TECHNIQUES D'ENQUÊTE

Nº 12-001-XPB au catalogue

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

DÉCEMBRE 2000

•

VOLUME 26

•

NUMÉRO 2

Statistique Canada Statistics Canada

Canada

TECHNIQUES D'ENQUÊTE

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

DÉCEMBRE 2000 • VOLUME 26 • NUMÉRO 2

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.

Février 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
F. Mayda (Directeur de la Production)
C. Patrick
R. Platek (Ancien président)
E. Rancourt
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*
P. Biemer, *Research Triangle Institute*
D.A. Binder, *Statistique Canada*
C. Clark, *U.S. Bureau of the Census*
J.-C. Deville, *INSEE*
J. Eltinge, *Texas A&M University*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
M.A. Hidiroglou, *Statistique Canada*
D. Holt, *Central Statistical Office, 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, *Central Bureau of Statistics, 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, W. Yung et D. Stukel, *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.

LESLIE KISH
(1910 - 2000)

Ce numéro est dédié à la mémoire de Leslie Kish. Sa joie de vivre contagieuse, sa préoccupation profonde pour les opprimés et les démunis et ses contributions importantes à la méthodologie d'enquête et aux statistiques ont été, et demeurent, une source d'inspiration pour beaucoup.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Volume 26, numéro 2, décembre 2000

TABLE DES MATIÈRES

Dans ce numéro 131

I.P. FELLEGI
Leslie Kish – Une vie de dévouement 133

D.E. HAINES, K.H. POLLOCK et S.G. PANTULA
Estimation de la taille et des chiffres de population pour des échantillons tirés de listes incomplètes avec probabilités d'inclusion hétérogènes 135

J.-F. BEAUMONT
Une méthode d'estimation en présence de non-réponse non-ignorable 145

B.D. SPENCER
Un effet de plan de sondage approximatif pour une pondération inégale en cas de corrélation possible entre les mesures et les probabilités de sélection 153

P.P. BIEMER et J.M. BUSHERY
Validité de l'analyse markovienne de structure latente pour l'estimation de l'erreur de classification des données sur la population active 157

K.J. THOMPSON et R.S. SIGMAN
L'estimation et l'estimation de la variance par répliques des prix de vente médians des maisons vendues 173

C.H. McLAREN et D.G. STEEL
L'effet de divers plans de renouvellement sur la variance d'échantillonnage des estimations désaisonnalisées et des estimations de la tendance 185

Y. YOU et J.N.K. RAO
Estimation bayésienne hiérarchique des moyennes pour petites régions à l'aide de modèles à plusieurs niveaux 197

F.C. OKAFOR et H. LEE
Échantillonnage à deux phases pour l'estimation par quotient ou par régression avec sous-échantillonnage des non-répondants 207

J. PICKERY et G. LOOSVELDT
Modélisation des effets d'intervieweur dans le cas des enquêtes par panel: Une application 213

M. FUCHS
La conception des écrans et l'ordre des questions dans un module IAO. Résultats d'une expérience de facilité d'utilisation menée sur le terrain 225

Remerciements 235

Techniques d'enquête, décembre 2000
Vol. 26, n° 1, pp. 131-132
Statistique Canada

Dans ce numéro

Le présent numéro est dédié à Leslie Kish, qui est décédé cet automne à l'âge de 90 ans. Chose remarquable, jusqu'à la fin de sa vie le professeur Kish a continué de présenter de nouvelles idées statistiques et méthodologiques, comme en témoigne son article sur le cumul et la combinaison des enquêtes démographiques publié il y a un an dans le numéro marquant le 25e anniversaire de *Techniques d'enquête*. On trouvera au début du présent numéro une réflexion d'Ivan Fellegi sur la vie et les réalisations statistiques du professeur Kish.

L'article de Haines, Pollock et Pantula examine les estimations d'un total lorsque l'on combine deux listes incomplètes et une base aréolaire. Les auteurs proposent des chiffres de population appropriés pouvant rendre compte du caractère incomplet des bases de sondage. De plus, leurs modèles tiennent compte du fait que les listes incomplètes ont de meilleures chances d'inclure des unités d'échantillonnage plus grandes.

Beaumont propose une méthode d'estimation permettant de réduire le biais dû à un mécanisme de réponse qui dépend de la variable d'intérêt, qu'on appelle un mécanisme de réponse non-ignorable. La méthode proposée requiert un modèle pour la variable d'intérêt et un modèle pour la probabilité de répondre. La méthode est considérée robuste par rapport à l'hypothèse de normalité puisqu'elle est construite de telle sorte qu'elle n'exige pas de spécifier la distribution des erreurs du modèle impliquant la variable d'intérêt, ce qui n'est pas le cas de la méthode du maximum de vraisemblance. Il propose également une méthode simple permettant de vérifier la validité de l'hypothèse de normalité des erreurs quand la non-réponse n'est pas ignorable.

Spencer aborde le problème de l'estimation des effets de plan de sondage attribuables à la pondération lorsqu'il existe une corrélation entre les probabilités de sélection et la variable d'intérêt. À l'aide d'une représentation de la population par régression, Spencer présente une approximation de l'effet de plan de sondage lorsqu'il existe une corrélation entre les probabilités de sélection et la variable d'intérêt.

Biemer et Bushery ont recours à l'hypothèse de Markov sur les changements de situation vis-à-vis de l'activité afin de cerner les erreurs de classification dans des données sur la population active. À l'aide de cette méthode, ils estiment les taux d'erreur de réponse pour des panels fournissant des données mensuelles sur la population active dans le cadre de la Current Population Survey (CPS). La cohérence globale des résultats est considérée comme un indicateur de l'utilité de l'analyse markovienne de la structure latente comme moyen d'évaluer l'exactitude des réponses à la CPS. Un aspect critique de cette analyse est la confirmation de l'hypothèse de Markov; les auteurs présentent des données empiriques intéressantes sur sa validité à court terme relativement à la CPS.

De nombreux bureaux de la statistique font appel à la méthode de répétition MHS (demi-échantillon modifié) pour l'estimation de la variance d'échantillonnage des médianes. C'est là un problème concret important puisque le calcul direct des médianes d'échantillon exige souvent une énorme capacité de calcul. Une autre méthode d'estimation consiste à grouper les données continues en intervalles discrets et à utiliser une interpolation linéaire pour l'intervalle comportant la médiane. Dans leur article, Thompson et Sigman comparent les effets d'une absence de groupement (c'est-à-dire la médiane de l'échantillon), d'un groupement avec intervalles de taille fixe et d'un groupement avec des intervalles dont la taille dépend des données, sur les médianes et les estimations connexes de la variance MHS. Leur étude empirique indique que les intervalles dont la taille dépend des données fournissent les estimations de la variance qui comportent le moins de biais, la meilleure stabilité et les meilleurs intervalles de confiance.

McLaren et Steel examinent les répercussions de différents modèles de chevauchement sur la variance d'échantillonnage des estimations désaisonnalisées et des estimations de tendance obtenues de séries chronologiques fondées sur des sondages lorsqu'on utilise les méthodes de désaisonnalisation X-11 et X-11-ARIMA du Recensement. Les profils de renouvellement «dans pour 8», «dans pour 6», «dans pour 4, hors pour 4, dans pour 4» sont raisonnables si la statistique clé à analyser est le changement mensuel des estimations désaisonnalisées. Toutefois, si les statistiques clés sont le niveau tendanciel et la différence entre deux estimations de tendances consécutives, le profil de renouvellement «dans pour 1, hors pour 2, dans pour 1, pour un total de 8 mois» est préférable lorsque l'on veut réduire la variance de l'échantillonnage. Les auteurs indiquent également que le profil «dans pour 2, hors pour 2, dans pour 2, pour un total de 8 mois» est un compromis raisonnable si le niveau et le changement mensuel des estimations de tendance et des estimations désaisonnalisées sont tous deux considérés comme importants.

You et Rao présentent des modèles hiérarchiques de Bayes à plusieurs niveaux pour les estimations régionales. Les modèles admettent des paramètres de régression aléatoires qui dépendent également de covariables régionales. La moyenne régionale est estimée à l'aide de la moyenne a posteriori, et la variance a posteriori est prise comme mesure de la précision. Trois modèles de variance sont examinés: fixe égal, fixe inégal, aléatoire. Les auteurs présentent des détails sur l'échantillonnage de Gibbs pour ces modèles et s'en servent comme outil d'inférence. Ils illustrent les procédures à l'aide de données sur le revenu des ménages au Brésil au niveau des comtés.

Okafor et Lee considèrent un schéma d'échantillonnage double, comportant un sous-échantillon de non-répondants qui sont interrogés de nouveau au cours de la deuxième phase en fonction d'un taux d'échantillonnage fixe. Suivant ce schéma, les auteurs proposent des versions modifiées des estimateurs par quotient et par régression. Ils déterminent des valeurs optimales pour la taille des échantillons et le taux d'échantillonnage fixe, d'après des fonctions de coûts, de façon à minimiser la variance. De plus, ils présentent les variances et leurs estimateurs. Une petite étude empirique permet d'étudier l'efficacité relative des estimateurs modifiés par quotient et par régression en fonction de l'estimateur standard de Hansen-Hurwitz.

Pickery et Loosveldt intègrent une importante technique analytique à l'étude de la non-réponse partielle. Leurs modèles offrent un aperçu plus complet des facteurs qui influencent la non-réponse partielle que ne le font les travaux antérieurs. Un aspect important de la stratégie est que les auteurs établissent une distinction entre la variation propre à l'intervieweur ou au répondant, la variation attribuable à des caractéristiques de l'intervieweur ou du répondant et la variance de l'erreur.

Fuchs examine l'influence que la conception de l'écran et l'ordre des questions exercent sur le comportement des intervieweurs dans le cadre d'une interview assistée par ordinateur (IAO). À l'aide d'expériences de laboratoire, on a pu montrer que la conception de l'écran et l'ordre des questions influencent bel et bien le comportement des intervieweurs. Dans son exposé, Fuchs présente les résultats d'une expérience sur le terrain dans laquelle on a analysé deux conceptions de l'écran ainsi que deux ordres des questions dans un plan factoriel 2x2. Les résultats se fondent sur des mesures du temps intégrées à l'application ITAO et sur 234 interviews choisies au hasard et enregistrées sur bande magnétique puis analysées en fonction d'un schéma de codage.

M.P. Singh

Techniques d'enquête, décembre 2000
Vol. 26, n° 2, pp. 133-134
Statistique Canada

Leslie Kish – Une vie de dévouement

IVAN P. FELLEGI[1]

1. INTRODUCTION

Je ne peux croire qu'il me faut écrire un article à la mémoire de Leslie Kish. Il y a tout juste quelques mois, j'ai rédigé une courte allocution quelque peu humoristique à l'occasion de son 90e anniversaire de naissance. J'avais alors demandé en plaisantant pourquoi on faisait tant de cas d'un 90e anniversaire. Après tout, la Reine-Mère venait à peine de célébrer son centenaire. J'avais alors remarqué que *cela* méritait d'être souligné. Il avait ri de bon coeur avec, dans les yeux, ce pétillement bien connu qui lui était propre. J'étais de nouveau surpris de constater combien il aimait encore s'amuser, la force de son dynamisme, de sa perspicacité et, en fait, de sa jeunesse dans toutes les facettes de son comportement et ce, malgré une mobilité quelque peu réduite. Il m'avait parlé de l'arthroplastie partielle du genou qu'il devait subir et m'avait confié que, selon son médecin, il devait choisir entre l'opération ou la marchette pour se déplacer. Bien entendu, il n'était pas question pour lui d'opter pour la marchette: il lui fallait jouir d'une mobilité complète. Et cette mobilité, à 90 ans, ne signifiait pas simplement la capacité de se déplacer à la maison; il s'agissait plutôt de voyager partout dans le monde plusieurs fois par année. Leslie Kish est mort des suites de complications postopératoires, après avoir lutté plusieurs semaines avec le courage inébranlable qui le caractérisait.

