression logistique. On retrouve également les méthodes d'ajustement basées sur le calage aux marges qui utilisent les distributions marginales de l'échantillon initial ou de la population. Singh, Wu et Boyer (1995) ont utilisé cette approche afin de dériver une méthode d'ajustement ayant la propriété de produire des estimations cohérentes d'une vague à l'autre pour les enquêtes longitudinales. Deville (1998) propose une méthode de correction de non-réponse par calage ou par échantillonnage équilibré. Pour une revue des méthodes d'ajustement de non-réponse, se référer à Kalton et Kasprzyk (1986), Platek, Singh et Tremblay (1978), Chapman, Bailey et Kasprzyk (1986) ainsi qu'à Little (1986). Dans le présent document, seules les méthodes reposant sur la création de GRH sont considérées.

4.1 Formation de GRH

Dans la plupart des enquêtes, l'information connue pour les non-répondants est minime, mis à part quelques variables de stratification de la base de sondage. Ainsi donc, le choix des GRH est très limité et les strates sont souvent utilisées comme GRH. Dans un tel cas, l'hypothèse que la probabilité de réponse est la même pour toutes les unités d'une strate donnée est émise. Toutefois, dans les enquêtes longitudinales, beaucoup d'information à propos des répondants et des non-répondants de la vague courante est disponible à partir des réponses aux vagues précédentes. Cette quantité d'information peut donc être utilisée pour créer des GRH pour lesquels l'hypothèse d'un mécanisme de réponse uniforme à l'intérieur de ceux-ci est plausible. Ceci résulte en un meilleur ajustement pour la non-réponse et conséquemment en une réduction du risque d'introduction d'un biais de non-réponse dans les estimations.

4.1.1 Méthode de sélection des variables pour la formation des GRH

Par définition, les GRH sont formés à partir d'un ensemble de variables pouvant prédire la propension à répondre. Si l'ensemble de variables initialement identifiées est trop vaste, des tests univariés peuvent être utilisés afin de garder les variables les plus importantes pour discriminer les caractéristiques des répondants de celles des non- répondants. À partir de cet ensemble de variables importantes, une méthode de sélection peut être utilisée pour garder les meilleures variables explicatives de réponse. Deux méthodes courantes de sélection de variables sont : un modèle de régression logistique (RL) et un modèle de segmentation (MS).

4.1.1.1 Régression logistique

Dans la méthode de la RL, l'utilisation combinée du « fait d'avoir répondu ou non à l'enquête » comme variable dépendante, des poids standardisés et de la procédure « stepwise », permet d'obtenir la liste des variables dichotomiques les plus significatives pour expliquer la propension à répondre. Règle générale, les GRH sont créés selon les 2^q combinaisons possibles à partir d'un ensemble de q variables explicatives retenues. On qualifie souvent la méthode de la RL d'approche symétrique. Cependant, le respect de certaines contraintes additionnelles lors de la création des GRH, peut entraîner une diminution de leur nombre. Par exemple, on peut exiger d'avoir un nombre minimum (n) d'unités de référence et un taux de réponse (TR) (pondéré ou non pondéré) supérieur à un certain seuil dans chacun des GRH. Kalton et Kasprzyk (1986) favorisent ces contraintes afin d'éviter l'augmentation de la variance associée aux poids extrêmes. Cependant, ces contraintes peuvent réduire l'efficacité de l'ajustement de non-réponse et se traduire par une augmentation du biais. Lorsqu'un GRH ne satisfait pas à une de ces contraintes, on doit le regrouper avec un autre GRH. Le regroupement des GRH s'effectue jusqu'à ce que tous les GRH satisfassent aux contraintes additionnelles imposées. De cette façon, on obtient $2^q - J$ combinaisons valides, où J représente la réduction causée par le regroupement des GRH.

Par exemple, à la figure 1, $2^q = 8$ GRH sont créés à partir de q=3 variables explicatives. Les cases ombrées de la figure 1 représentent les GRH. À l'intérieur de chaque GRH, un facteur d'ajustement est calculé et le poids w_{0k} de chaque unité de référence est ensuite ajusté correctement.

4.1.1.2 Modèle de segmentation

La méthode MS, dite non symétrique, est basée sur l'algorithme CHAID (Chi-square Automatic Interaction Detection) développé par Kass (1980). Elle partitionne l'échantillon en sous-groupes selon les taux de réponse des variables explicatives en se servant de tests du Khi-deux. La procédure de segmentation continue jusqu'à ce qu'une variable explicative significative ne puisse plus être trouvée. Les sous-groupes finaux créés par le MS deviennent les GRH, pour lesquels des ajustements de non-réponse sont calculés. Comme pour la RL, le respect de contraintes additionnelles peut être imposé.

À la figure 1, on peut observer que la méthode du MS a partitionné l'échantillon en plusieurs GRH à partir de différentes variables explicatives. Les GRH sont à nouveau représentés par les cases ombrées. La segmentation s'est poursuivie jusqu'à ce qu'il ne soit plus possible de trouver de variables explicatives.

4.1.2 Facteur d'ajustement pour la non-réponse

Que les GRH soient formés en faisant appel à la RL ou à l'aide du MS, on émet l'hypothèse d'un mécanisme de

réponse uniforme à l'intérieur de chacun des GRH. Ainsi, le facteur d'ajustement pour la non-réponse est donné par l'inverse du taux de réponse (pondéré par w_{0k} ou non-pondéré) du GRH.

5. ÉTUDE EMPIRIQUE BASÉE SUR L'ENQUÊTE SUR LA DYNAMIQUE DU TRAVAIL ET DU REVENU (EDTR)

Afin de comparer l'efficacité de la RL et du MS, les données de l'EDTR ont été utilisées pour une étude empirique. L'EDTR est une enquête longitudinale auprès des ménages qui a débuté en 1993, et dont l'un des objectifs consiste à comprendre le bien-être économique de la société canadienne (voir Lavigne et Michaud 1998).

Ces deux méthodes ont été mises à l'essai à l'aide d'une simulation en analysant des variables d'intérêt et différents domaines. Les composantes de la mesure de changement, les biais absolus et relatifs de même que les variances, ont été étudiés.

5.1 Description de l'étude empirique

Comme première étape de l'étude empirique, la probabilité de répondre à la première vague de l'enquête a été estimée pour chacune des unités de l'échantillon longitudinal. Des variables potentiellement explicatives de la propension à répondre (issues d'une interview préliminaire) ont été utilisées pour former un très grand nombre de GRH. Tous les individus de l'échantillon ont été assignés à un GRH à partir des valeurs des variables explicatives. Une probabilité de réponse a alors été estimée dans chacun des GRH en utilisant le taux de réponse pondéré. Par la suite, seuls les répondants, ainsi que leur probabilité de réponse estimée, ont été conservés dans l'échantillon de référence pour la simulation. De la non-réponse a ensuite été générée pour l'échantillon de référence à l'aide d'un échantillonnage de Poisson. Cette procédure, illustrée à la figure 2, a été répétée 100 fois de façon indépendante, créant ainsi 100 ensembles de répondants et de non-répondants. Le taux de réponse moyen de chaque répétition s'établissait à environ 90 %, soit le taux observé à la première vague de l'EDTR.

Pour chacune des 100 répétitions, un ajustement de non-réponse a été effectué en utilisant la méthode de la RL pour créer les GRH. Similairement, un ajustement de non-réponse a été effectué en utilisant la méthode du MS pour créer les GRH pour chacune des 20 premières répétitions. Pour l'approche du MS, le nombre de répétitions a été limité à 20 étant donné la stabilité des résultats et puisque plusieurs interventions manuelles et l'utilisation d'un logiciel spécifique (dans notre cas : Knowledge Seeker – ANGOSS Software 1995) étaient nécessaires.

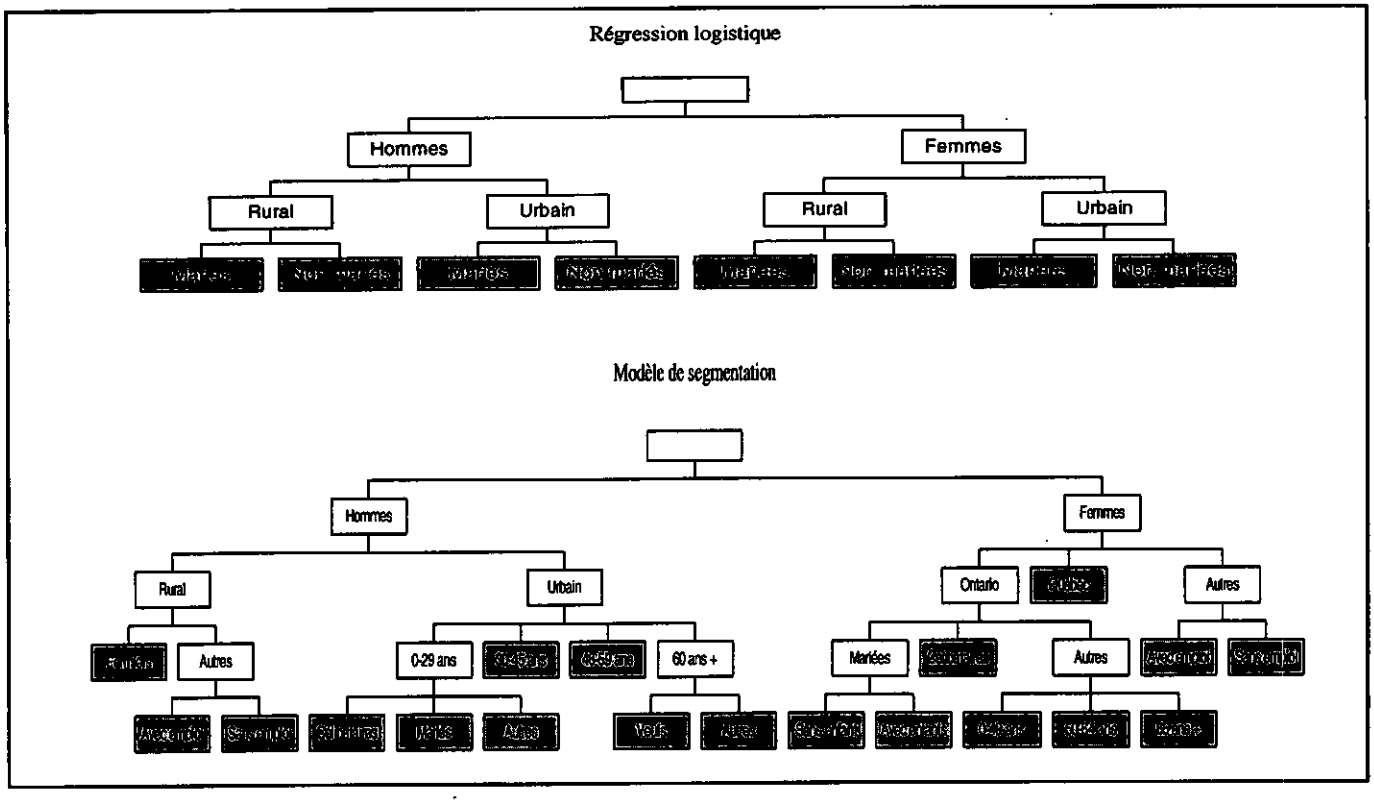


Figure 1. Visualisation de la formation des GRH selon la méthode

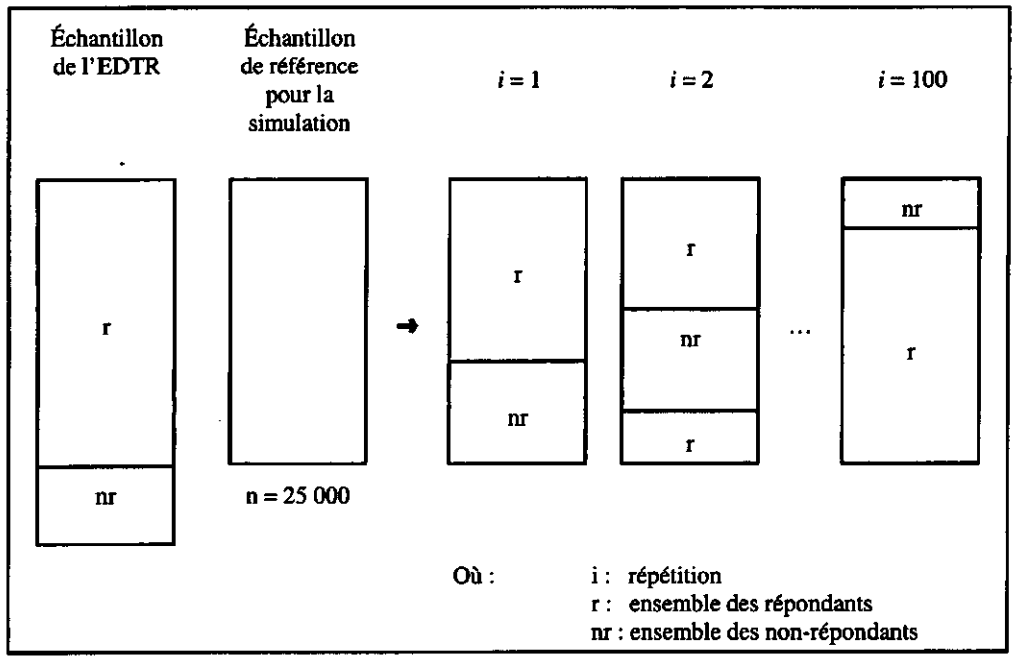


Figure 2. Illustration du procédé de simulation

Plusieurs variantes des méthodes de sélection des variables ont été étudiées :

a) RL_i, où i représente la moyenne approximative, sur les 100 répétitions, du nombre de GRH générés selon la méthode de la RL. Dans la présente étude, i=4, 16, 40, 60. Par exemple, pour RL_40, les q=6 variables explicatives les plus importantes de la propension à répondre ont d'abord été identifiées. Les GRH ont ensuite été formés en utilisant les combinaisons valides $(2^q - J)$ de ces q=6 variables explicatives. L'imposition de contraintes additionnelles ($n > 30$ et TR>50 %) dans chaque GRH a mené au regroupement de certains GRH. En moyenne, sur les 100 répétitions, 24 GRH ont dû être regroupés (J=24) et un total de $2^q - J = 2^6 - 24 = 40$ GRH ont été formés, d'où l'appellation RL_40. Dans l'étude de simulation, RL_i, où i=4, 16, 40, 60 GRH, correspond respectivement à q =2, 4, 6, 8 variables explicatives.

b) MS_ i, où i indique la moyenne approximative sur les 20 premières répétitions du nombre de GRH générés selon la méthode du MS. Dans la présente étude, i=16, 25, 40. Par exemple, pour MS_16, un MS a été utilisé avec un seuil de signification de 0,0001. Après l'imposition des mêmes contraintes additionnelles que pour la RL, 16 GRH ont en moyenne été créés. MS_i, où i=16, 25, 40 GHR, correspond respectivement aux seuils de signification 0,0001; 0,0005; 0,0025. Plus le seuil utilisé est élevé, plus il est facile d'identifier des différences significatives, ce qui permet une segmentation plus détaillée et en conséquence un plus grand nombre de GRH.

c) Une méthode avec un seul GRH (1_GRH) a également été utilisée à des fins de comparaison. Cette méthode consiste, pour chacune des 100 répétitions, à définir l'échantillon tout entier comme un seul GRH. Il est à noter que cette méthode est efficace seulement si le mécanisme de réponse est uniforme au sein de l'échantillon complet, ce qui est rarement le cas.

Dans un premier temps, les poids initiaux ont été normés tels que $\Sigma_s w_{0k} = N$ afin d'éliminer l'effet de la sous-couverture et de mieux isoler l'effet de la non-réponse. Ainsi, G mesurera uniquement le changement moyen causé par l'ajustement de non-réponse.

Une fois les poids initiaux normés, chacun des ensembles de poids finaux est ensuite le résultat de deux étapes : un ajustement pour la non-réponse (selon une des huit méthodes mentionnées : 1_GRH, RL_i, où i = 4, 16, 40, 60 et MS_i, où i = 16, 25, 40) et une même poststratification (14 groupes d'âge-sexe par province).

5.2 Analyse des résultats de l'étude empirique

Pour chacune des méthodes discutées à la section précédente, les composantes de la mesure de changement D ont été étudiées. De même, le biais de non-réponse moyen, absolu et relatif ainsi que la variance moyenne des estimations ont été analysés.

5.2.1 Mesure de changement (D)

Le tableau 1 présente la valeur moyenne de D et de ses composantes sur toutes les M répétitions (où M=100 pour la RL et M=20 pour le MS) ainsi que la contribution en pourcentage de chaque composante à la valeur moyenne de D. On remarque tout d'abord que pour la méthode 1_GRH, R_{01} est nul puisqu'un seul et même ajustement de non-réponse est apporté à l'ensemble des répondants. Ainsi, $w_{1k} = \alpha w_{0k}$,

où α est une constante, et par conséquent, $r_{01k} = 1$ pour tout $k \in r$ et $R_{01} = 0$. On remarque également que D augmente à mesure que le nombre de GRH augmente peu importe la méthode, RL ou MS. Ainsi, plus il y a de GRH pour pallier la non-réponse, plus le changement total que subissent les poids est grand. De plus, les valeurs de D sont plus élevées pour le MS que pour la RL.

Pour la RL et le MS, la contribution de R_{01} à la mesure de changement est de plus en plus grande à mesure que le nombre de GRH augmente étant donné qu'un plus grand nombre de GRH cible mieux la non-réponse. Par conséquent, l'ajustement de non-réponse prend de plus en plus d'importance et par le fait même, les poids varient de plus en plus. De plus, la contribution de R_{01} à la mesure de changement est beaucoup plus importante pour le MS que pour la RL. Ceci indique que le MS semble mieux modéliser la non-réponse et isoler mieux les tendances particulières que la RL.

Quant à R_{12}, il est presque constant, peu importe la méthode et le nombre de GRH utilisés. Cependant, malgré le fait qu'il varie très peu, sa contribution à la mesure de changement diminue à mesure que le nombre de GRH augmente. Ceci s'explique par le fait qu'une plus grande variation des poids survient lors de l'ajustement de non-réponse et que les modifications que la poststratification apporte aux poids sont de moins en moins importantes à mesure que le nombre de GRH augmente.

Dans le cas de R_{int} sa valeur est négligeable et sa contribution à la mesure de changement est très faible. Ceci signifie que l'interaction entre l'ajustement de non-réponse et la poststratification est quasiment nulle.

Enfin, G demeure constant, peu importe la méthode et le nombre de GRH utilisés. Comme pour R_{12}, la contribution de G à la mesure de changement diminue lorsque le nombre de GRH augmente. En effet, un plus grand nombre de GRH cible mieux la non-réponse, provoquant ainsi une plus grande variation dans le système de poids intermédiaires.

Puisque pour l'ensemble des méthodes, G est constant, R_{int} est près de zéro et R_{12} est presque constant, il est clair que les variations de D sont influencées en majeure partie par les variations de R_{01}.

Le graphique 1 montre la contribution moyenne en pourcentage de R_{01} et R_{12} à la mesure de changement. Pour la RL et le MS, la contribution de R_{01} augmente avec le nombre de GRH tandis que celle de R_{12} diminue. Aussi, la contribution de R_{01} est supérieure pour le MS que pour la RL, tandis que celle de R_{12} est inférieure pour le MS que pour la RL. En fait, le profil de la contribution de R_{01} est le même que le profil de D (tableau 1). Ceci confirme que les variations de la mesure de changement sont principalement causées par les variations de R_{01}.

Le graphique 2 permet de comparer la RL et le MS en ce qui a trait à la contribution moyenne en pourcentage de R_{01} à D. Pour un nombre de GRH donné, R_{01} contribue à un plus grand pourcentage de D pour la méthode du MS que pour la RL. Ceci signifie que les changements individuels dans les poids entre le système initial et le système intermédiaire sont supérieurs pour le MS que pour la RL.

5.2.2 Biais relatifs et absolus

Afin de comparer la performance de la RL face au MS dans la réduction du biais de non-réponse, le biais relatif (BR) et le biais absolu (BA) ont été utilisés :

$$\mathrm{BR}_i = 100\left(\frac{\hat{Y}_i - Y}{Y}\right) \text{ et } \mathrm{BA}_i = \hat{Y}_i - Y;$$

où $\hat{Y}_i$ est l'estimation de la variable d'intérêt obtenue pour la i-ème répétition, $i = 1, 2, \ldots, M$, M=100 pour la RL et M=20 pour le MS et Y est le total de la variable d'intérêt obtenu de l'échantillon de référence.

Le biais relatif moyen (BRM) et le biais absolu moyen (BAM) sont calculés en prenant respectivement la moyenne du BR et du BA pour l'ensemble des répétitions :

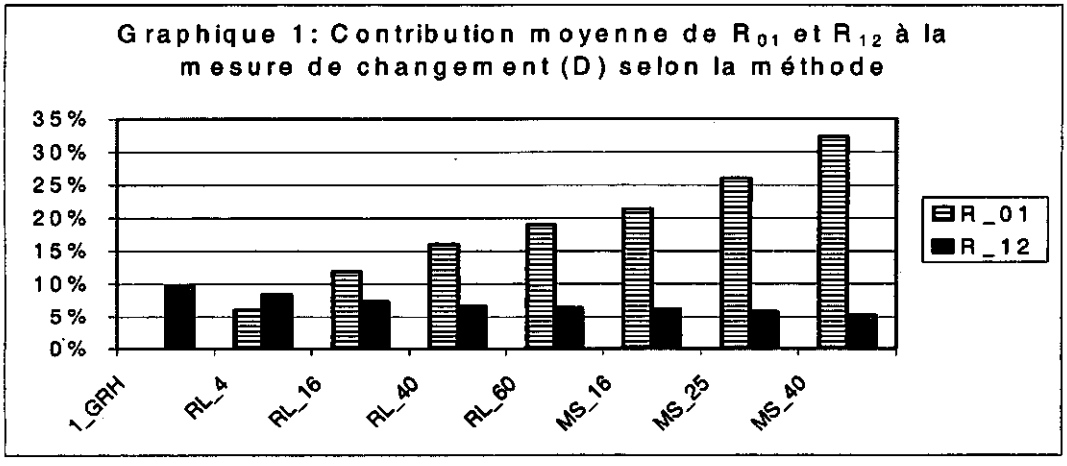
$$\mathrm{BRM} = \frac{1}{M}\sum_{i=1}^{M} \mathrm{BR}_i \text{ et } \mathrm{BAM} = \frac{1}{M}\sum_{i=1}^{M} \mathrm{BA}_i$$

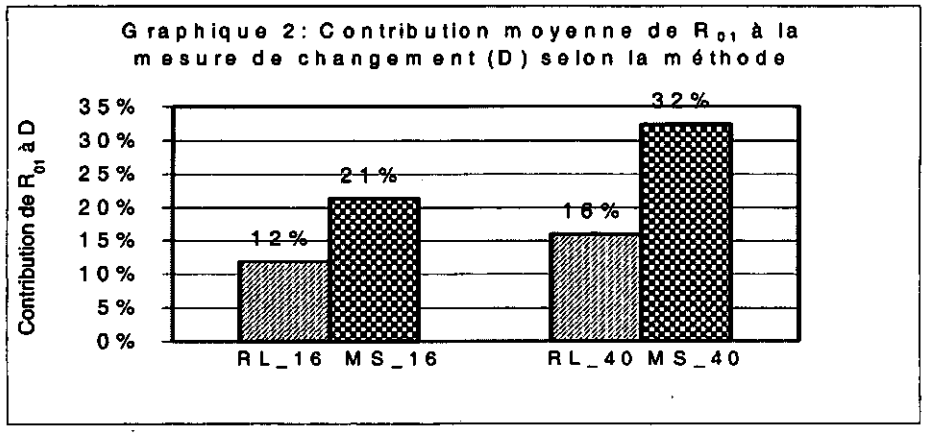
où M=100 dans le cas de la RL, et M=20 dans le cas du MS.

Tableau 1

Valeur moyenne de D sur les répétitions, de chaque composante et de leur contribution (en %) à la mesure de changement pour chacune des huit méthodes d'ajustement de la non-réponse

Méthode	D	R_{01} ($\times 10^{-3}$)	R_{01}/D (%)	R_{12} ($\times 10^{-3}$)	R_{12}/D (%)	R_{int} ($\times 10^{-5}$)	R_{int}/D (%)	G ($\times 10^{-2}$)	G/D (%)
1_GRH	0,012135	0,00	0,00	1,17	9,66	0,00	0,00	1,11	90,34
RL_4	0,012952	0,78	6,04	1,10	8,49	0,06	0,01	1,11	85,46
RL_16	0,013809	1,66	11,97	1,00	7,31	3,76	0,54	1,11	80,19
RL_40	0,014426	2,32	16,02	0,96	6,66	4,02	0,55	1,11	76,77
RL_60	0,014948	2,85	19,00	0,95	6,35	3,75	0,49	1,11	74,15
MS_16	0,015712	3,42	21,33	0,97	6,19	3,40	0,43	1,11	72,05
MS_25	0,016713	4,44	26,02	0,95	5,73	2,95	0,36	1,11	67,89
MS_40	0,018202	5,97	32,37	0,95	5,23	1,20	0,14	1,11	62,26





Pour les 100 répétitions, des estimations à l'échelle nationale ont été produites pour les trois variables suivantes : « personne vivant ou non dans une famille dont le revenu est inférieur au seuil de faible revenu (SFR) », « revenu total de la personne (RT) » et « salaires et traitements de la personne (ST) ». Le BRM de chaque estimation a été calculé pour les huit méthodes à l'étude. Étant donné la grande taille de l'échantillon, le faible taux de non-réponse (10 %) et le fait qu'un grand nombre de totaux de contrôle ont été utilisés pour la poststratification, le BRM est très faible (voir tableau 2) pour chacune des méthodes utilisées.