À mes yeux, le trait le plus distinctif de cet homme fut son dévouement intarissable. L'un de ses derniers actes d'altruisme a été d'inspirer ses amis et collègues à établir le *Leslie Kish International Fellows Fund* dans le but d'aider les étudiants de pays en développement à obtenir une formation en échantillonnage de populations.

Leslie Kish est né en 1910 à Poprad, qui faisait alors partie de l'Empire austro-hongrois et est maintenant située en Slovaquie. Il racontait que, à différentes époques de l'histoire, Poprad a appartenu à cinq pays différents; c'est là un symbole éloquent d'une vie motivée par l'amour des citoyens de toutes les régions du monde. En 1925, ses parents décident d'émigrer aux États-Unis, à l'instar de centaines de milliers de Hongrois quittant leur pays. Comme l'a dit le célèbre poète hongrois, Attila Jozsef, «un million et demi de nos citoyens sont sortis du pays en titubant pour gagner l'Amérique». Peu après leur arrivée, le père de Leslie décède. La mère de Leslie et les quatre enfants doivent décider s'ils resteront aux États-Unis. C'est ce qu'ils choisissent de faire, mais cela signifie que les deux aînés, dont Leslie alors âgé de 16 ans, devront travailler pour subvenir aux besoins de la famille.

Leslie poursuit ses études en cours du soir. En 1937, il ne lui reste qu'un an d'études pour obtenir un diplôme du premier cycle. Mais cet homme de 27 ans est alors de nouveau prêt à se sacrifier pour contribuer à la création d'un monde meilleur. Il interrompt ses études pour joindre les Brigades internationales et participer à la lutte contre les fascistes en Espagne. Son amour de ce pays et des opprimés l'a habité le reste de ses jours.

À la fin de la Guerre civile espagnole en 1939, il revient aux États-Unis, termine ses études à l'Université de New York et obtient un diplôme en mathématiques. Il déménage à Washington, où il a la chance de s'associer à des groupes de pionniers, d'abord au Bureau du recensement, puis au ministère de l'Agriculture.

Il interrompt de nouveau sa carrière pour s'engager comme volontaire dans l'armée. En 1947, il joint enfin les rangs de l'Université du Michigan à Ann Arbor où, en collaboration avec un petit groupe enthousiaste, il participe à la fondation de l' *Institute for Social Research*. Il dira plus tard qu'il n'a jamais autant travaillé que durant ces premières années: il obtient une maîtrise et un doctorat tout en occupant un poste à temps plein sans compter le temps qu'il consacre à l'enseignement.

En statistique, il nous aura légué plusieurs ouvrages remarquables au nombre desquels figure *Survey Sampling*, une oeuvre novatrice qui s'impose non seulement comme une bible dans le domaine (c'est-à-dire, à l'instar de la Bible, comme une source élevée d'inspiration), mais aussi comme un outil courant de la pratique. En ce sens, le système statistique mondial s'appuie, dans une large mesure, sur les précieux trésors de sagesse pratique de l'ouvrage *Survey Sampling*. En 1988 (lorsque Leslie Kish est encore un jeune homme de 78 ans), paraît *Statistical Design for Research* qui intègre et structure une expérience statistique acquise durant une vie entière. Entre-temps, avant et après cette année-là, ses travaux sont ponctués d'une masse d'articles, d'exposés et d'allocutions. Leslie Kish, travaillant quelquefois avec d'autres, introduit le concept des effets du plan de sondage dans nos esprits et ces termes dans notre discours; il est parmi les premiers à explorer la question de l'inférence dans les échantillons complexes et met au point le concept novateur désigné aujourd'hui par le terme répétition compensée – *balanced repeated replication* (en collaboration avec Marty Frankel); il fait oeuvre de pionnier dans l'étude de l'erreur de réponse; il se fait l'apôtre des recensements et échantillons avec renouvellement; il lance le concept du choix contrôlé; il formule le concept des plans à objectifs multiples; il

[1] Ivan P. Fellegi, Statisticien en chef, Statistique Canada, 26ième étage, section A, Immeuble R.-H. Coats, Ottawa (Ontario), Canada, K1A 0T6.

s'intéresse, dès le début, aux estimations visant les petites régions; et j'en passe. Mais, sans vouloir déprécier l'importance de ces travaux, j'estime que certaines de ses autres contributions sont tout aussi fondamentales.

Il fut l'une des rares personnes dont les premiers travaux appliqués ont conféré à l'échantillonnage respect et admiration. En plus de compter au nombre des fondateurs de ce qui est devenu l'*Institute for Survey Research* à Ann Arbor, il enseigne à des générations de statisticiens, tant américains qu'étrangers, dans le cadre du légendaire programme d'été pour les statisticiens étrangers. Après avoir officiellement pris sa retraite, il poursuit ses activités par les cours qu'il donne dans le programme d'été, par le travail d'édition et de collaboration à l'une ou l'autre des chroniques de questions et de réponses du bulletin *The Survey Statistician* entrepris depuis des décennies ainsi que par une multitude de conférences et travaux de services de conseils. Dans les rencontres internationales, il m'arrivait de «tomber» sur ses anciens étudiants et amis du temps. Aujourd'hui, on ne «tombe» plus sur eux par hasard, ils sont littéralement omniprésents: je me demande combien de spécialistes étrangers de l'échantillonnage bien connus n'ont pas été à un moment donné des étudiants de Leslie Kish. Et je ne veux pas passer sous silence deux de ces nombreuses contributions qui me tiennent particulièrement à coeur: ses années de loyaux services à Statistique Canada en qualité de membre fondateur de notre Comité consultatif des méthodes statistiques et son allocution à titre de président de l'ASA en 1977 (laquelle a paru dans le *JASA* en mars 1978), l'allocution la plus marquante du président de l'ASA qu'il m'a été donné d'entendre.

Ses réalisations lui ont valu une reconnaissance à l'échelle internationale. Je me contenterai de souligner quelques-unes des douzaines de distinctions qu'il s'est méritées. Il s'est vu décerner un doctorat honorifique de l'Université de Bologne à l'occasion du 900e anniversaire de l'établissement, la médaille Samuel Wilks, soit la plus haute distinction de l'ASA, la chaire Henry Russell, soit la plus haute distinction de l'Université du Michigan, le titre de membre titulaire honoraire de l'IIS que je considère comme une forme de prix Nobel de la statistique et peut-être la reconnaissance la plus significative à ses yeux, sans compter la multitude de hautes distinctions conférées par la Hongrie (doctorat honorifique de la plus grande université de Budapest, titre de membre honoraire de l'Académie des sciences de la Hongrie et la Croix d'officier de l'Ordre du Mérite).

Outre ce qu'il nous lègue dans le domaine de la statistique, il nous laisse le phénomène «Leslie Kish, une force de la nature»: le combattant pendant la Guerre civile d'Espagne, le philosophe de tout ce qui est statistique, le défenseur des droits humains qui n'aura jamais vieilli, le conteur, le lecteur insatiable, l'auteur des plus belles lettres de Noël, l'époux et le père affectueux et l'ami de toujours de centaines et peut-être de milliers de personnes.

Lorsque j'ai pris la parole à son 90e anniversaire de naissance, j'ai terminé en disant espérer être présent au véritable anniversaire de Leslie, celui que la Reine-Mère venait de célébrer. Et je ne plaisantais pas. Il était si plein de vie qu'il semblait non seulement possible de le voir atteindre le centenaire, mais en fait il paraissait même impossible d'imaginer autre chose. Malheureusement, il s'est éteint. Son dernier acte de générosité aura consisté à offrir son corps à la recherche médicale. Ne serait-il pas pertinent que les travaux qui en résultent puissent nous permettre de mieux comprendre cet être exceptionnel que fut Leslie Kish?

Techniques d'enquête, décembre 2000
Vol. 26, n° 2, pp. 135-144
Statistique Canada

Estimation de la taille et des chiffres de population pour des échantillons tirés de listes incomplètes avec probabilités d'inclusion hétérogènes

DAWN E. HAINES, KENNETH H. POLLOCK et SASTRY G. PANTULA[1]

RÉSUMÉ

Les informations tirées de bases de sondage aréolaires et de listes sont combinées de façon à fournir des estimations efficaces de la taille et des chiffres de population. Les auteurs examinent le cas où les probabilités d'inclusion dans les listes sont hétérogènes et modélisées en fonction de covariables. Ils adaptent et modifient la méthode employée par Huggins (1989) et par Alho (1990) pour la modélisation de variables auxiliaires dans des études de type saisir-ressaisir faisant appel à un modèle de régression logistique. Les auteurs présentent les résultats d'une étude de simulation qui permet de comparer divers estimateurs de la taille des bases de sondage et des chiffres de population en ayant recours à la stratégie de régression logistique pour modéliser des probabilités d'inclusion hétérogènes.

MOTS CLÉS: Régression logistique; base de sondage sous forme de liste; base de sondage aréolaire; échantillonnage de type saisir-ressaisir.

1. INTRODUCTION

Dans le présent exposé, les auteurs estiment la taille de la population et les chiffres de population en présence d'informations tirées de plusieurs bases de sondage indépendantes. Il est supposé que les éléments de la population comportent différentes probabilités d'inclusion pour diverses bases de sondage. Ces probabilités d'inclusion hétérogènes peuvent dépendre d'une covariable. Ainsi, supposons qu'il s'agisse d'estimer le nombre d'exploitations porcines et le nombre de cochons en Caroline du Nord. Des mesures de covariable comme la superficie des explotations porcines ou le nombre d'employés sont une indication de la taille des exploitations porcines. Il se peut que de grandes exploitations aient de meilleures chances d'être incluses dans une liste servant de base de sondage que des exploitations plus petites. Dans des expériences de type saisir-ressaisir, il se peut que les probabilités de saisie soient inégales pour les animaux. Les probabilités de saisie (inclusion) pour les animaux peuvent varier selon l'âge, le sexe, la taille ou l'espèce.

On entend par listes des bases de sondage qui énumèrent des unités d'échantillonnage de la population cible. Les éléments qu'on y trouve peuvent comprendre, entre autres, des noms, des adresses, des numéros de téléphone, des numéros de sécurité sociale ou des descriptions matérielles d'emplacement. Ces variables et d'autres variables de stratification diverses servent à identifier des personnes, des animaux, des entreprises ou d'autres établissements. Les listes et les bases aréolaires servent à fournir des estimations de taille et de chiffres de population inconnus. Puisque, dans toute importante opération de collecte de données, les bases de sondage comportent inévitablement des imperfections comme des omissions, des dédoublements et des enregistrements inexacts (Hansen, Hurwitz et Madow 1953), on trouve dans la littérature diverses propositions de solution pour surmonter cette difficulté. Une stratégie, élaborée d'abord par Hartley (1962, 1974), combine une liste incomplète et une base aréolaire. D'autres développements théoriques sont dus à Cochran (1965), à Lund (1968), à Fuller et Burmeister (1972) et à Bosecker et Ford (1976). Haines et Pollock (1998a) ont appliqué la méthode des deux bases de sondage à une population de pigargues à tête blanche, tandis que Haines et Pollock (1998b) ont présenté une stratégie théorique plus générale pour ce qui est de la combinaison de plusieurs bases de sondage. Ces deux exposés n'ont pas examiné le cas des probabilités d'inclusion hétérogènes. Fienberg (1992) a présenté une bibliographie annotée de la documentation sur les démarches de type saisir-ressaisir traitant plus particulièrement du problème de sous-dénombrement du recensement, qui comprend les exposés de Wolter (1986, 1990) et de Cowan et Malec (1986).

Le NASS (*National Agricultural Statistics Service*) utilise actuellement plusieurs bases de sondage pour ses démarches d'échantillonnage et d'estimation de nombreux produits agricoles. Le NASS recueille et résume des données sur les superficies ensemencées, le bétail, la production et les stocks de céréales, les coûts de production, les dépenses agricoles et d'autres caractéristiques agricoles. Fecso, Tortora et Vogel (1986) ont passé en revue les bases de sondage utilisées pour le secteur agricole des États-Unis, tandis que Nealon (1984) a fourni des détails sur les estimateurs liés à des bases de sondage multiples et aréolaires qui sont utilisés par le Department of Agriculture des États-Unis. Pollock, Turner et Brown (1994) ont présenté

[1] Dawn E. Haines, U. S. Bureau of the Census, Washington, DC 20233; Kenneth H. Pollock et Sastry G. Pantula, North Carolina State University, Department of Statistics, Box 8203, Raleigh, NC 27695-8203, U.S.A.

une solution modélisée de type saisir-ressaisir en vue de l'estimation de la taille des bases de sondage en fonction d'informations tirées de deux listes incomplètes. D'après Cochran (1977), il est souvent difficile d'obtenir une liste qui correspond exactement à la population étudiée. Les listes que l'on établit habituellement pour un but quelconque sont souvent incomplètes et en partie illisibles, ou comportent une part inconnue de dédoublement. Puisque les listes sont en règle générale incomplètes, les estimations qui se fondent uniquement sur des listes risquent de sous-estimer la taille de la population. En complétant les renseignements disponibles à l'aide d'un échantillon tiré d'une base aréolaire, on pourra peut-être établir des estimations efficaces de la taille de la population et des chiffres de population.

Une base aréolaire est une collection de régions géographiques définies par des limites identifiables. Les chargés d'enquête utilisent souvent des bases aréolaires afin d'obtenir une couverture complète de la population cible. Des populations comme les exploitations agricoles sont naturellement associées aux subdivisions de terrain constituant une base aréolaire. Ainsi, dans une enquête agricole, la région d'intérêt est divisée en une série de régions géographiques appelées des segments. Les segments, qui sont les unités d'échantillonnage, sont alors sélectionnés à l'aide de plans de sondage à plusieurs degrés stratifiés (Kott et Vogel 1995). On définit des règles qui établissent les liens entre les exploitations agricoles de la population et les segments de la base aréolaire. Lorsque les exploitations agricoles, ou les unités déclarantes, ont été identifiées au sein de chaque segment échantillonné, on les dénombre en personne et on recueille les données pertinentes. Nealon (1984) a donné une description détaillée des estimateurs de segments ouverts, fermés et pondérés. Faulkenberry et Garoui (1991) ont formulé des estimateurs supplémentaires conçus précisément pour des bases aréolaires. On trouvera dans Fecso et coll. (1986) des méthodes plus complexes de construction et d'échantillonnage pour les bases aréolaires. L'échantillonnage et le sous-échantillonnage à partir de bases aréolaires sont examinés en détail dans Kott et Vogel (1995).

À la section 2, les auteurs examinent des listes indépendantes dont les éléments comportent des probabilités d'inclusion hétéroscédastiques. Il y est question de méthodes qui fournissent des estimateurs de la taille de la population et des chiffres de population en présence de renseignements tirés d'une ou plusieurs listes et d'une base aréolaire. La section 3 résume les résultats d'une étude de simulation qui a permis de comparer divers estimateurs de la taille de la base (population) et des chiffres de population. Enfin, les auteurs présentent des conclusions et un sommaire.

2. PROBABILITÉS D'INCLUSION HÉTÉROSCÉDASTIQUES

2.1 Estimation de la taille de la population à l'aide de listes

Dans des expériences de type saisir-ressaisir, il se peut que différents animaux comportent différentes probabilités de saisie. De même, des éléments individuels peuvent comporter différentes probabilités d'inclusion dans une base de sondage sous forme de liste. Des listes différentes peuvent être considérées comme des cycles de saisie différents. Le modèle M_h désigne le modèle d'hétérogénéité dans la documentation sur les démarches de type saisir-ressaisir pour une population fermée (Otis, Burnham, White et Anderson 1978). Dans un contexte de type saisir-ressaisir, les probabilités de saisie, bien que pouvant varier d'un animal à l'autre, sont supposées les mêmes pour tous les cycles de piégeage. Le modèle d'hétérogénéité peut comporter jusqu'à $N + 1$ paramètres au total, c'est-à-dire N et $p_i, i = 1, ..., N$, où N est la taille de la population et p_i désigne la probabilité d'inclusion pour la i-ième unité. Pour des listes multiples, cela correspond à l'hypothèse suivant laquelle la probabilité d'inclusion p_i pour l'élément i est constante pour l'ensemble des k listes, $B_1, B_2, ..., B_k$.

Burnham (1972) et Burnham et Overton (1978, 1979) ont examiné le problème de l'estimation de N dans un contexte de type saisir-ressaisir. L'estimateur proposé pour N que décrit Burnham (1972) se fonde sur la méthode jackknife de réduction du biais (Quenouille 1956). Chao (1988) a élaboré un autre estimateur de moment pour ce modèle en fonction de données sur la fréquence de saisie (Pollock 1991). Dans certaines circonstances, l'estimateur proposé par Chao est moins biaisé que l'estimateur jackknife de Burnham. En général, il est difficile de trouver un estimateur tout à fait satisfaisant de N dans le cadre du modèle M_h. Otis et coll. (1978) ont donc suggéré que l'on conçoive l'étude dans son ensemble de façon à minimiser l'hétérogénéité. Norris et Pollock (1996) ont proposé un estimateur du maximum de vraisemblance non paramétrique qui n'est pas tout à fait satisfaisant.

Dans des expériences de type saisir-ressaisir, le modèle exprimé sous forme du M-ième modèle admet des probabilités d'inclusion pouvant varier tant selon le cycle de piégeage (base sous forme de liste) que selon l'individu. Nous définissons p_{ij} comme la probabilité d'inclusion du i-ième élément de la j-ième liste. Le M-ième modèle n'est évidemment pas facile à estimer puisqu'il peut comporter jusqu'à $tN + 1$ paramètres, où $t = k$, le nombre de listes. Chao, Lee et Jeng (1992), employant l'idée de couverture de l'échantillon, ont proposé une méthode non paramétrique d'estimation de la taille de la population pour le M-ième modèle.

À part la stratégie non paramétrique, une autre façon de procéder est de modéliser les probabilités d'inclusion en fonction d'une variable auxiliaire. Pollock, Hines et Nichols (1984), Huggins (1989) et Alho (1990) ont examiné le rôle des variables auxiliaires dans des expériences de type saisir-ressaisir avec probabilités inégales de saisie (inclusion). Les expériences de type saisir-ressaisir pour des populations fermées comportent $i = 1, ..., N$ individus et $j = 1, ..., t$ cycles de piégeage. Encore une fois, les $j = 1, ..., t$ cycles de piégeage sont semblables à $t = k$, le nombre de listes indépendantes. Huggins (1989) et Alho (1990) ont proposé une procédure d'estimation conditionnelle en vue de l'estimation de la taille d'une population fermée se fondant sur une saisie et une seule ressaisie. Dans ces deux exposés, les auteurs ont adopté le modèle logistique pour les probabilités d'inclusion, données par

$$p_{ij} = \frac{\exp(\alpha_j + \beta_j x_i)}{1 + \exp(\alpha_j + \beta_j x_i)}, \qquad (1)$$

où x_i est une covariable et α_j et β_j sont des paramètres inconnus. À noter que cette paramétrisation donne $0 \le p_{ij} \le 1$ pour toutes les valeurs de α_j et β_j. Pour $\beta_j > 0$, la probabilité d'inclusion augmente en fonction de la covariable. Cette paramétrisation est différente de l'échantillonnage avec probabilité proportionnelle à la taille (ppt), où p_{ij} est supposée proportionnelle à x_i. Les estimateurs du maximum de vraisemblance de α_j et β_j sont obtenus en fonction de la vraisemblance conditionnelle que l'unité se retrouve dans une liste au moins. Haines (1997) a dérivé la fonction de vraisemblance conditionnelle pour trois listes indépendantes.

En traitant chaque individu comme une strate distincte, définissons les variables indicatrices ci-dessous pour $i = 1, ..., N$:

$$u_{ij} = \begin{cases} 1 & \text{l'individu } i \text{ est compris dans la base de sondage } j \text{ seulement} \\ 0 & \text{autrement} \end{cases}$$

et

$$j = B_1, B_2,$$

$$a_i = \begin{cases} 1 & \text{l'individu } i \text{ est compris dans les deux bases de sondage} \\ 0 & \text{autrement.} \end{cases}$$

La valeur de l'expression

$$M_i = u_{iB_1} + u_{iB_2} + a_i \qquad (2)$$

est 1 si l'individu i est compris dans au moins une des deux bases de sondage et 0 autrement.

Alho (1990) a présenté la fonction de vraisemblance conditionnelle pour deux listes comme suit:

$$\frac{\exp\{\alpha_{B_1} N_{B_1} + \alpha_{B_2} N_{B_2} + \beta_{B_1} \sum_{i \in B_1} x_i + \beta_{B_2} \sum_{i \in B_2} x_i\}}{\prod_{M_i = 1} K_i(\theta)}, \qquad (3)$$

où

$$K_i(\theta) = \exp\{\alpha_{B_1} + \beta_{B_1} x_i\} + \exp\{\alpha_{B_2} + \beta_{B_2} x_i\}$$

$$+ \exp\{\alpha_{B_1} + \alpha_{B_2} + (\beta_{B_1} + \beta_{B_2}) x_i\}$$

et $\theta = (\alpha_{B_1}, \beta_{B_1}, \alpha_{B_2}, \beta_{B_2})'$. Alho (1990) a utilisé une procédure itérative fondée sur les statistiques suffisantes pour maximiser (3) et nous appliquons la méthode de Newton pour calculer des estimateurs du maximum de vraisemblance conditionnelle pour θ, notés $\hat{\theta} = (\hat{\alpha}_{B_1}, \hat{\beta}_{B_1}, \hat{\alpha}_{B_2}, \hat{\beta}_{B_2})'$. On trouvera à l'annexe A dans Haines (1997) des détails sur la méthode de Newton. La probabilité estimée que l'individu i soit compris dans au moins une des listes est notée

$$\hat{\pi}_i = 1 - \left(\frac{1}{1 + \exp(\hat{\alpha}_{B_1} + \hat{\beta}_{B_1} x_i)} \right)$$

$$\times \left(\frac{1}{1 + \exp(\hat{\alpha}_{B_2} + \hat{\beta}_{B_2} x_i)} \right) = \pi_i(\hat{\theta}), \qquad (4)$$

où

$$\hat{p}_{ij} = \frac{\exp(\hat{\alpha}_j + \hat{\beta}_j x_i)}{1 + \exp(\hat{\alpha}_j + \hat{\beta}_j x_i)},$$

$$i = 1, ..., N \text{ et } j = B_1, B_2. \qquad (5)$$

Si θ était connu, l'estimateur de Horvitz-Thompson de N serait $\hat{N} = \sum_{M_i = 1} 1/\pi_i$ (Horvitz et Thompson 1952). D'après Cochran (1977), la variance de $\hat{N}$ est

$$V(\hat{N}) = \sum_{i=1}^{N} \frac{1 - \pi_i}{\pi_i}. \qquad (6)$$

Une estimation de la variance de $\hat{N}$ est

$$\hat{V}(\hat{N}) = \sum_{M_i = 1} \frac{1 - \pi_i}{\pi_i^2}.$$

Puisque θ est inconnu, nous considérons l'estimation de la taille de la population donnée par $\hat{N} = \sum_{M_i = 1} 1/\hat{\pi}_i$, où $\hat{\pi}_i$ est définie en (4). Une estimation de la variance de $\hat{N}$ est obtenue à l'aide de la méthode de Taylor et se présente sous la forme

$$\hat{V}(\hat{N}) = \sum_{M_i = 1} \frac{1 - \pi_i(\hat{\theta})}{\pi_i^2(\hat{\theta})} + \hat{A}\hat{\sum}(\hat{\theta})\hat{A}',$$

où (7)

$$\hat{A} = \sum_{M_i = 1} \left[\frac{1}{\pi_i^2(\hat{\theta})} \frac{\partial \pi_i(\hat{\theta})}{\partial \hat{\theta}'} \right]$$

et $\hat{\sum}(\hat{\theta})$ est l'inverse de la matrice hessienne. Le second membre en (7) est attribuable à l'estimation de $\pi_i(\theta)$ par $\pi_i(\hat{\theta})$.