On observe au tableau 2 que pour les trois variables, le BRM est à peu près constant pour le MS peu importe le nombre de GRH utilisés. De même, pour la RL, le BRM du RT et des ST est à peu près constant peu importe le nombre de GRH utilisés. Par contre, pour le SFR, le BRM de la méthode RL_4 est nettement plus petit que le BRM des trois autres méthodes de la RL. Ceci peut s'expliquer par le fait que le SFR est une variable dérivée à partir de plusieurs autres variables, contrairement au RT et aux ST qui sont des variables observées. Le BRM des trois variables pour la méthode 1_GRH est beaucoup plus grand que le BRM produit par le MS et par la RL sauf pour le SFR où dans ce cas, le BRM est environ équivalent au BRM de la RL. Il semble donc que la méthode 1_GRH performe moins bien que le MS et la RL. Dans le meilleur des cas, elle est environ équivalente à la RL. Contrairement au MS, on note que la progression du BRM n'est pas strictement décroissante pour la RL à mesure que le nombre de GRH augmente.

Malgré le fait que le BRM soit minime pour les variables étudiées pour le Canada, il peut augmenter rapidement pour de petits domaines. Pour la présente étude, d'autres domaines ont également été étudiés. Quoique certains écarts soient observés pour plusieurs d'entre eux, il semble que le BRM du MS soit généralement plus petit que le BRM de la RL et la méthode 1_GRH. Une étude plus approfondie sur un plus grand nombre de variables d'intérêt et de domaines serait bénéfique afin de corroborer ces conclusions.

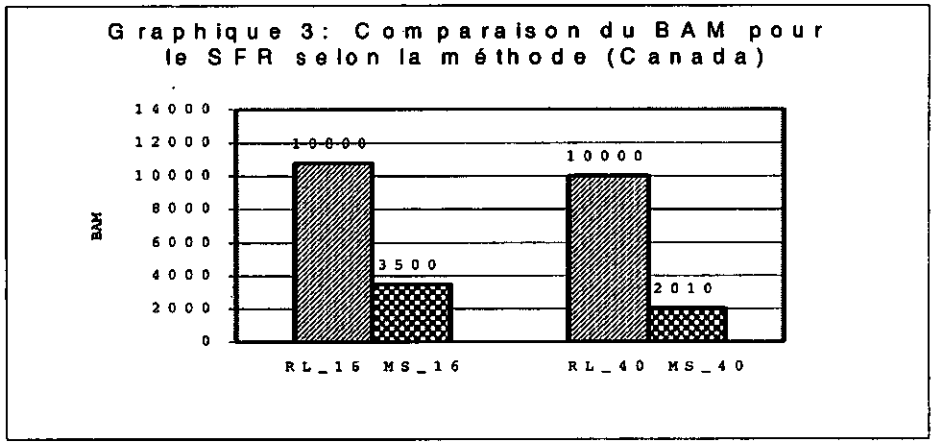
Tel que mentionné précédemment, les changements individuels dans les poids causés par l'ajustement de non-réponse sont supérieurs pour le MS que pour la RL (voir graphique 2). Ceci semble indiquer que le MS est plus efficace pour réduire le BA de non-réponse, pour un nombre fixe de GRH. Le graphique 3 confirme cette observation; on y observe que le BAM pour le SFR est inférieur selon la méthode du MS que pour la RL.

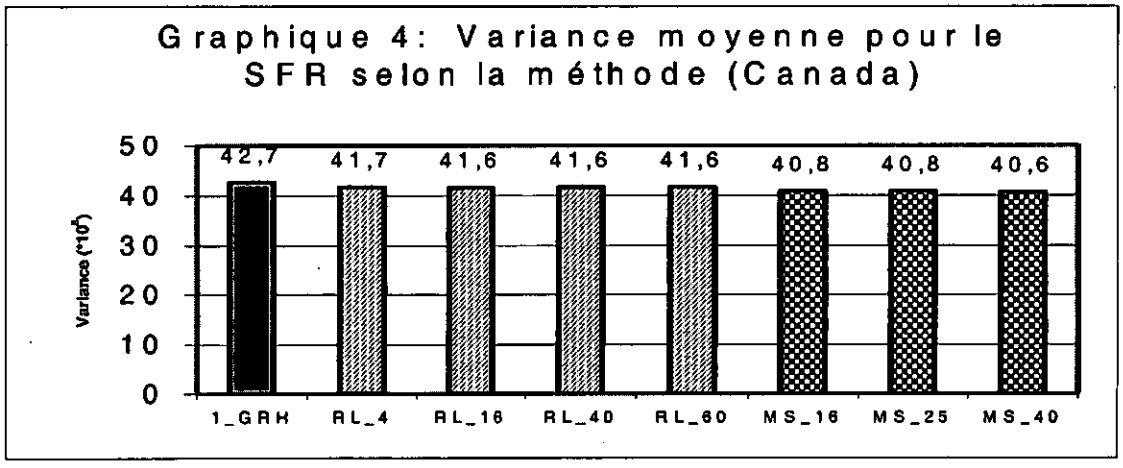
5.2.3 Estimations de variance

Des estimations de variance ont été produites pour les trois variables d'intérêt à l'aide de la méthode du jackknife. Pour le SFR (graphique 4), la variance moyenne des estimations est de façon approximative la même, peu importe la méthode étudiée. Il y a cependant une légère baisse lorsque le nombre de GRH augmente, autant pour la RL que pour le MS. De plus, selon l'étude empirique, les estimations de la variance moyenne du MS sont légèrement plus petites que celles de la RL. On constate qu'une plus grande dispersion dans les poids (une valeur de *D* plus élevée) n'entraîne donc pas une augmentation de la variance.

Tableau 2
BRM (en %) pour différentes variables selon les méthodes étudiées – Canada

Variable	MÉTHODE UTILISÉE							
	1_GRH	RL_4	RL_16	RL_40	RL_60	MS_16	MS_25	MS_40
SFR	0,37	0,15	0,43	0,37	0,31	0,14	0,12	0,08
RT	-0,32	-0,09	-0,06	-0,05	-0,06	-0,006	-0,005	0,002
ST	-0,44	-0,13	-0,15	-0,19	-0,14	-0,10	-0,09	-0,09





6. APPLICATION AUX DONNÉES DE L'ENQUÊTE NATIONALE LONGITUDINALE SUR LES ENFANTS ET LES JEUNES (ENLEJ)

Dans cette section, la majorité des analyses réalisées à l'aide de la RL et du MS lors de l'étude empirique avec les données de l'EDTR sont reproduites en utilisant l'information obtenue de l'ENLEJ. Tout comme l'EDTR, l'ENLEJ est une enquête longitudinale auprès des ménages. Elle a débuté en 1994 et son but est de recueillir de l'information pour analyser les politiques et développer des programmes portant sur les facteurs critiques qui affectent le développement des enfants au Canada (voir Michaud, Morin, Clermont et Laflamme 1998).

6.1 Description et analyse des résultats de l'application

Pour cette étude, les méthodes utilisées sont les suivantes : RL_i, où i=4, 14, 41, 70 avec respectivement q=2, 4, 6, 8 variables, et MS_i, où i=19, 36 avec des seuils de signification de 0,001 et 0,005 respectivement. Les deux mêmes contraintes imposées pour l'EDTR ont été réappliquées lors de la création des GRH. Pour chacune des méthodes à l'étude, la même poststratification a été utilisée (22 groupes d'âge-sexe par province).

Contrairement à l'étude empirique basée sur l'EDTR, seules les données recueillies aux deux premières vagues de l'ENLEJ ont été utilisées. Aucune simulation n'a été effectuée et les poids initiaux n'ont pas été normés ($\Sigma_s w_{0k} = \hat{N} < N$). Il est à noter que la sous-couverture de l'ENLEJ est d'environ 13 % et la non-réponse aux alentours de 8 %.

Les résultats présentés au tableau 3 permettent de tirer des conclusions similaires à celles obtenues lors de la simulation (tableau 1). Cependant, on observe que la contribution relative de R_{01} à la mesure de changement est plus faible pour l'ENLEJ que dans le cas de l'EDTR. Ce résultat indique que l'ajustement de non-réponse de l'EDTR cause de plus grands changements individuels dans les poids, entraînant une plus grande contribution du terme R_{01}. Dans le cas de l'ENLEJ, l'ajustement de non-réponse n'a donc pas d'effet majeur sur les changements individuels dans les poids, contrairement à ce qui a été observé dans le cas de l'EDTR.

La contribution relative de R_{12} à la mesure de changement est plus élevée pour l'ENLEJ que pour l'EDTR. Ce résultat indique que la poststratification plus raffinée de l'ENLEJ entraîne de plus grands changements individuels dans les poids, ce qui se traduit par une plus grande contribution de R_{12}. L'ENLEJ profite donc beaucoup de la poststratification alors qu'elle a une importance beaucoup moindre pour l'EDTR.

Pour ce qui est de R_{int}, tout comme pour l'EDTR, sa contribution à la mesure de changement est négligeable. Contrairement à l'EDTR, le signe de R_{int} est négatif ce qui signifie que l'interaction entre R_{01} et R_{12} est négative.

Pour ce qui est de G, comme dans le cas de l'EDTR, il est la principale source de contribution à la mesure de changement. Dans le cas de l'ENLEJ, G inclut non seulement le changement de poids moyen causé par l'ajustement de non-réponse mais aussi le changement de poids moyen causé par la correction pour la sous-couverture via la poststratification.

En comparant tous ces résultats, on s'aperçoit que les deux enquêtes se ressemblent beaucoup étant donné que $R_{int} \approx 0$ et que la somme des contributions à la mesure de changement de R_{01} et R_{12} se situe approximativement autour de 35 % dans chacun des cas. Par contre, l'ENLEJ est aussi très différente de l'EDTR car R_{12} prédomine pour la première alors que R_{01} prédomine pour la seconde.

Tout comme pour l'EDTR, D augmente avec le nombre de GRH et cette mesure est plus élevée pour le MS que pour la RL. La valeur de D est d'ailleurs plus élevée pour l'ENLEJ que pour l'EDTR principalement à cause de la sous-couverture de l'ENLEJ, ce qui a pour effet d'augmenter G et conséquemment D.

La contribution moyenne de R_{01} pour la RL et le MS augmente avec le nombre de GRH tandis que celle de R_{12} diminue (graphique 5). La contribution de R_{01} est également plus élevée pour le MS que pour la RL, contrairement à la contribution de R_{12} qui est plus petite pour le MS que pour la RL.

Tableau 3

Valeur de D, de chaque composante et de leur contribution (en %) à la mesure de changement pour chacune des six méthodes d'ajustement de la non-réponse

Méthode	D	R_{01}	R_{01}/D (%)	R_{12}	R_{12}/D (%)	R_{int} ($\times 10^{-4}$)	R_{int}/D (%)	G	G/D (%)
RL_4	0,1475	0,0052	3,51	0,0369	25,05	-4,63	-0,31	0,1058	71,76
RL_14	0,1497	0,0075	5,00	0,0367	24,69	-5,50	-0,37	0,1058	70,68
RL_41	0,1530	0,0112	7,29	0,0369	24,13	-9,16	-0,60	0,1058	69,18
RL_70	0,1564	0,0144	9,21	0,0362	23,13	-0,19	-0,01	0,1058	67,67
MS_19	0,1608	0,0187	11,63	0,0371	23,07	-8,24	-0,51	0,1058	65,81
MS_36	0,1640	0,0220	13,41	0,0373	22,76	-11,30	-0,69	0,1058	64,52

Tout comme observé lors de l'étude empirique, le profil de la contribution de R_{01} à la mesure de changement est le même que celui de la mesure elle-même. Cela montre que les variations de D dépendent directement de R_{01}.

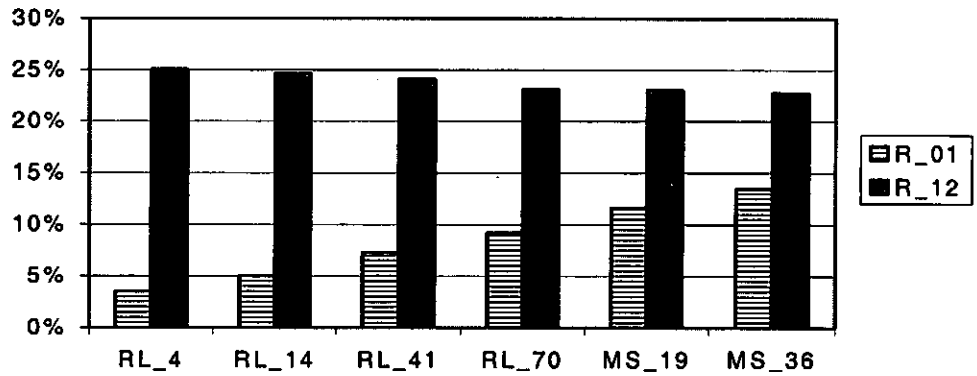
Le graphique 6 permet de comparer la RL et le MS en présentant la contribution moyenne de R_{01} à la mesure de changement pour les méthodes avec un nombre sensiblement équivalent de GRH. Comme pour l'EDTR, les résultats montrent que la non-réponse semble être mieux ciblée avec la méthode du MS qu'avec la RL.

Contrairement à l'étude de simulation de l'EDTR, le biais n'a pas été évalué étant donné qu'aucune source externe de données n'était disponible pour fins d'évaluation.

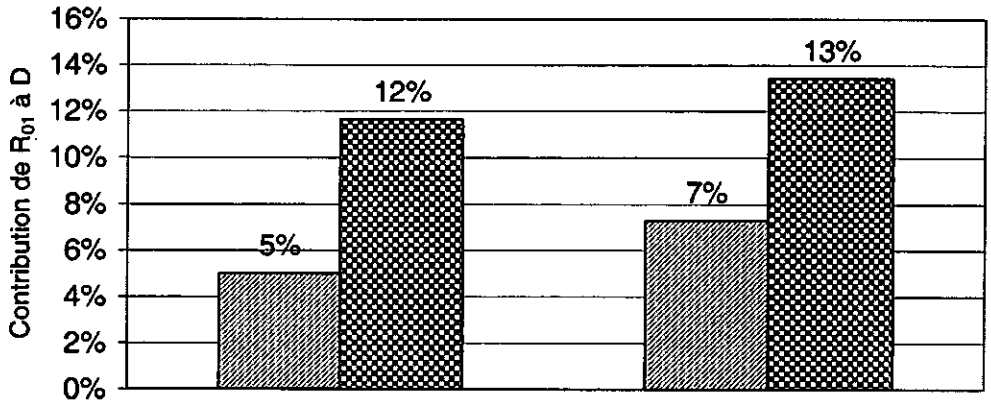
7. CONCLUSION

Ce document met en relief le fait que le choix des GRH et de la méthode pour les définir dépendent : i) de la disponibilité de l'information auxiliaire, ii) du souci de réduire le biais de non-réponse pour toutes les estimations, et iii) du temps et des contraintes opérationnelles. L'étude empirique, de même que les données de l'ENLEJ, ont démontré que la méthode du MS semble meilleure que celle de la RL pour réduire le biais de non-réponse. Les résultats ont également démontré que la mesure de changement proposée peut être un outil très utile pour comparer différentes stratégies de pondération.

Graphique 5 : Contribution moyenne de R_{01} et R_{12} à la mesure de changement (D) selon la méthode



Graphique 6 : Contribution moyenne de R_{01} à la mesure de changement (D) selon la méthode



En particulier, il semble que plus la valeur du terme R_{01} est élevée, plus la réduction du biais obtenue en utilisant les GRH est importante. Étant donné la difficulté d'obtenir une estimation fiable du biais de non-réponse dans une enquête, la relation identifiée entre la taille de la composante R_{01} et la réduction du biais suggère d'utiliser R_{01} comme un outil pour évaluer les méthodes d'ajustement de non-réponse. Pour ce faire, il faut tout d'abord déterminer R_{01} pour différents ensembles de GRH. Ensuite, l'ensemble affichant la plus grande valeur de R_{01} est sujet à être plus efficace que les autres alternatives pour réduire le biais de non-réponse, pour la plupart des variables d'intérêt.

La mesure de changement présentée pourrait aussi être utilisée pour comparer différentes stratégies de calage. Dans ce cas, l'ajustement de non-réponse pourrait demeurer le même pour toutes les méthodes de poststratification à l'étude. Une étude détaillée du comportement du terme R_{12} pourrait être effectuée et permettrait sans doute de tirer certaines conclusions comme la présente étude a permis de tirer des conclusions à propos du terme R_{01}. Ce genre d'étude n'aurait pas nécessairement besoin de se limiter au contexte longitudinal mais pourrait fort bien s'effectuer dans le cas d'une enquête transversale. De même, la mesure de changement pourrait aussi s'avérer utile pour évaluer différentes méthodes d'ajustement de non-réponse d'enquêtes transversales.

REMERCIEMENTS

Les auteurs aimeraient remercier M. Hladky, M. Latouche, C. Nadeau et N. Tremblay pour leur importante contribution à ce projet.

BIBLIOGRAPHIE

ANGOSS SOFTWARE (1995). Knowledge Seeker IV for Windows – User's Guide. ANGOSS Software International Limited.

CHAPMAN, D.W., BAILEY, L. et KASPRZYK, D. (1986). Méthodes de compensation de la non-réponse au U.S. Bureau of the Census. *Techniques d'enquête*, 12, 167-187.

DEVILLE, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. *Actes du Congrès de la Société Statistique du Canada, Recueil de la Section des méthodes d'enquête*, 103-110.

DEVILLE, J.-C., et SÄRNDAL, C.-E. (1992). Calibration estimators and generalized raking techniques in survey sampling. *Journal of the American Statistical Association*, 87, 418, 376-382.

KALTON, G., et KASPRZYK, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.

KASS, G.V. (1980). An exploratory technique for investigating large quantities of categorial data. *Applied Statistics*, 29, 119-127.

LAVIGNE, M., et MICHAUD, S. (1998). Aspects généraux de l'Enquête sur la dynamique du travail et du revenu. Document de recherche de l'EDTR, Statistique Canada, catalogue 98-05.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de la Statistique*, 54, 137-139.

MICHAUD, S., MORIN, Y., CLERMONT, Y. et LAFLAMME, G. (1998). Issues in the design of a survey to measure child development : The Experience of the Canadian National Longitudinal Survey of Children and Youth. Statistique Canada, document interne.

PLATEK, R., SINGH, M.P. et TREMBLAY, V. (1978). Adjustment for nonresponse in surveys. *Survey Sampling and Measurement*. N.K. Namboodiri, Éd. Academic Press, 157-174.

RIZZO, L., KALTON, G. et BRICK, M. (1996). Comparaison de quelques méthodes de correction de la non-réponse d'un panel. *Techniques d'enquête*, 22, 43-53.

SINGH, A. C., WU, S. et BOYER, R. (1995). Longitudinal survey nonresponse adjustment by weight calibration for estimation of gross flows. *Actes de la Conférence de l'American Statistical Association*, 396-401.

Techniques d'enquête, juin 2001
Vol. 27, n° 1, pp. 117-127
Statistique Canada

Échantillonnage et pondération d'une enquête auprès de personnes sans domicile : un exemple français

PASCAL ARDILLY et DAVID LE BLANC[1]

RÉSUMÉ

L'INSEE a réalisé en 2001 une enquête destinée à mieux connaître la population sans domicile. En l'absence de base de sondage permettant d'atteindre directement les personnes sans domicile, le principe de l'enquête est d'échantillonner des prestations qui leur sont destinées et d'interroger les individus qui bénéficient de ces prestations. Lorsque l'on désire pondérer les observations individuelles issues de l'enquête, une difficulté surgit du fait qu'un individu peut bénéficier de plusieurs prestations pendant la période de référence considérée. Cet article montre comment il est possible d'appliquer la méthode du partage des poids pour résoudre ce problème. Dans ce type d'enquête, une même variable peut donner lieu à plusieurs paramètres d'intérêt, correspondant à des populations variant avec le temps. A chaque définition des paramètres correspond un jeu de poids. L'article insiste particulièrement sur le calcul de poids « un jour moyen » et « une semaine moyenne ». On donne également des éléments sur les données de fréquentation à collecter et la correction de la non-réponse.

MOTS CLÉS : Partage des poids; base incomplète; personnes sans-domicile.

1. INTRODUCTION

L'INSEE a réalisé en 2001 une enquête destinée à mieux connaître la population des sans-domicile. Cette enquête est la première enquête représentative de ce type en France (Une enquête de ce type a été menée aux États-Unis en 1991 par le *Research Triangle Institute* (RTI) dans la région métropolitaine de Washington (RTI 1993)). Le principe de l'enquête est d'atteindre les personnes sans domicile par le biais de prestations qui leur sont destinées, hébergement de nuit et repas. Evidemment, une personne peut fréquenter une ou plusieurs prestations de la base de sondage pendant la période de référence considérée, ce qui pose une difficulté lorsque l'on désire pondérer le fichier d'observations individuelles issu de l'enquête. Dans cet article, nous montrons comment la méthode du partage des poids peut être appliquée à ce problème. Dans ce type d'enquête, contrairement à la plupart des enquêtes traditionnelles auprès des ménages, une même variable peut donner lieu à plusieurs paramètres d'intérêt, correspondant à différents concepts de population : les plus utilisés par les praticiens sont les paramètres « un jour moyen » et « une semaine moyenne ». A chaque définition des paramètres correspond un jeu de poids. Nous définissons précisément ces concepts, et insistons particulièrement sur le calcul pratique des poids correspondants. Le plan de l'article est le suivant : nous rappelons d'abord les objectifs de l'enquête, sa population de référence et son plan de sondage. Nous introduisons ensuite les paramètres d'intérêt et dérivons les estimateurs de ces paramètres issus de la méthode du partage des poids. Nous décrivons la mise en oeuvre pratique de calculs de poids « un jour moyen » et « une semaine moyenne ». Enfin, nous donnons des considérations pratiques sur la correction de la non-réponse.

2. L'ENQUÊTE « SANS DOMICILE »

2.1 Objectifs de l'enquête

L'enquête réalisée par l'INSEE en février 2001 vise à mieux connaître la population des « sans-domicile ». Cette population est généralement définie par défaut comme l'ensemble des personnes qui ne disposent pas d'un domicile fixe. Cette population échappe aux enquêtes traditionnelles auprès des ménages menées par l'Institut, celles-ci reposant sur une base de sondage de logements. En l'absence d'une base de sondage recensant cette population, le principe de l'enquête consiste à atteindre la population visée par le biais de prestations destinées aux personnes en difficulté, correspondant à des services d'hébergement et de repas. Ces prestations sont fournies sur des bases temporelles qui varient selon leur nature : les repas sont fournis chaque jour midi et soir, les nuitées une fois par jour.

Cet échantillonnage indirect induit deux distorsions entre la population visée initialement et la population atteinte par l'enquête. D'une part, la population visée n'est pas atteinte dans sa totalité : seuls ses membres qui utilisent les prestations rentrant dans le champ de l'enquête sont potentiellement échantillonnables. D'autre part, la population atteinte par l'enquête contient des personnes qui n'appartiennent pas à la population visée initialement, dans la mesure où les services destinés en priorité aux personnes sans domicile sont aussi utilisés par des personnes qui vivent en ménage ordinaire mais sont en situation précaire (c'est surtout le cas pour les repas). Dans tout ce qui suit, tout en gardant cette distinction à l'esprit, on parlera cependant parfois de « sans-domicile » pour désigner les personnes fréquentant les prestations du champ de l'enquête.

[1] Pascal Ardilly et David Le Blanc, Institut National de la Statistique et des Études Economiques, 18 boulevard Adolphe Pinard, 75675, Paris, Cedex, France. Courrier électronique : pascal.ardilly@insee.fr, leblanc@ensae.fr .

2.2 La population de référence

La caractéristique principale des services considérés est qu'ils sont fournis dans un lieu précis ; ce lieu est appelé par la suite *centre*. A un centre donné correspond un ou plusieurs types de services. L'unité statistique échantillonnée, que nous appellerons par la suite *prestation*, sera définie comme un quadruplet (service, jour, intervalle de temps, personne) : il s'agit d'un service de type donné dans un centre donné, un jour donné, dans un intervalle de temps donné, à une personne donnée. Une personne peut bien sûr bénéficier de plusieurs prestations la même journée, et *a fortiori* une semaine donnée ou pendant le mois d'enquête.

La *période de référence* de l'enquête s'étend sur un mois (du 15 janvier au 15 février 2001). On note J l'ensemble des jours de la période de référence de l'enquête, repérés par l'indice j.

Le *champ géographique* de l'enquête est celui des agglomérations de plus de 20 000 habitants.

Les *prestations dans le champ de l'enquête* sont celles qui relèvent d'un des deux types de services retenus : repas et hébergement, dès lors qu'ils sont assurés au moins une journée pendant la période de référence de l'enquête.

La *population de référence*, notée $P(J)$, est constituée des personnes qui ont bénéficié d'au moins une prestation du champ de l'enquête pendant la période de référence.

Cette population d'intérêt dépend fondamentalement de la période de référence. Sa taille croît avec la longueur de cette période, mais « moins vite » que le temps : en effet, d'un jour sur l'autre, on retrouve certaines personnes dans les centres. En réalité, l'évolution de $P(J)$ avec J est complexe, car deux phénomènes distincts interviennent, dont on peut penser qu'ils ont des temps caractéristiques différents :

- la population « sans-domicile » à un moment donné ne fréquente qu'épisodiquement les centres de la base : pour prétendre la couvrir, il faut donc enquêter sur une période de temps où toutes les personnes de cette population ont au moins une fois recours à des services, (cette période n'est pas connue, mais il est admis en France, « à dire d'expert », que la population non couverte pendant un mois d'hiver complet est de taille négligeable).
- la population « sans-domicile » se renouvelle dans le temps. D'une année sur l'autre, des entrées et des sorties, sans doute nombreuses, interviennent, liées aux mouvements démographiques ou aux évolutions conjoncturelles ou structurelles de la société (entrées et sorties de situations de précarité).

La question de la détermination de J revient finalement à savoir si on s'intéresse plutôt à une notion de sans-domicile « à un instant donné » (J plutôt court), ou à une notion de sans-domicile sur une longue période (J plutôt long). L'approche retenue par l'INSEE constitue un compromis entre les deux.