Un autre estimateur de la taille de la population qui est utilisé couramment dans des expériences de type saisir -ressaisir est l'estimateur de Lincoln-Petersen. Cet estimateur classique est attribuable à Lincoln (1930) et à Petersen (1896) et se présente comme suit:

$$\hat{N}_{L-P} = \frac{N_{B_1} N_{B_2}}{N_{b_1 b_2}},$$

où N_{B_1} et N_{B_2} désignent la taille des listes B_1 et B_2, respectivement, et $N_{b_1 b_2}$ désigne le nombre d'unités communes aux deux bases de sondage. Il s'agit d'une méthode simple d'estimation de moments fondée sur l'hypothèse voulant que toutes les unités comportent des probabilités d'inclusion homogènes pour chacune des deux listes indépendantes. Il se peut que le dénominateur $N_{b_1 b_2}$ soit nul. Chapman (1951) a proposé une version modifiée de l'estimateur de Lincoln-Petersen, donnée par

$$\hat{N}_{CH} = \frac{(N_{B_1} + 1)(N_{B_2} + 1)}{(N_{b_1 b_2} + 1)} - 1. \quad (8)$$

Cet estimateur est moins biaisé que l'estimateur de Lincoln-Petersen (Chapman 1951). D'après Sekar et Deming (1949), l'erreur-type asymptotique de $\hat{N}_{CH}$ est

$$\sqrt{\hat{V}(\hat{N}_{CH})} = \sqrt{\frac{N_{B_1} N_{B_2} N_{b_1} N_{b_2}}{(N_{b_1 b_2})^3}},$$

où N_{b_1} et N_{b_2} désignent le nombre d'unités appartenant uniquement aux listes B_1 et B_2, respectivement.

L'estimateur de Lincoln-Petersen est l'estimateur du maximum de vraisemblance inconditionnelle de la taille de la population N lorsqu'il y a deux listes indépendantes et que les probabilités d'inclusion sont homogènes. Haines (1997) étend les procédures d'estimation à k listes, chacune comportant des probabilités d'inclusion homogènes. Cet estimateur, toutefois, n'est pas approprié lorsque les probabilités d'inclusion sont hétérogènes. On trouvera à la section 3 les résultats de la simulation.

2.2 Estimation de la taille de la population à l'aide de listes et de bases aréolaires

Supposons que nous avons accès à une base aréolaire en plus des deux listes B_1 et B_2. La base aréolaire est constituée de U_A segments qui couvrent la population entière. Un échantillon aléatoire simple de u_A segments est sélectionné. Nous supposons que toutes les unités des segments échantillonnés sont observées. La probabilité d'inclusion dans l'échantillon de la base aréolaire est la même pour toutes les unités et représente la quantité connue $p_A = u_A/U_A$. Ensuite, nous maximisons la vraisemblance conditionnelle (3) relativement à θ et nous calculons la probabilité estimée qu'un individu i soit inclus dans au moins une liste ou dans la base aréolaire. Cette probabilité est notée $\tilde{\pi}_i = \hat{\pi}_i + p_A(1 - \hat{\pi}_i)$. Les probabilités $\hat{\pi}_i$ et $\hat{p}_{ij}$ sont définies en (4) et en (5), respectivement. Un estimateur de Horvitz-Thompson estimé pour la taille de la population est

$$\hat{\hat{N}} = \sum_{i \in \text{échantillon}} \frac{1}{\tilde{\pi}_i}. \quad (9)$$

Cet estimateur se laisse facilement étendre au cas comportant k listes, $B_1, \ldots, B_k$, et une base aréolaire indépendante.

D'après Cochran (1977), une estimation de la variance de $\hat{\hat{N}}$ est donnée par

$$\hat{V}(\hat{\hat{N}}) = \sum_{M_i = 1} \frac{1 - \tilde{\pi}_i}{\tilde{\pi}_i^2} + 2 \sum_{i} \sum_{l} \frac{(\tilde{\pi}_{il} - \tilde{\pi}_i \tilde{\pi}_l)}{\tilde{\pi}_{il} \tilde{\pi}_i \tilde{\pi}_l} + \hat{A} \hat{\sum} \hat{A}', \quad (10)$$

où $\hat{A}$ est défini en (7) et $\hat{\sum}$ est l'inverse de la matrice hessienne. La formule de variance pour $\hat{N}$ en (6) et son estimation sont valides uniquement lorsque π_{il}, la probabilité que les unités i et l soient incluses dans l'échantillon, est égale à $\pi_i \pi_l$. Lorsqu'on inclut un échantillon de base aréolaire, toutefois, π_{il} n'est pas nécessairement égale à $\pi_i \pi_l$. Supposons que les unités i et l appartiennent au même segment de base aréolaire. Dans un tel cas, les unités i et l sont toutes deux incluses ou non incluses dans l'échantillon, selon que leur segment correspondant est sélectionné ou non. Il est possible de montrer que la probabilité d'inclusion combinée, π_{il}, se laisse estimer comme suit:

$$\tilde{\pi}_{il} = \begin{cases} \tilde{\pi}_i \tilde{\pi}_l & \text{si } i \text{ et } l \text{ appartiennent à des segments aréolaires différents} \\ p_A + \hat{\pi}_i \hat{\pi}_l (1 - p_A) & \text{si } i \text{ et } l \text{ appartiennent au même segment aréolaire} \end{cases} \quad (11)$$

où $\hat{\pi}_i$ est définie en (4) et $\tilde{\pi}_i = \hat{\pi}_i + p_A(1 - \hat{\pi}_i)$. Ainsi, lorsque i et l appartiennent au même segment, $\tilde{\pi}_{il} \neq \tilde{\pi}_i \tilde{\pi}_l$. Toutefois, si p_A est petite et que $\hat{\pi}_i$ et $\hat{\pi}_l$ sont grandes, alors $(\tilde{\pi}_{il} - \tilde{\pi}_i \tilde{\pi}_l)$ se rapproche de zéro.

2.3 Estimation du total de la population à l'aide de listes

Supposons que des valeurs y_i sont disponibles pour tous les éléments de deux listes indépendantes B_1 et B_2. Pour θ connu, une estimation du total de la population, Y, est l'estimateur de Horvitz-Thompson

$$\hat{Y}_{H-T} = \sum_{M_i=1} \frac{y_i}{\pi_i(\theta)}. \quad (12)$$

D'après Cochran (1977), la variance estimée de $\hat{Y}_{H-T}$ est

$$\hat{V}(\hat{Y}_{H-T}) = \sum_{M_i=1} \frac{y_i^2(1-\pi_i(\theta))}{\pi_i^2(\theta)}.$$

Lorsque θ est inconnu et estimé par $\hat{\theta}$, une estimation du total de la population est

$$\hat{\hat{Y}}_{H-T} = \sum_{M_i=1} \frac{y_i}{\pi_i(\hat{\theta})}.$$

Une estimation de la variance de $\hat{\hat{Y}}_{H-T}$ est dérivée à l'aide de la méthode de Taylor et se présente sous la forme

$$\hat{V}(\hat{\hat{Y}}_{H-T}) = \sum_{M_i=1} \frac{y_i^2(1-\pi_i(\hat{\theta}))}{\pi_i^2(\hat{\theta})} + \hat{B}\hat{\Sigma}(\hat{\theta})\hat{B}',$$

où

$$\hat{B} = \sum_{M_i=1} \left[\frac{y_i}{\pi_i^2(\hat{\theta})} \frac{\partial \pi_i(\hat{\theta})}{\partial \hat{\theta}'} \right] \quad (13)$$

et $\hat{\Sigma}(\hat{\theta})$ est l'inverse de la matrice hessienne évaluée sur $\hat{\theta}$. Ces idées s'étendent facilement de façon à incorporer k listes indépendantes.

Dans la pratique, les y_i ne sont pas nécessairement observées pour toutes les unités des listes. Considérons le cas dans lequel les y_i sont disponibles pour un échantillon aléatoire seulement de n_{B_1} et n_{B_2} unités des listes B_1 et B_2, respectivement. De par leur construction, les probabilités d'inclusion, p_{ij}, varient selon l'individu i et la base j. Toutefois, une fois des individus inclus dans une liste, on en tire un sous-échantillon par échantillonnage aléatoire simple. Par conséquent, tous les éléments de la liste B_j ont des chances égales (n_{B_j}/N_{B_j}) d'être inclus dans le sous-échantillon. À noter que nous tirons des échantillons de chaque liste au lieu de tirer un seul échantillon d'une base de listes combinées. Puisque les listes sont supposées indépendantes, la probabilité estimée que le i-ième individu soit inclus dans au moins une des deux listes est

$$\hat{\pi}_i = \hat{p}_{iB_1}\frac{n_{B_1}}{N_{B_1}} + \hat{p}_{iB_2}\frac{n_{B_2}}{N_{B_2}} - \hat{p}_{iB_1}\hat{p}_{iB_2}\frac{n_{B_1}}{N_{B_1}}\frac{n_{B_2}}{N_{B_2}}. \quad (14)$$

On obtient une estimation de Horvitz-Thompson estimée pour Y en substituant (14) en (12).

Dans ce cas, un autre estimateur du total de la population, Y, est

$$\hat{Y} = \hat{N}_{CH} \frac{\sum_{M_i=1} y_i}{N_{b_1} + N_{b_2} + N_{b_1 b_2}},$$

qui est l'estimateur de Chapman multiplié par la moyenne des réponses pour les éléments inclus dans au moins un sous-échantillon de liste. Encore une fois, cet estimateur est valide uniquement lorsque les probabilités d'inclusion sont homogènes. Il existe $N_{b_1} + N_{b_2} + N_{b_1 b_2}$ éléments uniques dans les bases B_1 et B_2. Un estimateur semblable se laisse définir lorsque des renseignements sont disponibles uniquement pour des sous-échantillons des listes.

2.4 Estimation du total de la population à l'aide de bases aréolaires et de listes

Considérons le cas dans lequel, en plus de valeurs y_i pour les unités des listes (ou les sous-échantillons des listes), des valeurs y_i sont disponibles pour tous les éléments d'un échantillon aléatoire de segments d'une base aréolaire. L'inclusion de l'information de la base aréolaire donne lieu à la probabilité d'inclusion estimée pour le i-ième individu, c'est-à-dire

$$\tilde{\pi}_i = \hat{\pi}_i + p_A(1-\hat{\pi}_i), \quad (15)$$

où $\hat{\pi}_i$ est définie en (4) ou en (14), selon que y_i est observée pour toutes les unités des listes ou seulement pour un sous-échantillon d'unités, respectivement. Dans ce cas, un estimateur de Horvitz-Thompson estimé pour le total de la population est

$$\hat{\hat{Y}}_{H-T} = \sum_{i \in \text{échantillon}} \frac{y_i}{\tilde{\pi}_i}. \quad (16)$$

Une estimation de la variance de $\hat{\hat{Y}}_{H-T}$ est donnée par

$$\hat{V}(\hat{\hat{Y}}_{H-T}) = \sum_{M_i=1} \frac{y_i^2(1-\tilde{\pi}_i)}{\tilde{\pi}_i^2} + 2\sum_{i<l}\sum \frac{(\tilde{\pi}_{il} - \tilde{\pi}_i\tilde{\pi}_l)}{\tilde{\pi}_{il}\tilde{\pi}_i\tilde{\pi}_l} y_i y_l + \hat{B}\hat{\Sigma}\hat{B}', \quad (17)$$

où $\tilde{\pi}_{il}$ est définie en (11) et $\hat{B}$ et $\hat{\Sigma}$ sont définis en (13).