2.3 Le plan de sondage de l'enquête

Le plan de sondage de l'enquête comprend trois degrés : tirage d'agglomérations, tirage de centres et d'intervalles de temps, et enfin tirage de prestations.

2.3.1 Tirage des agglomérations

Le premier degré du plan de sondage consiste à tirer des agglomérations, proportionnellement à un critère de taille défini comme une combinaison de la population des agglomérations et des capacités d'accueil telles qu'elles ont pu être recensées dans les fichiers des associations et les fichiers du Ministère de la Santé. Ce premier degré de tirage est effectué plusieurs mois avant les autres. Ce décalage s'impose car le recensement exhaustif des centres et des informations les concernant (type de service rendu, capacité moyenne, jours d'ouverture, ...) est entrepris sur les agglomérations tirées. Cette opération est réalisée en deux fois : une enquête lourde l'année précédant la collecte, et une mise à jour juste avant le début de la collecte. On obtient ainsi une base de sondage de centres. Cette base joue un rôle fondamental : des personnes qui ne fréquenteraient que des centres non recensés seraient de fait non échantillonnables.

2.3.2 Tirage des centres, des jours et des intervalles de temps

Pour des raisons pratiques, il n'est pas possible d'enquêter l'ensemble des centres et de maintenir sur le terrain, dans un centre donné, un enquêteur durant une journée entière. Enfin, on ne peut interroger toutes les personnes dans un centre. Il est donc incontournable d'échantillonner :

- des centres dans les agglomérations tirées (indice c)
- des jours d'enquête pendant la période de collecte (indice j)
- des intervalles de temps pendant les jours d'enquête (indice t).
- des personnes au sein d'un (centre, jour, intervalle de temps) tiré.

Pour des raisons théoriques, *les intervalles de temps sont définis de façon qu'un individu ne puisse pas bénéficier de deux prestations différentes durant cet intervalle de temps* (par exemple, un de ces intervalle de temps est la période de 11 à 14 heures). En effet, la mesure des liens avec la base de sondage ne peut raisonnablement s'effectuer qu'en permettant aux personnes interrogées de repérer facilement dans le temps et l'espace les prestations qui leur ont été servies au cours de la période d'enquête. Pour les centres offrant des repas, un intervalle de temps recouvrira les repas du midi et un intervalle les repas du soir. On considère qu'une personne ne peut fréquenter qu'un seul centre durant l'intervalle de temps correspondant au repas de midi, faute de quoi il faudrait lui demander si elle n'a pas déjà pris un repas ailleurs, ou si elle ne mange pas deux fois dans

le même centre. Il se trouve par ailleurs que la largeur d'un intervalle assurant une telle propriété correspond à la durée au cours de laquelle on peut raisonnablement demander à un enquêteur d'interroger sur place (soit 2 à 3 heures au maximum). (On remarquera que les accueils de jour ne font pas partie des services retenus dans le champ de l'enquête. Cette restriction de champ correspond à deux préoccupations. D'une part, il serait très difficile de découper la journée en intervalles de temps de trois ou quatre heures et de collecter les liens sur la base de ce découpage (l'effort de mémoire demandé à l'enquêté serait considérable et n'a pas paru raisonnable aux concepteurs de l'enquête). D'autre part, les fréquentations de ces services sont très peu prévisibles. On a voulu éviter à une équipe d'enquêteurs de se déplacer et de ne réaliser aucun entretien faute de fréquentation.)

En fait, il n'y a pas de différence fondamentale entre l'échantillonnage des centres et l'échantillonnage des périodes de temps : les unités pertinentes à considérer sont les triplets (c, j, t) correspondant au croisement d'un centre, d'un jour et d'un intervalle de temps. Certaines cases du tableau croisant « temps » et « centres » seront éliminées *a priori* avant le tirage, soit parce que le centre est fermé durant le créneau horaire considéré, soit parce que la fréquentation y est manifestement très faible. (Dans ce dernier cas, il faut prendre garde à l'éventuelle restriction du champ couvert, s'il s'avérait que des personnes ne fréquentent que ce centre et ne sont présentes que dans ce créneau horaire. Si ces dernières sont atypiques, des biais seront introduits dans les estimations.)

Le mode de tirage retenu a consisté à tirer au hasard des triplets (centres, jour, intervalles de temps) proportionnellement à la taille des centres obtenue lors du recensement des centres. (En pratique, des regroupements d'intervalles de temps ont eu lieu dès lors qu'un centre était échantillonné plus de quatre fois au cours de la période d'enquête, pour des raisons d'acceptabilité par les responsables de centre.) Une stratification par type de centre a été effectuée. (Pour les services d'hébergement, une stratification sur le critère hommes seulement/femmes seulement/accueil mixte a été introduite.) Toutefois, cette stratification « de précaution », ne portant pas directement sur les unités d'observation, n'aura été utile que si le comportement des personnes diffère sensiblement selon le type de centre où on les trouve.

2.3.3 Tirage des prestations

Ce dernier degré consiste à achever l'échantillonnage des prestations, c'est-à-dire à tirer des personnes dans un centre sélectionné un jour donné dans un intervalle de temps donné. Les données recueillies lors du recensement des centres ne suffisent pas en général à constituer une base de sondage de prestations. Dans certains centres d'hébergement, il peut exister des listes ; c'est le cas le plus favorable, un tirage des personnes pouvant être conduit à partir de ces listes. En revanche, dans la majorité des centres (par exemple, dans un point-soupe), on ne connaît même pas le nombre de personnes qui vont se présenter durant un intervalle de temps donné : on ne peut donc pas faire de base de sondage des prestations. L'échantillonnage des prestations s'effectue à probabilités égales. Comme traditionnellement dans les sondages à plusieurs degrés, tirer un nombre constant de prestations (dernier degré) permet d'assurer des probabilités de tirage constantes, et donc de limiter les risques d'explosion de variance.

En pratique, la méthode de tirage retenue varie d'un type de centre à un autre, selon la topographie des lieux : liste existante, file d'attente, arrivées espacées dans le temps, population « groupée » sans ordre dans un même lieu au même moment, *etc.* Elle tient aussi compte du nombre maximal d'interviews raisonnablement assurables par le ou les enquêteurs pendant l'intervalle de temps de l'enquête, et du fait qu'il n'est pas souhaitable de retenir des personnes échantillonnées trop longtemps après la fermeture d'un centre ou l'arrêt de distribution de repas, sous peine d'augmenter la non-réponse.

Dans tous les cas, un « dénombreur » compte pendant la période d'échantillonnage le nombre N de prestations servies. Ce rôle est essentiel pour déterminer la probabilité de tirage des prestations échantillonnées. Parallèlement, il procède à un tirage de type systématique (Dans l'idéal, le tirage devrait être assuré par une autre personne (ou « échantillonneur »), afin d'éviter les erreurs de mesure sur la fréquentation. Des raisons budgétaires ont conduit à ne pas retenir cette solution.) selon la méthode suivante :

- dans les centres où une liste est disponible, on tire n prestations, n étant fixé avant l'enquête.
- dans les centres où aucune liste n'est disponible, on tire les prestations avec un taux de sondage f fixe. f est déterminé selon le nombre de prestations attendues $\tilde{N}$ et le nombre de prestations que l'on désire échantillonner $\tilde{n}$, afin d'assurer des probabilités de tirage égales. Dans ce cas, la taille de l'échantillon est inconnue *a priori*.

3. LES PARAMÈTRES D'INTÉRÊT

Les quantités d'intérêt sont essentiellement du type totaux ou ratios. Nous désirons estimer un total relatif à une variable y définie sur la population $P(J)$,

$$Y_J = \sum_{k \varepsilon P(J)} y_k. \qquad (1)$$

Un cas particulier de ces totaux est l'effectif de $P(J)$, $N_J = \text{card}(P(J)) = \Sigma_{k \varepsilon P(J)} 1$.

Nous désirons également estimer la moyenne de y dans la population de référence,

$$\bar{Y}_J = \frac{Y_J}{N_J} = \frac{1}{N_J} \sum_{k \varepsilon P(J)} y_k. \qquad (2)$$

Par exemple, y peut être la nationalité de l'individu, l'âge auquel il a terminé ses études, ou le nombre de centres qu'il a fréquentés le jour de l'entretien.

Nous serons par la suite amenés à distinguer deux types de variables :

- les variables fixes au cours de la période de référence de l'enquête (par exemple, l'âge de fin d'études).
- les variables qui varient au cours de la période de référence de l'enquête ($y_k = y_k(j)$). Le nombre de centres fréquentés le jour de l'enquête appartient à cette catégorie.

Nous traitons d'abord le cas des variables fixes au cours de la période de référence de l'enquête. La section 6 aborde brièvement le cas des variables qui varient au cours du temps.

4. ESTIMATION D'UN TOTAL OU D'UN RATIO DANS LE CAS OÙ LA VARIABLE D'INTÉRÊT EST CONSTANTE SUR LA PÉRIODE D'ENQUÊTE

Pour la commodité de l'exposé, nous ne faisons pas apparaître explicitement tous les degrés de tirage. Nous nous plaçons au niveau d'une agglomération échantillonnée au premier degré du tirage.

On note :

C : ensemble des centres de l'agglomération ouverts au moins un jour de la période d'enquête, repérés par l'indice c

$\Pi_{c,j,t}$: ensemble des prestations servies dans le centre c le jour j pendant l'intervalle de temps t, repérées par l'indice i.

$\Pi_{j,t}$: ensemble des prestations servies dans l'agglomération le jour j pendant l'intervalle de temps t.

$P_{c,j,t}$: l'ensemble des personnes se présentant dans le centre c le jour j pendant l'intervalle de temps t, repérées par l'indice k

$P_{j,t}$: l'ensemble des personnes se présentant dans un des centres de l'agglomération le jour j pendant l'intervalle de temps t.

De la définition des intervalles de temps, il ressort qu'à chaque individu $k \in P_{j,t}$, correspond une et une seule prestation i. Ainsi, il existe une correspondance biunivoque entre $P_{j,t}$ et $\Pi_{j,t}$. Dit autrement, pour tout couple (j, t), les $P_{c,j,t}$ sont disjointes. En revanche, $P_{c,j,t}$ et $P_{c',j',t'}$ peuvent avoir une intersection non vide, dès que $t \neq t'$.

La population d'intérêt s'écrit alors

$$P(J) = \bigcup_{c,j,t} P_{c,j,t} = \bigcup \Big(\coprod_{c \in C} P_{c,j,t} \Big).$$

Le point central du raisonnement consiste à exprimer le total d'une variable sur la population des *individus* (qui est notre total d'intérêt) comme le total d'une autre variable sur la population des *prestations* (qui sont les unités échantillonnées), l'estimation de ce dernier ne posant aucune difficulté particulière. Pour obtenir ce résultat, on peut recourir à un raisonnement direct, ou appliquer la méthode du partage des poids, l'un ou l'autre pouvant sembler plus naturel.

En raisonnant directement, nous définissons l'application K, qui à toute prestation i servie durant la période de référence J dans l'ensemble des centres du champ de l'enquête, associe l'individu bénéficiaire de cette prestation.

$$K : \{\text{prestations}\} \to \{\text{individus}\}.$$
$$i \to K(i)$$

La population d'intérêt $P(J)$ est l'image par K de $\Pi(J)$, l'ensemble des prestations servies durant la période de référence dans l'ensemble des centres du champ de l'enquête. Pour tout $k \in P(J)$, on définit $r_k(J) = \text{card}(K^{-1}(k))$, le nombre de prestations servies à l'individu k durant la période J dans l'ensemble des centres du champ de l'enquête, que nous appellerons aussi « nombre de liens ».

On a l'égalité fondamentale :

$$Y_J = \sum_{k \in P(J)} y_k = \sum_{i \in \Pi(J)} \frac{y_{K(i)}}{r_{K(i)}(J)}. \tag{3}$$

En effet, la variable y prenant la même valeur pour toutes les prestations i « pointant » sur l'individu k, c'est-à-dire telles que $K(i) = k$, le membre de droite peut s'écrire

$$\sum_{k \in P(J)} \left[\sum_{i \in \Pi(J); K(i) = k} \frac{y_k}{r_k(J)} \right] = \sum_{k \in P(J)} \frac{y_k}{r_k(J)} \left[\sum_{i \in \Pi(J); K(i) = k} 1 \right].$$

Mais la quantité entre crochets est le nombre de prestations servies à l'individu k durant la période J, soit $r_k(J)$, ce qui prouve l'égalité.

On peut alors voir $y_{K(i)}$ comme attaché à la prestation i correspondante et noter y_i au lieu de $y_{K(i)}$, et $r_i(J)$ au lieu de $r_{K(i)}(J)$. En notant $z_i = y_i / r_i(J)$, $Z = \sum_{i \in \Pi(J)} z_i$, on a $Z = Y_J$.

La formule (3) n'est autre que celle du partage des poids. Le raisonnement ci-dessus est d'ailleurs celui qui sous-tend cette méthode. (Seules les notations changent; la méthode du partage des poids décrit les liens entre la population échantillonnée et la population d'intérêt par une matrice plutôt qu'une application, une même unité de la population échantillonnée pouvant « pointer » vers plusieurs unités de la population d'intérêt). Le principe de cette dernière est rappelé en annexe 1.