3. ÉTUDE DE SIMULATION

3.1 Hypothèses de l'étude

Afin d'étudier les propriétés de la taille de la population et des estimateurs du total, Haines (1997) a considéré quatre-vingt modèles différents. Nous présentons ici des détails pour deux de ces modèles seulement. Il est supposé que les probabilités d'inclusion pour deux listes dépendent

d'une covariable x_i. Deuxièmement, nous supposons que la covariable peut être corrélée avec la variable de réponse y_i. De plus, nous supposons que x_i et y_i suivent une distribution normale logarithmique avec corrélation ρ_{xy}. La distribution normale logarithmique est utilisée de façon à admettre une distribution asymétrique des covariables. Nous générons x_i comme e^{u_i} et y_i comme e^{v_i}, où u_i et v_i sont des variables aléatoires normales à deux dimensions avec moyennes nulles, des variances unitaires et une corrélation ρ_{uv}. Il est possible de montrer que $\rho_{uv} = \log[\rho_{xy}(e-1)+1]$.

Considérons une population de taille N. Supposons qu'il existe deux listes indépendantes, B_1 et B_2, et une base aréolaire, A. La base aréolaire est supposée complète en ce sens qu'elle couvre la population entière. On tire un échantillon de segments de la base aréolaire et on observe les unités au sein de chaque segment aréolaire. Soit p_A, la probabilité d'inclusion d'un élément quelconque dans l'échantillon de la base aréolaire, où p_A est supposée identique pour tous les individus.

La probabilité que le i-ième élément soit inclus dans la j-ième liste est donnée par le modèle de régression logistique (1) pour $i = 1, ..., N$ et $j = B_1, B_2$. Nous supposons que la probabilité que le i-ième élément soit inclus dans la liste B_1 est indépendante de son état d'inclusion dans la liste B_2 et dans l'échantillon de la base aréolaire.

3.2 Réglage des paramètres

Nous considérons diverses valeurs paramétriques. Pour la taille de la population, N, nous posons $N = 300$ ou 1 000. Nous utilisons $\rho_{xy} = -0{,}3, 0{,}0, 0{,}5$ et 1 correspondant à une corrélation négative, nulle, positive et parfaite entre la variable de réponse et la covariable. Ici, $\rho_{xy} = 1$ correspond à $x_i = y_i$, indiquant que la probabilité d'inclusion est liée directement à la variable de réponse.

Pour chacun des $2 \times 4 = 8$ réglages paramétriques ci-dessus de N et ρ_{xy}, nous considérons deux modèles correspondant à différents choix de $\alpha_{B_j}, \beta_{B_1}, \alpha_{B_2}$ et β_{B_2}. Rappelons que $E(x_i) = E[e^{u_i}] = e^{0,5}$. Considérons un élément à valeur de covariable donnée par la valeur moyenne $e^{0,5}$. La probabilité que cet élément soit inclus dans la j-ième liste est

$$p_j^{(E)} = \frac{\exp(\alpha_j + \beta_j e^{0,5})}{1 + \exp(\alpha_j + \beta_j e^{0,5})}, \quad j = B_1, B_2.$$

Si $\alpha_j = -\beta_j e^{0,5}$, cet élément a 50% de chances d'être inclus dans la liste j. Nous utilisons cette relation dans le modèle 1.

Nous étendons l'idée ci-dessus; si nous posons

$$\alpha_j = \log\left(\frac{p}{1-p}\right) - \beta_j e^{0,5},$$

alors l'unité à valeur de covariable moyenne comporte une probabilité p d'être incluse dans la liste j. Si nous supposons que les probabilités d'inclusion sont les mêmes pour les listes B_1 et B_2, alors les chances d'inclusion dans au moins une des deux listes est donnée par $1 - (1-p)^2$. Cette relation est utilisée dans le modèle 2. Les valeurs particulières de α_j et β_j pour les deux modèles sont résumées dans le tableau 1.

Tableau 1
Sommaire des paramètres de modèle

Modèle	α_{B_1}	β_{B_1}	α_{B_2}	β_{B_2}	$p_{B_1}^{(E)}$	$p_{B_2}^{(E)}$	$1-(1-p_{B_1}^{(E)}).(1-p_{B_2}^{(E)})$
1	0	0	0	0	0,5	0,5	0,75
2	-0,5478	0,8	-0,5478	0,8	0,6838	0,6838	0,90

Pour chacun des $2 \times 4 \times 2 = 16$ modèles, nous considérons trois valeurs p_A données par 0, 0,05 et 0,20. Ici, $p_A = 0$ correspond à l'utilisation des seuls renseignements des listes B_1 et B_2.

3.3 Génération des données

Pour chacun des seize modèles ci-dessus, nous générons d'abord (x_i, y_i) à l'aide de la distribution normale logarithmique à deux dimensions pour $i = 1, ..., N$. Nous «générons» (identifions) ensuite les unités qui appartiennent aux listes B_1 et B_2. Nous utilisons la probabilité p_{ij} pour inclure le i-ième élément dans la liste j. Enfin, à l'aide de $p_A = 0{,}05$, nous identifions les éléments qui appartiennent à la base aréolaire A. Nous répétons le processus pour le cas $p_A = 0{,}20$. Pour chaque combinaison paramétrique, nous générons 1 000 répétitions de Monte Carlo.

3.4 Estimateurs

Pour la taille de la population, nous considérons l'estimateur de Chapman, $\hat{N}_{CH}$, donné en (8). Cet estimateur suppose que $\beta_{B_1} = \beta_{B_2} = 0$ et n'utilise pas l'information de l'échantillon de la base aréolaire. Nous considérons également les estimateurs de Horvitz-Thompson estimés dont il est question à la section 2.

Pour l'estimation du total de la population d'une variable de réponse, nous considérons le cas dans lequel la réponse est observée pour tous les éléments de la liste. Les éléments d'une base aréolaire sont échantillonnés avec des probabilités $p_A = 0$, 0,05 et 0,20. Nous ne considérons pas d'estimations du total de la population fondées sur des sous-échantillons de chaque liste. L'estimation du total de la population, $\hat{Y}_{p_A}$, se présente sous la même forme que (16) avec $\hat{\pi}_i$ définie en (15). De même, l'estimation de la taille de la population, $\hat{N}_{p_A}$, se présente sous la même forme que (9).

L'estimateur

$$\hat{Y}_{\mathrm{CHs},p_A} = \hat{N}_{\mathrm{CH}} \bar{y}_{(p_A)}, \; p_A = 0, 0{,}05, 0{,}20$$

est également considéré, où $\bar{y}_{(p_A)}$ est la moyenne de l'échantillon des y_i incluses dans l'«échantillon». Le rendement de $\hat{Y}_{\mathrm{CHs},p_A}$ dépend de $\hat{N}_{\mathrm{CH}}$, qui a sous-estimé N appréciablement pour le modèle 2. Les résultats pour $\hat{Y}_{\mathrm{CHs},p_A}$ ne sont pas inclus ici. Un autre estimateur non biaisé pour le plan de Y est donné par

$$\hat{Y}_A = \sum_{i \in \text{"échantillon de la base aéolaire"}} \frac{y_i}{p_A}.$$

Il s'agit de l'estimateur de Horvitz-Thompson fondé sur l'échantillon de la base aréolaire seulement. Puisque le dénombrement complet des segments aréolaires coûte cher, p_A est typiquement petite dans la pratique. Pour une petite p_A, $\hat{Y}_A$ devrait avoir une variance beaucoup plus grande que $\hat{Y}_{p_A}$ puisque l'estimateur $\hat{Y}_{p_A}$ inclut des renseignements des listes en plus de renseignements des échantillons de la base aréolaire. Les résultats pour $\hat{Y}_A$ ne sont donc pas inclus.

3.5 Variance estimée de l'estimateur

Dans notre étude de simulation, les valeurs de p_A considérées sont très petites. Par contre, la probabilité d'inclusion dans au moins une des listes est grande pour chaque individu. Par conséquent, $\tilde{\pi}_i$ est proche de $\hat{\pi}_i$ et $\tilde{\pi}_{il}$ en (11) est proche de $\tilde{\pi}_i \tilde{\pi}_l$. Ainsi, le second membre des équations (10) et (17), comportant $\tilde{\pi}_{il} - \tilde{\pi}_i \tilde{\pi}_l$, devrait être petit. Nous n'avons pas inclus ce membre dans notre estimation de la variance. Malgré cette omission, nous remarquons que la variance estimée est très proche de la variance empirique de l'estimateur pour les modèles que nous considérons.

3.6 Statistiques sommaires

Pour les estimations de la taille de la population, nous présentons des moyennes établies pour les 4 000 répétitions correspondant aux quatre valeurs de ρ_{xy} et les 1 000 répétitions de Monte Carlo pour chaque ρ_{xy}. Pour chaque modèle, nous résumons la moyenne et l'écart-type des estimations, la moyenne des erreurs-types estimées des estimateurs, la racine carrée de l'erreur quadratique moyenne relative (REQMR) en pourcentage et les probabilités de couverture empiriques pour un intervalle de confiance de 95%. Ces mesures sont toutes normalisées selon la taille de la population N. Nous présentons des résultats pour les modèles 1 et 2 dans les tableaux 2 et 3, respectivement.

De même, pour les estimations du total de la population, nous présentons des données statistiques sommaires établies sous forme de moyennes pour les 1 000 répétitions correspondant à chaque combinaison paramétrique. Nous résumons la moyenne et l'écart-type des estimations de même que la moyenne des erreurs-types estimées des estimateurs, les estimations étant mises à l'échelle en fonction du vrai total (Y) pour cette répétition. Autrement dit, pour chaque répétition, nous divisons l'estimation par son total de répétition, Y. Nous calculons alors la moyenne et les écarts-types de ces estimations normalisées. De même, pour chaque répétition, nous calculons l'erreur-type estimée de l'estimateur du total divisée par le total pour la répétition, et nous calculons ensuite la moyenne de ces valeurs normalisées. Celles-ci sont signalées parce que les totaux changent d'une répétition à l'autre. Enfin, nous signalons les probabilités de couverture des intervalles de confiance de 95% pour le total. Les résultats pour les modèles 1 et 2 sont présentés dans les tableaux 4 et 5 respectivement.