4.1 Estimation d'un total

Supposons maintenant que l'on dispose d'un échantillon s_Π de prestations, auquel est associé un jeu de poids $(w_i)_{i \in s_\Pi}$. Nous supposons ces poids sans biais (il s'agit de l'inverse des probabilités d'inclusion des prestations dans

l'échantillon). s_Π définit implicitement un échantillon d'individus s_P, précisément l'ensemble des individus destinataires des prestations échantillonnées. La formule du partage des poids (voir annexe 1) assure que l'estimateur

$$\hat{Y}_J = \sum_{s_P} y_k \tilde{w}_k$$

est sans biais, où l'on a posé pour tout $k \varepsilon s_P$:

$$\tilde{w}_k = \frac{1}{r_k(J)} \sum_{s_\Pi : K(i) = k} w_i. \qquad (4)$$

La formule (4) énonce simplement que le poids d'un individu est égal à la somme des poids des prestations qui ont servi à « l'attraper », divisée par le nombre de liens avec la base de sondage, $r_k(J)$. On peut donc travailler *directement sur les individus échantillonnés* : pour chaque individu k, on calcule le poids $\tilde{w}_k$, et on estime le total Y_J par $\hat{Y}_J$.

La figure 1 donne un exemple fictif d'échantillonnage. L'univers des prestations contient 13 prestations, atteignant 8 personnes. 6 prestations sont échantillonnées. L'échantillon d'individus contient 5 personnes, l'individu numéro 2 ayant été « attrapé » par deux prestations différentes. Selon la formule (4), les poids des individus échantillonnés seront égaux à :

$$\tilde{w}_1 = w_1, \tilde{w}_2 = \frac{1}{2}(w_2 + w_8), \tilde{w}_3 = w_{10}, \tilde{w}_6 = w_7, \tilde{w}_7 = \frac{1}{3} w_9.$$

Si les prestations ont toutes le même poids égal à 13/6 (par exemple si les prestations ont été tirées par sondage aléatoire simple), le nombre de personnes ayant fréquenté les services pendant la durée de l'enquête est estimé par

$$\hat{Y}_J = \sum_{s_P} \tilde{w}_k = \frac{13}{6}\left[1 + \frac{1}{2}.2 + 1 + 1 + \frac{1}{3}\right] = \frac{169}{18} \approx 9.39.$$

Dans le cas présent où la variable étudiée ne varie pas au cours de la période d'enquête, il est indifférent pour le biais de l'estimateur d'identifier les personnes fréquentant les prestations. Considérons en effet un individu « attrapé » par deux prestations différentes de poids w_1 et w_2. Deux cas peuvent se produire en pratique :

- on repère que l'individu est le même; la pondération associée à cet individu sera égale à $(w_1 + w_2)/r_{k(J)}$, et le terme correspondant à l'individu dans l'estimateur sera égal à $y_k\ (w_1 + w_2)/r_{k(J)}$.
- on ne repère pas que l'individu a déjà été interrogé; on comptera deux individus différents; les pondérations associées à ces individus seront égales à $w_1/r_{k(J)}$ et $w_2/r_{k(J)}$, et le terme correspondant à ces deux pseudo-individus dans l'estimateur sera encore égal à $y_k\ (w_1 + w_2)/r_{k(J)}$.

Cela suppose bien sûr que les informations données par la même personne enquêtée à deux endroits/jours différents soient les mêmes, ce qui est loin d'être acquis.

En revanche, le repérage des individus peut s'avérer important pour limiter la non-réponse (voir section 7).

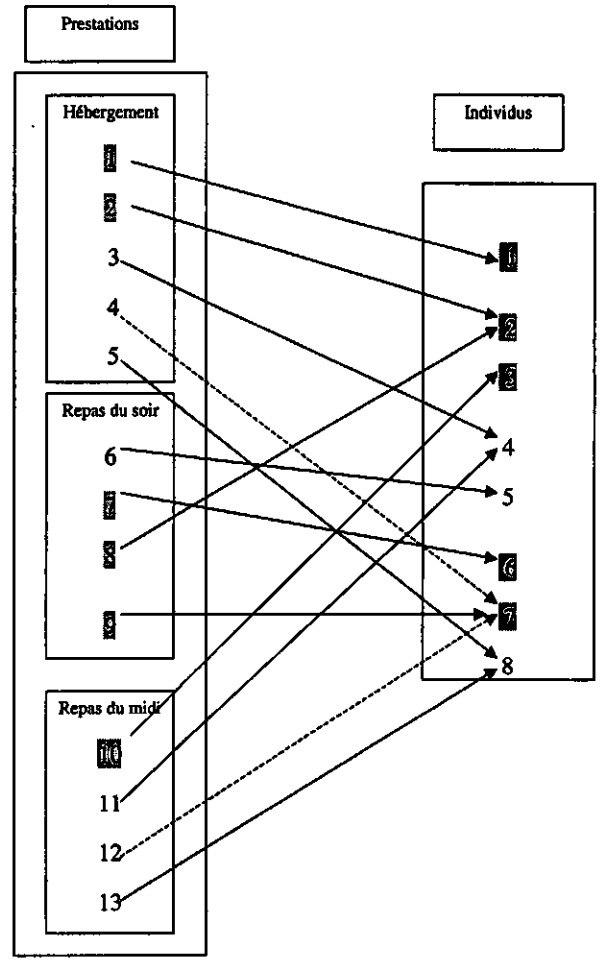


Figure 1. Les flèches représentent les liens entre les prestations et les individus. Les prestations dont l'identifiant est en grisé sont échantillonnées. Elles renvoient aux individus en grisé. Les traits en pointillé représentent les liens déclarés par l'individu 7 qui n'ont pas servi à l'échantillonner.

4.2 Estimation d'un ratio

On suppose maintenant que l'on s'intéresse à l'estimation de la moyenne $\bar{Y}_J$ (voir la formule 2). $\bar{Y}_J$ peut être estimé par l'estimateur de Hajek,

$$\hat{\bar{Y}}_J = \frac{\hat{Y}_J}{\hat{N}_J}$$

où $\hat{N}_J = \Sigma_{k \varepsilon s_P} \tilde{w}_k$.

4.3 Calcul de variance

La variance des estimateurs présentés ci-dessus se calcule classiquement à condition de raisonner à partir des

prestations. Le calcul demeure complexe puisqu'il s'agit d'un plan à plusieurs degrés à probabilités inégales. Il est évidemment indispensable, pour ne pas sous-estimer la vraie variance, de conserver toutes les prestations dans les cas où plusieurs prestations échantillonnées renvoient au même individu.

4.4 Comparaison avec d'autres méthodes d'estimation

Ayant introduit des estimateurs de type « partage des poids », il convient de s'interroger sur une méthode d'estimation alternative, où l'on tenterait d'estimer directement les probabilités de sélection des individus dans l'échantillon. (L'estimateur du partage des poids n'est pas l'estimateur classique de Horvitz-Thompson : en effet, les poids de cet estimateur dépendent clairement de l'échantillon complet de prestations (voir la formule 4). Cette méthode peut en effet sembler plus naturelle. Deux remarques s'imposent :

- pour obtenir les probabilités de sélection des individus physiques, il n'est pas raisonnablement possible de procéder autrement que par l'intermédiaire des prestations qui renvoient à l'individu, à partir d'informations fournies par ce dernier quant à sa fréquentation des différents centres. Avec les notations précédentes, nous avons :

$$\text{Prob}\,(k \,\varepsilon\, s_p) = \text{Prob}\,(\bigcup_{i \varepsilon \Pi(J);\, K(i)=k} i).$$

La formule de Poincaré permet d'exprimer cette probabilité à partir des probabilités d'inclusions simples, doubles, triples, *etc*, des prestations. Mises à part les probabilités d'inclusion simples, ces probabilités sont complexes car issues de tirages à probabilités inégales et sans remise. Il ne faut donc pas espérer obtenir une expression calculable pour Prob $(k \,\varepsilon\, s_p)$. Par contraste, la méthode du partage des poids est très simple à mettre en oeuvre.

- de manière plus structurelle, un problème vient du fait que les probabilités de sélection des prestations non échantillonnées ne sont pas connues *a priori*, du fait du tirage à plusieurs degrés. Aux degrés inférieurs, les probabilités de tirage dépendent du tirage précédent. Dans notre cas, on ne connaît pas la fréquentation des centres qui ne sont pas enquêtés. Pour obtenir la probabilité de sélection d'un individu, il faut connaître les probabilités d'inclusion de *toutes* les prestations dont il a bénéficié. Par contraste, une des forces de la méthode du partage des poids est de ne faire dépendre les poids des unités atteintes indirectement (ici les individus) que des poids des unités échantillonnées directement (les prestations). Cet avantage de la méthode est mentionné dans Lavallée (1995).

5. PROBLÈMES D'ESTIMATION ET SOLUTIONS PRATIQUES DANS LA CAS D'UNE VARIABLE CONSTANTE

Dans les formules présentées précédemment, la connaissance des liens des personnes avec l'univers des prestations est indispensable. Or, ces quantités ne sont pas connues, pour plusieurs raisons :

- une raison théorique : parce que la collecte est étalée dans le temps, et qu'un individu interrogé en début de période ne peut pas prévoir les services qu'il va fréquenter après la date d'entretien (Notons que la collecte doit nécessairement être étalée dans le temps, si l'on vise une bonne couverture la population-cible; une collecte synchrone, même si elle était techniquement réalisable, n'attendrait pas l'ensemble de la population-cible mais seulement les personnes qui fréquentent les services à cette date.),
- des raisons pratiques : parce que la mémoire des personnes interrogées fait défaut au-delà de quelques jours, et parce que la détection par l'enquêteur ou le concepteur d'enquête de prestations servies dans des centres n'appartenant pas à la base de sondage s'avère très difficile.

En pratique, il est donc impossible d'estimer sans biais un total d'intérêt sur la période de l'enquête (un mois) sans faire des hypothèses *a priori* (voir la section 5.3).

5.1 Estimation « un jour moyen », « une semaine moyenne »

On est donc amené à s'intéresser à des quantités qui font intervenir les liens sur une période courte, par exemple le jour ou la semaine. La population des personnes qui fréquentent les services du champ de l'enquête un jour j donné est $P_j = \cup_{c,t} P_{c,j,t}$. Introduisons les quantités suivantes relatives au jour j :

$$\Theta_j = \sum_{k \varepsilon P_j} y_k$$

$$N_j = \sum_{k \varepsilon P_j} 1 = \text{card}(P_j).$$

Si $\tau = \text{card}\,(J)$ est le nombre de jours de la période de référence de l'enquête, nous définissons les paramètres d'intérêt suivants :

- le total de y sur la population des personnes qui fréquentent les services du champ de l'enquête un jour « moyen », dans le sens suivant :

$$\Theta = \frac{1}{\tau} \sum_{j=1}^{\tau} \Theta_j. \qquad (5)$$

Un cas particulier est le nombre de personnes qui fréquentent les services du champ de l'enquête un jour « moyen », $\bar{N} = 1/\tau \sum_{j=1}^{\tau} N_j$.

De manière identique, la moyenne de y sur la population des personnes qui fréquentent les services du champ de l'enquête un jour « moyen » est définie comme :

$$\psi = \frac{\Theta}{\bar{N}} = \frac{\sum_{j=1}^{\tau} \Theta_j}{\sum_{j=1}^{\tau} N_j}. \qquad (6)$$

La définition des totaux ou moyennes une semaine donnée ou une « semaine moyenne » suit le même principe.

Pour estimer ces paramètres, il suffit d'adapter les formules de la section précédente, en constatant que *les* $r_k(J)$ *doivent être remplacés par le nombre de prestations du champ de l'enquête dont la personne échantillonnée a bénéficié le jour (resp. la semaine) d'enquête.*

Notons s_j l'échantillon des personnes interrogées le jour j, $r_k(j)$ le nombre de prestations de l'univers reçues par l'individu k le jour j uniquement, et $s_k(j)$ les prestations échantillonnées le jour j qui renvoient à l'individu k.

$$\Theta_j \text{ sera estimé par } \hat{\Theta}_j = \sum_{k \varepsilon s_j} y_k \tilde{w}_k,$$

$$\text{où } \tilde{w}_k = \frac{1}{r_k(j)} \sum_{i \varepsilon s_k(j)} w_i.$$

Ici, les poids des individus dépendent du jour j. (Mais *pas* les poids des prestations, w_i, qui sont fixés une fois pour toutes (en l'absence de non-réponse, il s'agit de l'inverse des probabilités de sélection des prestations)). Pour se convaincre de la différence entre Θ et Y_J, l'analogie suivante est commode : on considère un guichet où chaque personne qui arrive doit remplir un dossier. Le cas de Y_J correspond à un fonctionnement où une personne remplit un dossier la première fois où elle se présente au guichet, et n'en remplit plus les fois suivantes; le cas du « jour moyen » correspond à un fonctionnement où toute personne se présentant doit remplir un dossier, qu'elle soit déjà venue un jour précédent ou pas. Au bout d'une semaine par exemple, l'analyse des caractéristiques des personnes ayant rempli des dossiers sera très différente dans les deux cas : dans le deuxième cas, *les personnes qui viennent souvent au guichet seront surreprésentées par rapport au premier cas.* Il est possible de formaliser cette approche. Nous renvoyons le lecteur intéressé à Ardilly et Le Blanc (1999).