Tableau 2
Estimations de la taille de la population pour le modèle 1

$N = 300$	$\hat{N}_{CH}$	$\hat{N}_0$	$\hat{N}_{0,05}$	$\hat{N}_{0,20}$
Moyenne des estimations divisée par N	0,999	1,011	1,007	1,004
Écart-type des estimations divisé par N	0,059	0,077	0,059	0,048
Moyenne de l'écart-type estimé de l'estimateur divisée par N	0,059	0,072	0,059	0,047
REQMR en pourcentage	0,003	0,006	0,004	0,002
Couverture	0,947	0,955	0,957	0,950
$N = 1\,000$				
Moyenne des estimations divisée par N	1,000	1,003	1,002	1,002
Écart-type des estimations divisé par N	0,031	0,035	0,030	0,025
Moyenne de l'écart-type estimé de l'estimateur divisée par N	0,032	0,034	0,030	0,025
REQMR en pourcentage	0,001	0,001	0,001	0,001
Couverture	0,954	0,959	0,958	0,956

Tableau 3
Estimations de la taille de la population pour le modèle 2

$N = 300$	$\hat{N}_{CH}$	$\hat{N}_0$	$\hat{N}_{0,05}$	$\hat{N}_{0,20}$
Moyenne des estimations divisée par N	0,922	1,006	1,005	1,003
Écart-type des estimations divisé par N	0,032	0,052	0,049	0,040
Moyenne de l'écart-type estimé de l'estimateur divisée par N	0,028	0,052	0,048	0,040
REQMR en pourcentage	0,007	0,003	0,002	0,002
Couverture	0,271	0,953	0,954	0,951
$N = 1\,000$				
Moyenne des estimations divisée par N	0,921	1,001	1,001	1,001
Écart-type des estimations divisé par N	0,018	0,028	0,027	0,022
Moyenne de l'écart-type estimé de l'estimateur divisée par N	0,015	0,027	0,026	0,021
REQMR en pourcentage	0,007	0,0008	0,0007	0,0005
Couverture	0,009	0,949	0,949	0,949

3.7. Conclusions

3.7.1 Estimation de la taille de la population

Dans le modèle 1, les probabilités d'inclusion ne dépendent pas de la covariable. Dans ce cas, l'estimateur de Chapman $\hat{N}_{CH}$ est très proche de l'estimateur du maximum de vraisemblance (Lincoln-Petersen) et on s'attend donc à ce qu'il donne de meilleurs résultats que $\hat{N}_0$. L'estimateur $\hat{N}_0$ perd de son efficacité puisqu'il sert à estimer les paramètres α_{B_1}, β_{B_1}, α_{B_2}, et β_{B_2}, dont la valeur est nulle dans ce modèle. L'estimateur $\hat{N}_{0,05}$ est à peu près aussi efficace que $\hat{N}_{CH}$. Le biais pour toutes les estimations est minime. Pour le modèle 1, nous remarquons que la moyenne de l'écart-type estimé est proche de l'écart-type des estimations. Cela indique que l'estimation de l'erreur-type que nous utilisons donne de bons résultats. De même, nous remarquons que les probabilités de couverture empiriques se trouvent toutes à trois erreurs-types près de 0,95. Autrement dit, toutes les probabilités de couverture empiriques se trouvent à $(0,95 \pm 3\,[0,95 \times 0,05/4\,000]^{1/2}) = (0,94, 0,96)$ près.

Pour le modèle 2, la probabilité d'inclusion est fonction de la covariable. Par conséquent, $\hat{N}_{CH}$ n'est pas un estimateur approprié pour N. Nous observons que $\hat{N}_{CH}$ sous-estime appréciablement la vraie taille de la population. Par contre, $\hat{N}_{p_A}$ fournit une bonne estimation de N. Le biais dans $\hat{N}_{p_A}$ diminue à mesure que p_A augmente dans le modèle 2. De plus, le biais relatif diminue à mesure que la taille de la population augmente.

Comme on pouvait s'y attendre, l'écart-type de $\hat{N}_{p_A}$ diminue à mesure que la probabilité d'inclusion de la base aréolaire p_A augmente. Par exemple, dans le modèle 1 où $N = 300$, l'inclusion d'un échantillon de base aréolaire de 5% fait baisser l'écart-type relatif de 0,077 à 0,059, une réduction de 23%. Lorsqu'on utilise un échantillon de base aréolaire de 20%, l'écart-type relatif diminue de 0,077 à 0,048, une réduction de 38%. Lorsque $N = 1\,000$, l'inclusion d'un échantillon de base aréolaire de 5% fait baisser l'écart-type relatif de 0,035 à 0,030, une réduction de 14%. Le fait d'augmenter l'échantillon de base aréolaire à 20% fait baisser l'écart-type relatif de 0,035 à 0,025, une diminution de 29%. De façon générale, les erreurs-types relatives diminuent à mesure que la taille de la population augmente.

Tableau 4

Estimations du total de la population par sous-échantillonnage mises à l'échelle selon Y pour le modèle 1

		$N = 300$			$N = 1\,000$		
ρ_{xy}		$\hat{Y}_0/Y$	$\hat{Y}_{0,05}/Y$	$\hat{Y}_{0,20}/Y$	$\hat{Y}_0/Y$	$\hat{Y}_{0,05}/Y$	$\hat{Y}_{0,20}/Y$
-0,3	Moyenne des estimations	1,004	1,003	1,001	1,002	1,002	1,001
	Écart-type des estimations	0,077	0,073	0,062	0,042	0,040	0,035
	Moyenne de l'erreur-type estimée	0,076	0,072	0,061	0,041	0,039	0,033
	Couverture	0,953	0,951	0,949	0,946	0,942	0,942
0	Moyenne des estimations	1,013	1,012	1,008	1,001	1,001	1,001
	Écart-type des estimations	0,080	0,070	0,059	0,041	0,039	0,033
	Moyenne de l'erreur-type estimée	0,081	0,072	0,060	0,040	0,038	0,033
	Couverture	0,951	0,954	0,951	0,944	0,942	0,946
0,5	Moyenne des estimations	1,053	1,018	1,009	1,004	1,003	1,002
	Écart-type des estimations	0,586	0,104	0,072	0,057	0,045	0,037
	Moyenne de l'erreur-type estimée	0,233	0,094	0,070	0,051	0,045	0,036
	Couverture	0,950	0,951	0,945	0,950	0,955	0,955
1,0	Moyenne des estimations	1,064	1,030	1,013	1,013	1,009	1,006
	Écart-type des estimations	0,515	0,162	0,090	0,094	0,066	0,047
	Moyenne de l'erreur-type estimée	0,277	0,128	0,086	0,070	0,059	0,046
	Couverture	0,930	0,929	0,930	0,946	0,949	0,951

Tableau 5

Estimations du total de la population par sous-échantillonnage mises à l'échelle selon Y pour le modèle 2

		$N = 300$			$N = 1\,000$		
ρ_{xy}		$\hat{Y}_0/Y$	$\hat{Y}_{0,05}/Y$	$\hat{Y}_{0,20}/Y$	$\hat{Y}_0/Y$	$\hat{Y}_{0,05}/Y$	$\hat{Y}_{0,20}/Y$
-0,3	Moyenne des estimations	1,010	1,009	1,006	1,003	1,003	1,002
	Écart-type des estimations	0,098	0,092	0,078	0,052	0,049	0,041
	Moyenne de l'erreur-type estimée	0,094	0,089	0,074	0,051	0,048	0,040
	Couverture	0,935	0,926	0,931	0,952	0,955	0,955
0	Moyenne des estimations	1,008	1,007	1,005	1,002	1,002	1,001
	Écart-type des estimations	0,065	0,062	0,050	0,034	0,032	0,027
	Moyenne de l'erreur-type estimée	0,064	0,061	0,051	0,034	0,032	0,028
	Couverture	0,953	0,950	0,952	0,947	0,951	0,955
0,5	Moyenne des estimations	1,002	1,002	1,001	1,001	1,001	1,001
	Écart-type des estimations	0,035	0,033	0,028	0,019	0,018	0,015
	Moyenne de l'erreur-type estimée	0,035	0,034	0,029	0,019	0,018	0,016
	Couverture	0,954	0,950	0,951	0,965	0,967	0,965
1,0	Moyenne des estimations	1,001	1,001	1,001	1,000	1,000	1,000
	Écart-type des estimations	0,021	0,020	0,017	0,012	0,011	0,010
	Moyenne de l'erreur-type estimée	0,021	0,020	0,017	0,012	0,011	0,009
	Couverture	0,952	0,949	0,954	0,947	0,947	0,943

Même si la moyenne de l'erreur-type estimée de $\hat{N}_{p_A}$ est inférieure à l'écart-type empirique, la différence est relativement faible. De même, les probabilités de couverture de l'intervalle de confiance de 95% fondé sur $\hat{N}_{p_A}$ sont très proches de 0,95. Par contre, les probabilités de couverture de l'intervalle de confiance de 95% fondé sur $\hat{N}_{CH}$ sont de 0,271 et 0,009 pour $N = 300$ et $N = 1\ 000$, respectivement.

Compte tenu de nos simulations, nous recommandons l'utilisation de $\hat{N}_{p_A}$ avec une grande valeur de p_A. Le choix de p_A est déterminé dans la pratique par les coûts d'échantillonnage de la base aréolaire, qui ne sont pas pris en considération dans notre étude.

3.7.2 Estimation du total de la population

Pour les totaux de la population, nous obtenons des résultats qui sont très semblables à ce que nous avons observé pour la taille de la population. En général, le biais relatif et l'erreur-type diminuent à mesure que p_A augmente et que la taille de la population augmente. Nous remarquons également que l'erreur-type estimée relative moyenne est très proche de l'écart-type empirique de l'estimateur normalisé en fonction du total. Cela indique que la formule approximative de l'erreur-type en (7) représente une bonne estimation de l'erreur-type. À noter également que les probabilités de couverture empiriques se trouvent pour la plupart à trois erreurs-types près de 0,95. Autrement dit, la plupart des probabilités de couverture empiriques se trouvent à $(0{,}95 \pm 3[0{,}95 \times 0{,}05/1\ 000]^{1/2}) = (0{,}929,\ 0{,}971)$ près.

4. SOMMAIRE

Dans le présent exposé, nous avons étudié le rendement de l'estimateur de Horvitz-Thompson estimé de la taille de la population et du total de la population en fonction d'échantillons tirés de listes et de bases aréolaires. Nous avons présenté des méthodes d'estimation des paramètres du modèle de régression logistique pour les probabilités d'inclusion. Même si de nombreux modèles et autres estimateurs ont été considérés dans Haines (1997), nous présentons des résultats d'étude de simulation pour deux modèles seulement et quelques estimateurs.

Nous croyons que les méthodes utilisées dans le présent exposé pourraient être très utiles aux analystes d'enquête parce que le caractère incomplet des listes est une réalité de tous les jours. Nos résultats comptent parmi les premiers à proposer une méthode d'estimation des chiffres de population qui puisse tenir compte du caractère incomplet et modéliser les probabilités d'inclusion en fonction des covariables.

REMERCIEMENTS

Les auteurs tiennent à remercier le rédacteur et un rédacteur adjoint de remarques qui ont permis d'améliorer le contenu et la présentation de l'exposé. Les auteurs remercient également Christine Bunck, gestionnaire du programme BEST, Biological Resources Division, U.S. Geological Survey, de l'appui financier accordé à la présente recherche dans le cadre d'un bon de travail remis à la North Carolina State University. Les vues exprimées sont celles des auteurs et ne reflètent pas nécessairement les vues du Bureau of the Census.

BIBLIOGRAPHIE

ALHO, J.M. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46, 623-635.

BOSECKER, R.R., et FORD, B.L. (1976). Multiple frame estimation with stratified overlap domain. *Proceedings of the Social Statistics Section, American Statistical Association*, 219-224.

BURNHAM, K.P. (1972). Estimation of Population Size in Multiple Capture Studies when Capture Probabilities Vary Among Animals. Thèse de doctorat, Oregon State University.

BURNHAM, K.P., et OVERTON, W.S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, 65, 625-633.

BURNHAM, K.P., et OVERTON, W.S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology*, 60, 927-936.

CHAO, A. (1988). Estimating animal abundance with capture frequency data. *Journal of Wildlife Management*, 52, 295-300.

CHAO, A., LEE, S.-M., et JENG, S.-L. (1992). Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics*, 48, 201-216.

CHAPMAN, D.G. (1951). Some Properties of the Hypergeometric Distribution with Applications to Zoological Censuses. University of California, University of California Publication in Statistics.