5.2 Estimation pratique des liens avec la base de sondage

Même si l'on se restreint à estimer des quantités de type « semaine moyenne » ou « jour moyen », il n'est pas en général possible de connaître les liens avec la base de sondage un jour donné (et *a fortiori* une semaine donnée ou sur toute la période de l'enquête).

5.2.1 Estimation « un jour moyen »

Pour partager les poids, il faut estimer les liens relatifs au jour de l'enquête; le cas le plus problématique est celui de personnes enquêtées le midi dans un lieu de restauration; pour ces personnes, les centres fréquentés le soir même (repas et / ou hébergement) ne sont pas connus. Une option, non retenue par les concepteurs de l'enquête de l'INSEE, consiste à inclure dans le questionnaire des questions du type « Où allez-vous dîner (resp. dormir) ce soir ? ». Dans ce cas, les réponses peuvent être utilisées pour imputer des liens. La question est bien entendu de savoir si les réponses à ces questions reflètent les vrais liens, et d'autre part si la non-réponse à cette question ne sera pas trop élevée. D'un point de vue plus statistique, on peut utiliser (en faisant l'hypothèse d'une certaine régularité des comportements) des informations portant sur le même intervalle de temps le jour précédent l'enquête. Les liens correspondants sont sans doute des approximations convenables des vrais liens. Le problème pratique concerne l'éventuelle différentiation des jours de la semaine en matière de fréquentation des centres : par exemple, certains centres ne sont pas ouverts en fin de semaine, d'autres n'ouvrent que certains jours précis.

5.2.2 Estimation « une semaine moyenne »

Pour partager les poids, on garde tous les liens relatifs à la semaine. La première option décrite en 5.2.1 est évidemment à proscrire. Pour les estimations une semaine donnée, on peut prendre comme approximation pour les services fréquentés un jour j postérieur à la date d'entretien les services fréquentés par l'individu le jour $(j - 7)$. Cela est cohérent si l'on suppose qu'il existe une certaine saisonnalité des services fréquentés selon le jour de la semaine. Cela revient à remplacer dans les estimateurs la semaine civile de référence par une semaine glissante, c'est-à-dire les sept derniers jours à compter de la date d'interview. C'est l'option qui a été prise pour l'enquête, le questionnaire étant prévu pour récolter les liens sur les 7 jours précédant l'entretien.

5.3 Estimation sur l'ensemble de la période d'enquête

Estimer des totaux et des moyennes portant sur la population $P(J)$ peut apparaître comme un des objectifs de l'enquête. Cette estimation fait intervenir les liens des individus échantillonnés avec les prestations du champ de l'enquête pendant l'ensemble de la période de collecte, qui ne sont pas connus. Il est donc nécessaire de modéliser l'évolution des liens au-delà d'une semaine, ou, ce qui revient au même, de modéliser le comportement de passage des individus dans les centres.

La solution à adopter n'est pas simple. Par exemple, l'hypothèse qui peut venir à l'esprit, à savoir poser

$$\forall k, r_k(J) = A.r_k(S) \qquad (7)$$

où A est le nombre de semaines de l'enquête et $r_k(S)$ le nombre de liens de l'individu k avec les prestations du

champ de l'enquête pendant une semaine S, conduit à des estimateurs sur l'ensemble de la période identiques aux estimateurs sur une semaine moyenne. En effet, un estimateur « semaine moyenne » pondère l'individu k par

$$\sum_{i \varepsilon s_k(J)} \frac{w_i}{A.r_k(S_i)}$$

où S_i est la semaine durant laquelle la prestation i lui est servie, et $s_k(J)$ est l'échantillon de prestations qui renvoient à l'individu k, alors qu'un estimateur théorique « ensemble de la période » pondère l'individu k par

$$\sum_{i \varepsilon s_k(J)} \frac{w_i}{r_k(J)}.$$

La relation 7 est donc une condition suffisante d'égalité de ces estimateurs. Cette condition est notamment satisfaite si pour tout j et tout k

$$r_k(J) = \text{card}\,(J).r_k(j) \qquad (8)$$

c'est-à-dire si le nombre de liens journaliers ne dépend pas de j.

Cette hypothèse est certainement trop forte. Pour aller plus loin sur ce point, il faut sans doute exploiter les informations fournies par l'enquête elle-même sur le comportement des individus en matière de fréquentation des centres.

Le chiffre le plus demandé de l'enquête, dans le contexte français, sera sans doute une estimation de la taille de la population « sans-domicile », soit une estimation de la taille de $P(J)$. Au-delà des questions de comptage des liens déjà abondamment évoquées, cette estimation se heurte à plusieurs insuffisances de la base de sondage ainsi qu'au caractère indirect de l'échantillonnage.

- Le risque d'oublier certaines structures lors du dénombrement des centres est important. Même si l'inventaire est exhaustif, le décalage temporel entre cet inventaire et l'enquête à proprement parler rend probable l'apparition de nouvelles structures non recensées dans la base de sondage. Cela peut générer un biais dans la mesure où certains des individus qui fréquenteraient ces structures ne fréquenteraient par ailleurs aucun service de la base de sondage (On pourrait également s'attendre à un refus de coopération de la part des responsables de certains centres : pour l'enquête INSEE, le refus des institutions a été pratiquement inexistant (moins de 1 % de refus). Cela est dû en grande partie à un important travail de sensibilisation au moment du recensement des centres et juste avant l'enquête). Par ailleurs, l'absence de biais est conditionnée par un calcul correct des liens, les passages dans des centres non recensés ne devant pas être comptabilisés dans ces liens.
- Les individus qui fréquenteraient des centres uniquement en dehors des heures « classiques » (concrètement, celles où on se sera donné les moyens de compter les prestations) sont hors champ de l'enquête. (Leur dénombrement poserait des problèmes considérables de mise en oeuvre sur le terrain).
- Une autre source de biais peut provenir du délicat comptage du nombre total de prestations servies dans les centres lors de l'enquête, ces nombres servant à calculer la probabilité pour une prestation d'être échantillonnée. Pour des raisons budgétaires, une seule personne assure le comptage des prestations et l'échantillonnage, ce qui peut poser des problèmes de rigueur d'échantillonnage en cas de situation confuse sur le terrain.
- Au niveau des concepts, il demeure une difficulté puisque l'enquête doit se dérouler sur un mois et que la population-cible évolue au cours de la période.

L'estimation de la taille de la population est donc particulièrement fragile. Pour cette raison, on peut s'attendre à ce que les erreurs commises soient plus importantes pour les totaux que pour les moyennes.

6. ESTIMATION DANS LE CAS DE VARIABLES D'INTÉRÊT NON CONSTANTES AU COURS DE LA PÉRIODE D'ENQUÊTE

Certaines variables d'intérêt de l'enquête dépendent de la date d'observation, et ne sont donc pas constantes au cours de la période d'enquête. Ce peut être le cas de réponses à des questions portant sur la journée précédant l'interview, par exemple « Combien de repas avez-vous pris hier ? », « Combien de fois avez-vous dormi dans la rue la semaine dernière? », *etc.* Les questions sur les liens sont également dans ce cas de figure. Il est donc important de voir dans quelle mesure on peut adapter le formalisme précédent à des estimations portant sur ce type de variables. Soit donc y une telle variable d'intérêt.

Si nous revenons à l'expression (3), il est facile de voir que la constance des y_k au cours de la période d'enquête est la condition qui permet de factoriser y_k et de faire apparaître les liens $r_k(J)$. On en déduit que *le type de calcul mené ci-dessus est toujours valable pour des estimations portant sur des périodes plus courtes que la période sur laquelle les* y_k *sont constants.*

Ainsi, pour des variables constantes sur un jour, on pourra parfaitement utiliser des estimateurs « un jour moyen ». Pour des variables constantes sur la semaine, on pourra utiliser des estimateurs « un jour moyen » ou « une semaine moyenne ».

7. CORRECTION DE LA NON-RÉPONSE TOTALE

Pour décrire complètement l'opération, il reste à préciser comment passer d'un jeu de probabilités d'inclusion (et donc de poids initiaux des prestations incluses dans

l'échantillon) à un jeu de poids sur les prestations répondantes. En effet, certaines personnes vont accepter l'entretien, d'autres non. On parlera dans le premier cas de prestation répondante, dans le deuxième de prestation non répondante. Les méthodes habituelles de correction de la non-réponse totale peuvent être mises en oeuvre. Nous suggérons une correction de la non-réponse par sous-groupes homogènes (pour une description de la méthode, voir par exemple Chambaz et Legendre 1999).

Concrètement, la difficulté majeure tient au fait qu'il n'y a pas de base de sondage d'individus, et donc pas d'information *a priori* sur les non-répondants. Dans un monde probablement très hétérogène, c'est un handicap considérable. On modélise donc le comportement de réponse des prestations. On sait depuis les enquêtes expérimentales de l'INED (Institut National des Etudes Démographiques) que la non-réponse varie fortement selon le type de centre (Firdion et Marpsat 1997). D'autres variables de la base de sondage peuvent être utilisées pour constituer des groupes homogènes (jour de la semaine, période du jour, groupes d'agglomérations, ...).

Une repondération des prestations répondantes conduit à des poids pour les prestations répondantes du type

$w_i = 1/\delta_i\pi_i$, où

π_i est la probabilité d'inclusion de la prestation i dans l'échantillon

δ_i est la probabilité estimée *a posteriori* que la prestation i donne lieu à réponse.

On obtient ainsi un jeu de poids pour les prestations répondantes.

En fait, certaines non-réponses viennent du fait qu'un même individu est échantillonné plusieurs fois : on peut penser qu'un individu échantillonné deux fois réponde lors du premier tirage, mais pas lors du second. (La fréquence d'occurrence de cet événement n'était pas connue au moment de la rédaction de cet article). Le second tirage génère alors une « fausse non-réponse ». Si celle-ci n'est pas détectée, la procédure de correction de la non-réponse totale amène à repondérer à tort, alors que la vraie valeur peut être récupérée dans un questionnaire déjà rempli. Pour éviter cela, l'enquêteur cherche à connaître le motif des refus et il doit cocher une case spécifique lorsque l'individu déclare avoir déjà été interrogé. Dans ce cas, il collecte quelques informations, dont le prénom et la date de naissance, qui doivent servir à relier ce questionnaire avec un questionnaire déjà rempli. (L'idéal serait de disposer d'un identifiant des répondants. Les impératifs de confidentialité et la prise en compte de l'accueil d'une telle mesure par les personnes interrogées ont conduit à ne pas retenir cette idée). Sur le terrain cependant, il est difficile d'obtenir le motif du refus. Même si on en dispose, des difficultés peuvent subsister. (Il est difficile de vérifier qu'un individu qui déclare avoir déjà été interrogé l'a effectivement déjà été. Même si l'individu est de bonne fois, il peut avoir été interrogé quelques jours auparavant pour une toute autre enquête que l'enquête de l'INSEE).

8. CONCLUSION

Dans cet article, nous montrons comment la méthode du partage des poids peut être utilisée pour pondérer l'enquête menée par l'INSEE pour mieux connaître les personnes sans-domicile. La méthode présente un grand nombre d'avantages. Elle permet de travailler sur un fichier d'individus, c'est-à-dire sur les unités statistiques naturelles utilisées dans la définition des paramètres d'intérêt. Simple à mettre en oeuvre, elle permet en outre de passer aisément d'une période de référence à une autre (estimation « un jour moyen », « une semaine moyenne »). Les opérations en aval de l'enquête comme la correction de la non-réponse et le calcul de variance peuvent être réalisées dans un cadre classique, car elles se font sur les unités échantillonnées (les prestations), dont on maîtrise les probabilités de tirage, et non sur les individus, dont les probabilités de tirage sont inconnues. Nous montrons qu'un critère crucial de qualité d'une telle enquête est le recueil fidèle des données de fréquentation des services par les personnes interrogées. Sans ces données, il n'est pas possible de pondérer l'enquête. La méthode du partage des poids parait un bon compromis pour une enquête dont le but n'est pas seulement de dénombrer une population, mais de mieux la connaître en passant un questionnaire. Pour une enquête visant principalement à un dénombrement des personnes sans-domicile, des méthodologies alternatives pourraient être envisagées. La première tourne autour des techniques de capture-recapture, utilisées pour connaître les effectifs de populations animales (à ce sujet, voir par exemple Pollock, Turner et Brown 1994). Ces techniques ne sont pas aisées à mettre en oeuvre dans une population souvent réticente à toute tentative d'identification, perçue négativement. Une autre technique est celle de l'échantillonnage « boule de neige », qui consiste à aller chercher les individus d'intérêt par l'intermédiaire d'individus déjà échantillonnes (Franck et Snijders 1994). Elle s'appuie sur un système de connaissances mutuelles des personnes vraisemblablement illusoire dans ce milieu. Ces méthodes se heurtent toujours à la question du repérage des individus. Dans notre cas, les seuls lieux où l'on peut trouver les personnes avec une probabilité suffisamment importante sont les centres : passer par l'intermédiaire des centres est incontournable.