COCHRAN, R.S. (1965). Theory and Applications of Multiple Frame Surveys. Thèse de doctorat, Iowa State University.

COCHRAN, W.G. (1977). *Sampling Techniques*. 3-ième édition. New York: John Wiley & Sons.

COWAN, C.D., et MALEC, D. (1986). Capture-recapture models when both sources have clustered observations. *Journal of the American Statistical Association*, 81, 347-353.

FAULKENBERRY, G.D., et GAROUI, A. (1991). Estimating a population total using an area frame. *Journal of the American Statistical Association*, 86, 445-449.

FECSO, R., TORTORA, R.D., et VOGEL, F.A. (1986). Sampling frames for agriculture in the United States. *Journal of Official Statistics*, 2, 279-292.

FIENBERG, S.E. (1992). Bibliographie sur la modélisation à l'aide de la saisie-resaisie avec application au redressement des chiffres du recensement pour éliminer le sous-dénombrement. *Techniques d'enquête*, 18, 157-169.

FULLER, W.A., et BURMEISTER, L.F. (1972). Estimators for samples selected from two overlapping frames. *Proceedings of the Social Statistics Section, American Statistical Association*, 245-249.

HAINES, D.E. (1997). Estimating Population Parameters Using Multiple Frame and Capture-Recapture Methodology. Thèse de doctorat, North Carolina State University.

HAINES, D.E., et POLLOCK, K.H. (1998a).Combinaison de bases multiples pour estimer la taille et les chiffres de la population. *Techniques d'enquête*, 24, 81-9188.

HAINES, D.E., et POLLOCK, K.H. (1998b). Estimating the number of active and successful bald eagle nests: an application of the dual frame method. *Environmental and Ecological Statistics*, 5, 245-256.

HANSEN, M.H., HURWITZ, W.N., et MADOW, W.G. (1953). *Sample Survey Methods and Theory 1*. New York: John Wiley & Sons.

HARTLEY, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.

HARTLEY, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, 36, 3, C, 99-118.

HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

HUGGINS, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76, 133-140.

KOTT, P.S., et VOGEL, F.A. (1995). Multiple-frame business surveys. *Business Survey Methods* (Éd. B.G. Cox.). New York: John Wiley & Sons. 185-203.

LINCOLN, F.C. (1930). Calculating Waterfowl Abundance on the Basis of Banding Returns. U.S. Department of Agriculture, Circular, 118.

LUND, R.E. (1968). Estimators in multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 282-288.

NEALON, J.P. (1984). Review of the Multiple and Area Frame Estimators, Staff Report 80. U.S. Department of Agriculture, Statistical Reporting Service, Washington, D. C.

NORRIS, J.L., et POLLOCK, K.H. (1996). Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics*, 52, 639-649.

OTIS, D.L., BURNHAM, K.P., WHITE, G.C., et ANDERSON, D.R. (1978). Statistical inference for capture data on closed animal populations. *Wildlife Monographs*, 62, 1-135.

PETERSEN, C.G.J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. *Rep. Danish Biol. Sta.*, 6, 1-48.

POLLOCK, K.H. (1991). Modeling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present, and future. *Journal of the American Statistical Association*, 86, 225-238.

POLLOCK, K.H., HINES, J.E., et NICHOLS, J.D. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics*, 40, 329-340.

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.

QUENOUILLE, M.H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353-360.

SEKAR, C.C., et DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.

WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.

WOLTER, K.M. (1990). Capture-recapture estimation in the presence of a known sex ratio. *Biometrics*, 46, 157-162.

Techniques d'enquête, décembre 2000
Vol. 26, n° 2, pp. 145-151
Statistique Canada

Une méthode d'estimation en présence de non-réponse non-ignorable

JEAN-FRANÇOIS BEAUMONT[1]

RÉSUMÉ

Lorsque le mécanisme de réponse dans une enquête dépend d'une variable d'intérêt mesurée dans cette même enquête et qui n'est observée que pour une partie de l'échantillon seulement, on dit qu'on est en présence de non-réponse non-ignorable. Dans une telle situation, ne pas tenir compte de la non-réponse peut engendrer un biais important dans l'estimation d'une moyenne ou d'un total. Pour contrer ce problème, on peut modéliser conjointement le mécanisme de réponse et la variable d'intérêt et effectuer l'estimation par la méthode du maximum de vraisemblance. La critique principale de cette méthode est que l'estimation par la méthode du maximum de vraisemblance est basée sur l'hypothèse difficilement vérifiable de normalité des erreurs pour le modèle impliquant la variable d'intérêt. Dans cet article, on propose une méthode d'estimation robuste par rapport à l'hypothèse de normalité puisqu'elle est construite de telle sorte qu'elle n'exige pas de spécifier la distribution des erreurs. La méthode est évaluée au moyen de simulations de Monte Carlo. On propose également une méthode simple permettant de vérifier la validité de l'hypothèse de normalité des erreurs quand la non-réponse n'est pas ignorable.

MOTS CLÉS: Non-réponse non-ignorable; maximum de vraisemblance; équations d'estimation; imputation par la régression; repondération.

1. INTRODUCTION

Lorsque le mécanisme de réponse dans une enquête dépend d'une variable d'intérêt mesurée dans cette même enquête et qui n'est observée que pour une partie de l'échantillon seulement, on dit qu'on est en présence de non-réponse non-ignorable. Par exemple, lorsqu'on mesure le revenu, il peut être réaliste de supposer que les personnes ayant un faible revenu ont moins tendance à répondre que les personnes ayant un revenu élevé ou vice-versa. Le lecteur est renvoyé à Little (1982) pour une définition formelle du concept de non-réponse non-ignorable. Dans une telle situation, ne pas tenir compte de la non-réponse peut engendrer un biais important dans l'estimation d'une moyenne ou d'un total. Pour contrer ce problème, on peut modéliser conjointement le mécanisme de réponse et la variable d'intérêt, effectuer l'estimation par la méthode du maximum de vraisemblance comme, entre autres, dans Greenlees, Reece et Zieschang (1982) et imputer les valeurs manquantes. La critique principale de cette méthode est que l'estimation par la méthode du maximum de vraisemblance est basée sur l'hypothèse difficilement vérifiable de normalité des erreurs pour le modèle impliquant la variable d'intérêt.

Rancourt, Lee et Särndal (1994) présentent des facteurs de correction simples pour réduire le biais engendré par la non-réponse qui n'est pas ignorable sans recourir à l'hypothèse de normalité et sans modèle pour le mécanisme de réponse. Ces facteurs de correction ne sont toutefois disponibles que pour l'imputation par le quotient.

Dans cet article, on propose une méthode d'estimation robuste par rapport à l'hypothèse de normalité puisqu'elle est construite de telle sorte qu'elle n'exige pas de spécifier la distribution des erreurs. On propose également une méthode simple permettant de vérifier la validité de l'hypothèse de normalité des erreurs quand la non-réponse n'est pas ignorable.

À la deuxième section, on définit le problème et on introduit un peu de notation. La troisième section sert à présenter différents estimateurs de la moyenne d'une population sous différentes hypothèses concernant le mécanisme de réponse et la distribution des données. À la quatrième section, on présente la méthode d'estimation proposée quand la non-réponse n'est pas ignorable. La cinquième section sert à décrire les résultats d'une étude de simulation permettant de comparer les différents estimateurs décrits aux deux sections précédentes. Finalement, la dernière section contient une brève discussion.

2. NOTATION

Dans ce qui va suivre, on tente d'estimer la moyenne de la variable Y pour une certaine population P. Pour y arriver, on sélectionne un échantillon S et la variable Y est observée pour une partie de l'échantillon seulement. L'échantillon des répondants est noté R et l'échantillon des non-répondants est noté O. On suppose qu'on dispose d'au moins une variable observée pour toutes les unités de l'échantillon et corrélée avec Y.

L'estimateur de la moyenne, $\mu = \sum_{i\in P} Y_i/N$, où N est la taille de la population, peut être obtenu en pondérant les unités répondantes:

$$\mu_P^* = \frac{\sum_{i\in R} w_i w_{R,i}^* Y_i}{\sum_{i\in R} w_i w_{R,i}^*}, \qquad (2.1)$$

[1] Jean-François Beaumont, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6.

où w_i sont les poids d'échantillonnage correspondant à l'inverse de la probabilité de sélection et $w_{R,i}^*$ sont des poids correspondant à l'inverse de la probabilité de réponse estimée. Un autre estimateur de la moyenne peut être obtenu en imputant les valeurs manquantes:

$$\mu_I^* = \frac{\sum_{i\in R} w_i Y_i + \sum_{i\in O} w_i Y_i^*}{\sum_{i\in S} w_i}, \qquad (2.2)$$

où les Y_i^* sont des valeurs imputées pour les unités non-répondantes.

Pour des raisons de simplicité, on va supposer, dans ce qui va suivre, que les poids d'échantillonnage sont constants pour toutes les unités de la population. Ainsi, on peut éliminer w_i des équations (2.1) et (2.2). On va aussi supposer qu'on ne dispose que d'une seule variable observée pour toutes les unités de l'échantillon. Cette variable sera notée X.

3. MÉTHODES D'ESTIMATION COURANTES

Cette section est consacrée à développer les expressions (2.1) et (2.2) sous différentes hypothèses concernant le mécanisme de réponse et la distribution des données et à décrire des méthodes d'estimation appropriées. À la section (3.1), on suppose un mécanisme de réponse uniforme, à la section (3.2), on suppose un mécanisme de réponse qui dépend de X et à la section (3.3), on suppose un mécanisme de réponse qui dépend de Y. Les mécanismes de réponse des sections (3.1) et (3.2) sont ignorables alors que celui de la section (3.3) n'est pas ignorable.

3.1 Mécanisme de réponse uniforme

Sous l'hypothèse d'un mécanisme de réponse uniforme, toutes les unités de l'échantillon ont la même probabilité de réponse. L'estimateur (2.1) devient alors:

$$\mu_{P,U}^* = \frac{\sum_{i\in R} Y_i}{n_R}, \qquad (3.1)$$

où n_R est le nombre total de répondants. Cet estimateur est le même que celui qu'on aurait obtenu en utilisant l'équation (2.2) et en imputant la moyenne des répondants à tous les non-répondants.

3.2 Mécanisme de réponse qui dépend de X

Lorsque le mécanisme de réponse dépend de la variable X (qui est corrélée avec Y), l'estimateur (3.1) peut être fortement biaisé. Il est alors préférable d'utiliser cette variable comme information supplémentaire pour l'estimation de la moyenne μ.

L'estimateur (2.1) peut être obtenu en substituant $1/w_{R,i}^*$ par la probabilité de réponse estimée au moyen d'une régression logistique. Un modèle est donc requis pour la probabilité de réponse. Si on ne dispose que d'une seule variable (X) observée pour toutes les unités de l'échantillon, ce modèle peut s'écrire:

$$P(R_i = 1 \mid X_i) = \frac{1}{1 + \exp[-(\alpha_0 + \alpha_1 X_i)]},$$

où α_0 et α_1 sont des paramètres à estimer (par la méthode du maximum de vraisemblance, par exemple) et R_i est une variable dichotomique qui égale 1 si l'unité i répond et 0 sinon. L'estimateur de la moyenne obtenu de cette façon sera noté $\mu_{P,X}^*$.