REMERCIEMENTS

Les auteurs remercient le rédacteur de la revue ainsi que deux rapporteurs anonymes, dont les remarques ont permis d'améliorer le fond et la forme de l'article. Les erreurs qui demeureraient nous sont entièrement imputables.

ANNEXE 1 : LA MÉTHODE DU PARTAGE DES POIDS APPLIQUÉE AU PROBLÈME

Cette annexe rappelle brièvement le principe de la méthode du partage des poids. Pour un exposé plus complet, le lecteur pourra consulter Lavallée (1995), ou Deville (1999) dont nous reprenons les notations.

1. On dispose d'une population U de n unités, et d'une population V de m unités. Ici, les unités de U sont les prestations dans le champ de l'enquête. Les unités de V sont les personnes ayant bénéficié d'au moins une prestation pendant la période de l'enquête (autrement dit dans le cas présent $V = P(J)$ avec les notations précédentes).
2. On suppose qu'il existe des liens entre les unités des deux populations. Ces liens peuvent s'écrire sous la forme d'une matrice

$$(r_{ik})\ 1 \le i \le n\ ,\quad 1 \le k \le m$$

où $r_{ik} = 1$ si l'unité k de V est reliée à l'unité i de U, $r_{ik} = 0$ sinon. Ici, les liens relient les prestations aux personnes ayant fréquenté ces prestations : $r_{ik} = 1$ si la personne k a fréquenté la prestation i de U, $r_{ik} = 0$ sinon.
3. Toutes les unités de U ont au moins un lien avec une unité de V. Cela est évidemment réalisé ici, par définition de la population V. De plus, ici, chaque unité de la population U pointe sur une unité et une seule de V.

Dans le cas général, on s'intéresse au total d'une variable d'intérêt y sur V,

$$Y = \sum_{k \varepsilon V} y_k.$$

Si par exemple on prend $y \equiv 1$, le total d'intérêt est le nombre de personnes ayant fréquenté un service du champ de l'enquête pendant le mois de l'enquête.

On note

$$r_k = \sum_{i \varepsilon U} r_{ik}.$$

L'identité $Y = \Sigma_{i \varepsilon U} \Sigma_{k \varepsilon V} (r_{ik}/r_k)\ y_k$ permet de définir pour tout $i \varepsilon U$ la variable $z_i = \Sigma_{k \varepsilon V} (r_{ik}/r_k)\ y_k$ et on a :

$$Z = \sum_{i \varepsilon U} z_i = \sum_{k \varepsilon V} y_k = Y.$$

Supposons maintenant que l'on dispose d'un échantillon s_U issu de la population U, auquel est associé un jeu de poids $(w_i)_{i \varepsilon s_U}$. Cet échantillon définit implicitement un échantillon dans V, s_V, précisément

$$s_V = \{k \varepsilon V; \exists i \varepsilon s_U, r_{ik} = 1\}.$$

On suppose que l'on a collecté les r_{ik} pour tous $k \varepsilon s_V$, c'est-à-dire que tous les liens des individus avec l'univers U sont connus (ce point est fondamental).

Le total $Z = Y$ est estimé par $\hat{Z} = \Sigma_{s_U} w_i z_i$.

Et donc, si les poids sont sans biais (c'est-à-dire, établis de manière que $\hat{Z}$ est sans biais), $\hat{Y}$ estime sans biais Y.

On peut réécrire $\hat{Z} = \Sigma_{s_U} w_i \Sigma_{k \varepsilon V}\ r_{ik}\ \ y_k/r_k\ = \hat{Y}$.

La deuxième somme ne porte que sur s_V par définition, et donc $\hat{Y} = \Sigma_{s_V} y_k (\Sigma_{s_U}\ w_i r_{ik}/r_k) = \Sigma_{s_V} y_k \tilde{w}_k$, où l'on a posé pour tout $k \varepsilon s_V$:

$$\tilde{w}_k = \frac{1}{r_k} \sum_{s_U} w_i r_{ik}. \tag{9}$$

On peut donc travailler directement sur les individus échantillonnés. Dans notre cas, r_k est le nombre de liens, c'est-à-dire le nombre de services fréquentés par la personne interrogée pendant la période de référence de l'enquête. C'est la quantité qui est notée $r_k(J)$ dans les sections précédentes, cette dépendance en J étant destinée à rappeler que les liens intervenant dans les poids peuvent varier selon le type d'estimateur (« jour moyen », « semaine moyenne ») que l'on considère. Ce nombre se déduit des données de fréquentation collectées à l'enquête.

ANNEXE 2 : TABLEAU RÉCAPITULATIF DES NOTATIONS

J	ensemble des jours de la période de référence de l'enquête
τ	$= \mathrm{card}(J)$, nombre de jours de la période de référence
$P(J)$	population d'intérêt, ensemble des personnes qui ont fréquenté au moins une prestation du champ de l'enquête pendant la période de référence
N_J	$= \mathrm{card}(P(J))$, effectif de la population d'intérêt
C	ensemble des centres de l'agglomération, repérés par l'indice c
$\Pi_{c,j,t}$	ensemble des prestations servies dans le centre c le jour j pendant l'intervalle de temps t, repérées par l'indice i
$\Pi_{j,t}$	ensemble des prestations servies dans l'agglomération le jour j pendant l'intervalle de temps t
$P_{c,j,t}$	ensemble des personnes se présentant dans le centre c le jour j pendant l'intervalle de temps t, repérées par l'indice k
$P_{j,t}$	ensemble des personnes se présentant dans un des centres de l'agglomération le jour j pendant l'intervalle de temps t
P_j	ensemble des personnes qui fréquentent les services du champ de l'enquête un jour j
y	variable d'intérêt
Y_J	total de la variable y dans la population de référence
$\bar{Y}_J$	moyenne de y dans la population de référence

$\Pi(J)$ ensemble des prestations servies durant la période de référence dans l'ensemble des centres du champ de l'enquête

$r_k(J)$ nombre de prestations servies à l'individu k durant la période J dans l'ensemble des centres du champ de l'enquête, ou "nombre de liens"

s_Π échantillon de prestations

w_i poids associés à l'échantillon de prestations

s_P échantillon d'individus, ensemble des individus destinataires des prestations échantillonnées

$\bar{w}_k$ poids associés à l'échantillon d'individus

Θ_j total de y sur P_j

N_j $= \text{card}(P_j)$

Θ total de y "un jour moyen"

$\bar{N}$ nombre de personnes "un jour moyen"

ψ $= \frac{\Theta}{\bar{N}}$, moyenne de y "un jour moyen"

$r_k(j)$ nombre de prestations reçues par l'individu k le jour j uniquement

s_j échantillon des personnes interrogées le jour j

$s_k(j)$ ensemble des prestations échantillonnées le jour j qui renvoient à l'individu k

$s_k(J)$ ensemble des prestations échantillonnées pendant la période J qui renvoient à l'individu k

BIBLIOGRAPHIE

ARDILLY, P., et LE BLANC, D. (1999). Enquête auprès des personnes sans-domicile : éléments techniques sur l'échantillonnage et le calcul de pondérations individuelles, une application de la méthode du partage des poids. Document de travail, *INSEE*, F9903.

CHAMBAZ, C., et LEGENDRE, N. (1999). Calcul des pondérations dans le panel européen de ménages. *Actes des journées de méthodologie statistique, INSEE Méthodes*, 84-86.

DEVILLE, J. C. (1999). Les enquêtes par panel : en quoi diffèrent-elles des autres enquêtes? suivi de : comment attraper une population en se servant d'une autre. *Actes des journées de méthodologie statistique, INSEE Méthodes*, 84-86.

FIRDION, J. M., et MARPSAT, M. (1997). Comptes rendus du groupe « pondérations » de l'enquête auprès des personnes sans-domicile, mimeo.

FRANCK, O., et SNIJDERS, T. (1994). Estimating the Size of Hidden Populations Using Snowball Sampling. *Journal of Official Statistics*, 10, 53-67.

LAVALLÉE, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27 -35.

POLLOCK, K.H., TURNER, S.C. et BROWN, C.A. (1994). Techniques de saisie-ressaisie pour l'estimation de la taille de la population et de totaux de population lorsqu'on ne dispose pas d'une base de sondage complète. *Techniques d'enquête*, 20, 121 -128.

RTI (1993). Prevalence of Drug Use in the Washington DC Metropolitan Area, Homeless and Transient Population : 1991. *Technical report*, 2.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 16, Number 4, 2000

Borrowing Strength When Explicit Data Pooling Is Prohibited
Jerome P. Reiter 295

Attrition and Misclassification of Drop-outs in the Analysis of Unemployment Duration
Johan Bring and Kenneth Carling 321

Income Measurement Error in Surveys: A Review
Jeffrey C. Moore, Linda L. Stinson, and Edward J. Welniak, Jr. 331

On Variance Estimation for Measures of Change When Samples Are Coordinated by the Use of Permanent Random Numbers
Lennart Nordberg 363

A Functional Form Approach to Calibration
Victor M. Estevao and Carl-Erik Särndal 379

An Assessment of the Current State of Dependent Interviewing in Household Surveys
Nancy A. Mathiowetz and Katherine A. McGonagle 401

Remembering Heads and Bushels: Cognitive Processes Involved in Agricultural Establishments' Reports of Inventories
Jaki Stanley McCarthy and Martin A. Safer 419

Accuracy of Using Pneumonia as an Underlying Cause in the Cause-of-Death Register
Boo Svartbo, Linda Nilsson, Anders Eriksson, Gösta Bucht, Lars Age Johansson, and Lars Olov Bygren 435

Index to Volume 16, 2000 445

Volume 17, Number 1, 2001

Can A Statistician Deliver?
Richard Platek and Carl-Erik Särndal 1

Comment
Barbara A. Bailar 21
Paul P. Biemer 25
Alain Desrosières, Jean-Claude Deville, and Olivier Sautory 33
Eva Elvers and Lennart Nordberg 39
Ivan P. Fellegi 43
Robert M. Groves and Nancy A. Mathiowetz 51
David Holt 55
Margarida Madaleno 63
David A. Marker and David R. Morganstein 71
Pilar Martín-Guzmán 73
Photis Nanopoulos 77
Svein Nordbotten 87
Erkki Pahkinen 93
Jaki Stanley McCarthy 99
Lynne Stokes 103
Dennis Trewin 107

Rejoinder
Richard Platek and Carl-Erik Särndal 113

Enumeration Accuracy in a Population Census: An Evaluation Using Latent Class Analysis
Paul P. Biemer, Henry Woltman, David Raglin, and Joan Hill 129

Estimation of Interviewer Effects on Multivariate Binary Responses in a Community Based Survey
Sujuan Gao 149

Swedish Employment in the 1950s – How to Fill the Lacuna
Gudmundur Gunnarsson and Thomas Lindh 163

Practicing What We Preach: The Application of Continuous Improvement in a Preclinical Statistics Department at a Pharmaceutical Company
V. Bill Pikounis, Joseph M. Antonello, Danielle Moore, Edith T. Senderak, and Keith A. Soper 187

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

The Canadian Journal of Statistics **La Revue Canadienne de Statistique**

CONTENTS TABLE DES MATIÈRES

Volume 29, No. 1, March/mars 2001, 1-172

Richard A. LOCKHART
Message from the new Editor/Un message du nouveau rédacteur en chef 1

Christian GENEST:
Report from the former Editor/Rapport du rédacteur en chef sortant 3

Malay GHOSH and Yeong-Hwa KIM
The Behrensÿ Fisher problem revisited: a Bayes-frequentist synthesis 5

Daniel R. ENO and Keying YE
Probability matching priors for an extended statistical calibration model 19

Patrick E. BROWN and Piet de JONG
Nonparametric smoothing using state space techniques 37

Kert VIELE
Evaluating fit in functional data analysis using model embeddings 51

Thomas J. DICICCIO, Michael A. MARTIN and Steven E. STERN
Simple and accurate one-sided inference from signed roots of likelihood ratios 67

Daniel B. HALL
On the application of extended quasi-likelihood to the clustered data case 77

Paul KABAILA and John BYRNE
Exact short Poisson confidence intervals 99

Vincent F. MELFI, Connie PAGE and Margarida GERALDES
An adaptive randomized design with application to estimation 107

Giseon HEO, Byron SCHMULAND and Douglas P. WIENS
Restricted minimax robust designs for misspecified regression models 117

Min TSAO and Julie ZHOU
On the robustness of empirical likelihood ratio confidence intervals for location 129

Marc HALLIN, Amal MELLOUK and Khalid RIFI
Projection de Hájek et polynômes de Bernstein 141

Faouzi EL BANTLI and Marc HALLIN
Asymptotic behaviour of M-estimators in AR(p) models under nonstandard conditions 155

Acknowledgement of referees' services/Remerciements aux membres des jurys 169

Corrigenda and addendum 170

Forthcoming Papers/Articles à paraître 171

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points ci-dessous. Les articles acceptés doivent être soumis sous forme de fichiers de traitement de texte, préférablement WordPerfect. Les autres logiciels sont acceptables, mais une version sur papier sera alors exigée pour le traitement des formules et des figures.

1. Présentation

1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
1.4 Les remerciements doivent paraître à la fin du texte.
1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.
3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O, 0; l, 1).
3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).
5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.