Si on préfère plutôt utiliser l'estimateur (2.2), on peut imputer les valeurs manquantes en se servant du modèle suivant:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad (3.2)$$

où β_0 et β_1 sont des paramètres inconnus et ε_i est un terme d'erreur aléatoire de moyenne nulle non corrélé avec les X_i. Les valeurs imputées sont données par: $Y_i^* = B_0^* + B_1^* X_i$, où B_0^* et B_1^* sont des estimations (obtenues par la méthode des moindres carrés au moyen des unités répondantes) de B_0 et B_1 qui sont elles-mêmes des estimations de β_0 et β_1. En fait, B_0 et B_1 sont les estimations que l'on aurait obtenues (par la méthode des moindres carrés) si on avait observé toutes les unités de l'échantillon S. L'estimateur obtenu de cette façon sera noté $\mu_{I,X}^*$. Il est à noter que tous les modèles considérés dans ce document sont supposés valides pour toutes les unités de l'échantillon S.

On pourrait également ajouter un résidu aux valeurs imputées dans le but de mieux estimer la variance due à l'échantillonnage (voir, par exemple, Gagnon, Lee, Rancourt et Särndal 1996). Cependant, cette technique ne permet toujours pas d'estimer la variance due à l'imputation. De plus, elle a tendance à produire des estimations de la moyenne moins précises que si l'on n'avait pas ajouté de résidus. Comme l'objet du présent article ne porte pas sur l'estimation de la variance, on a choisi de ne pas ajouter de résidus aux valeurs imputées. Ceci comporte également l'avantage de simplifier le calcul de $\mu_{I,X}^*$.

3.3 Mécanisme de réponse qui dépend de Y

Tous les estimateurs de la moyenne présentés jusqu'à maintenant peuvent être fortement biaisés lorsque le mécanisme de réponse dépend de Y (mécanisme de réponse non-ignorable). Pour un tel mécanisme de réponse, on peut modéliser la probabilité de réponse de la façon suivante:

$$P(R_i = 1 \mid Y_i) = \frac{1}{1 + \exp[-(\alpha_0 + \alpha_1 Y_i)]}. \qquad (3.3)$$

Puisque la variable Y n'est observée que pour les unités répondantes, il est impossible d'obtenir une estimation pour

α_0 et α_1 par la méthode du maximum de vraisemblance. On peut aussi utiliser le modèle (3.2). Cependant, les estimations des paramètres ne seront pas convergentes parce que $E(\varepsilon_i \mid R_i = 1)$ et $E(\varepsilon_i X_i \mid R_i = 1)$ ne sont pas nulles. Même si on disposait d'estimations convergentes, on ne pourrait pas imputer les valeurs manquantes de la façon décrite à la section (3.2) puisque $E(Y_i \mid R_i = 0, X_i) \neq \beta_0 + \beta_1 X_i$ (Greenlees, Reece et Zieschang 1982). Par exemple, si la probabilité de réponse est corrélée positivement avec la variable d'intérêt Y alors, pour une valeur de X donnée, la moyenne des unités non-répondantes sera inférieure à celle des unités répondantes et sera donc inférieure à la moyenne de toutes les unités réunies. Un raisonnement similaire peut être fait si la probabilité de réponse est corrélée négativement avec la variable d'intérêt. On peut d'ailleurs montrer que

$$E(Y_i \mid R_i = 0, X_i) = \beta_0 + \beta_1 X_i - \frac{\mathrm{cov}(Y_i, p(Y_i) \mid X_i)}{1 - E(p(Y_i) \mid X_i)},$$

où $p(Y_i) = P(R_i = 1 \mid Y_i)$.

Les deux approches de la section (3.2) ne sont donc pas valides lorsque le mécanisme de réponse n'est pas ignorable. Dans une telle situation, une meilleure approche consiste à estimer simultanément les paramètres des modèles (3.2) et (3.3). La méthode du maximum de vraisemblance peut être utilisée à cette fin. Cette méthode requiert cependant l'hypothèse supplémentaire que les erreurs ε_i soient normalement distribuées (ou toute autre distribution pertinente pour le type de données à analyser), aient une variance constante σ^2 et soient indépendantes entre elles. Le logarithme naturel de la fonction de vraisemblance l peut être écrit:

$$l = \sum_{i \in R} \ln[p(Y_i) f(Y_i \mid X_i)] + \sum_{i \in O} \ln[1 - E(p(Y_i) \mid X_i)], \quad (3.4)$$

où $f(Y_i \mid X_i)$ est la fonction de densité de probabilité d'une loi normale de moyenne $\beta_0 + \beta_1 X_i$ et de variance σ^2. La méthode du maximum de vraisemblance consiste à trouver les valeurs des paramètres qui maximisent l. Pour effectuer la maximisation, on doit être en mesure d'approximer $E(p(Y_i) \mid X_i)$. On peut y arriver en utilisant une méthode d'intégration numérique comme dans Greenlees, Reece et Zieschang (1982). Dans cet article, on a plutôt opté pour l'approximation suivante (Zeger, Liang et Albert 1988):

$$E(p(Y_i) \mid X_i) \approx \frac{1}{1 + \exp\{-k[\alpha_0 + \alpha_1(\beta_0 + \beta_1 X_i)]\}}, \quad (3.5)$$

où $k = 1/\sqrt{c^2 \sigma^2 \alpha_1^2 + 1}$ et $c = 16\sqrt{3}/15\pi$. Cette approximation est basée sur l'hypothèse que les erreurs sont normalement distribuées et ont une variance constante. On a préféré cette approximation à une méthode d'intégration numérique parce qu'elle est plus simple et plus rapide du point de vue informatique, ce qui est un avantage devant être sérieusement considéré avant d'entreprendre une étude de simulation. Finalement, on a effectué la maximisation de l'équation (3.4) à l'aide de l'algorithme Newton-Raphson et de la procédure NLIN du logiciel SAS (SAS Institute Inc. 1990).

Une fois que les paramètres des modèles (3.2) et (3.3) sont estimés, on peut choisir l'estimateur de la moyenne (2.1) ou (2.2). L'estimateur (2.1) est obtenu en remplaçant $w_{R,i}$ par $1/p^*(Y_i)$, où $p^*(Y_i)$ est la probabilité de réponse estimée. Cet estimateur sera noté $\hat{\mu}_{P,Y,ML}$. L'estimateur (2.2) peut être obtenu en déterminant les valeurs imputées Y_i^* de telle sorte qu'on minimise $\sum_{i \in S} e_i^2$ et que les contraintes $\sum_{i \in S} e_i = 0$ et $\sum_{i \in S} e_i X_i = 0$ soient respectées, où $e_i = Y_i - \beta_0^* - \beta_1^* X_i$, pour $i \in R$, $e_i = Y_i^* - \beta_0^* - \beta_1^* X_i$, pour $i \in O$, et β_0^* et β_1^* sont les estimations de β_0 et β_1 respectivement. L'estimateur de la moyenne peut alors s'écrire: $\hat{\mu}_{I,Y,ML}^* = \beta_0^* + \beta_1^* \sum_{i \in S} X_i / n$, où n est la taille de l'échantillon S. La logique derrière cette approche est que les deux contraintes précédentes auraient été respectées si on avait observé la variable Y pour toutes les unités de l'échantillon et qu'on avait modélisé cette variable à l'aide du modèle (3.2).

4. MÉTHODE D'ESTIMATION PROPOSÉE

Cette section sert à décrire la méthode d'estimation proposée pour un mécanisme de réponse non-ignorable (section 4.1) de même qu'une méthode graphique (section 4.2) permettant de vérifier l'hypothèse de normalité des erreurs du modèle (3.2).

4.1 Méthode d'estimation pour un mécanisme de réponse qui dépend de Y

La méthode du maximum de vraisemblance est valide quand les erreurs sont normalement distribuées et ont la même variance. Lorsque cette hypothèse est violée, il est préférable de recourir à une méthode d'estimation plus robuste.

Si les probabilités de réponse $p(Y_i)$ étaient connues et supérieures à zéro pour toutes les unités de l'échantillon, alors une méthode d'estimation robuste (autant par rapport à l'hypothèse de normalité des erreurs que par rapport au modèle 3.2) consisterait à minimiser la somme des erreurs au carré pondérées par l'inverse de la probabilité de réponse $p(Y_i)$. Cette minimisation est équivalente à résoudre le système d'équations

$$\sum_{i \in R} \frac{1}{p(Y_i)} (Y_i - \beta_0 - \beta_1 X_i) Z_{ik} = 0, \quad k = 1, 2, \quad (4.1)$$

où $Z_{i1} = 1$ et $Z_{i2} = X_i$. Cette approche est considérée robuste par rapport à l'hypothèse de normalité puisque la méthode des moindres carrés n'exige pas de spécifier la distribution des erreurs. Le fait de pondérer par l'inverse de

la probabilité de réponse procure également une certaine robustesse par rapport au modèle (3.2). En effet, les estimateurs B_0^* et B_1^* obtenus au moyen de l'équation (4.1) sont convergents par rapport au mécanisme de réponse pour B_0 et B_1 (qui sont les estimateurs de β_0 et β_1 que l'on aurait obtenus s'il n'y avait pas eu de non-réponse) peu importe la validité du modèle. Le lecteur est renvoyé à Särndal, Swensson et Wretman (1992, p. 519) pour un argument similaire, à l'exception qu'il est question du mécanisme de sélection de l'échantillon et non du mécanisme de réponse.

De même, si la fonction de densité de probabilité $f(Y_i \mid X_i)$ était connue (pas nécessairement normale mais ne doit pas dépendre des paramètres du modèle 3.3), alors on pourrait estimer les paramètres α_0 et α_1 du modèle (3.3) par la méthode du maximum de vraisemblance, par exemple, et résoudre le système d'équations

$$\sum_{i\in R} \frac{\partial}{\partial \alpha_k} \ln[p(Y_i)] + \sum_{i\in O} \frac{\partial}{\partial \alpha_k} \ln[1 - \mathrm{E}(p(Y_i) \mid X_i)] = 0, \quad (4.2)$$

pour $k = 0$ et $k = 1$.

Ainsi, les estimations des paramètres β_0, β_1, α_0 et α_1 sont obtenues en résolvant les équations d'estimation non biaisées (4.1) et (4.2). Un algorithme qui permet de trouver la solution consiste à résoudre alternativement les systèmes d'équation (4.1) et (4.2) jusqu'à ce que la convergence soit atteinte. Pour cela, il faut être en mesure de calculer $\mathrm{E}(p(Y_i) \mid X_i)$ dans l'équation (4.2). Cette dernière espérance requiert toutefois de connaître la distribution des erreurs ε_i, qui est vraisemblablement inconnue. Pour faire face à ce problème, on doit utiliser une approximation et plusieurs peuvent être envisagées, dont l'approximation (3.5). Il serait également possible de développer une stratégie basée sur la méthode du bootstrap en choisissant les unités répondantes proportionnellement à l'inverse de leur probabilité de réponse. Cependant, cette méthode requiert un temps d'exécution important du point de vue informatique et ne sera pas considérée dans ce qui va suivre. Dans cet article, on a plutôt opté pour l'approximation suivante obtenue en linéarisant $p(Y_i)$ au moyen d'une série de Taylor évaluée au point $\mathrm{E}(Y_i \mid X_i)$ et en prenant l'espérance des deux premiers termes de cette série:

$$\mathrm{E}(p(Y_i) \mid X_i) \approx p(\mathrm{E}(Y_i \mid X_i)) = p(\beta_0 + \beta_1 X_i). \quad (4.3)$$

Il est à noter que l'espérance du deuxième terme de la série est nulle. Cette approximation possède l'avantage de n'exiger que le premier moment de la distribution de Y_i conditionnelle à X_i. En ce sens, elle devrait être robuste par rapport à l'hypothèse de normalité des erreurs puisqu'elle ne requiert pas de spécifier la distribution des erreurs. Bien entendu, si la distribution des erreurs est connue ou peut être bien estimée, il est possible de trouver de meilleures approximations que (4.3) quoique, dans ce cas, il puisse être préférable d'utiliser la méthode du maximum de vraisemblance.

Une autre propriété intéressante de l'approximation (4.3) est que résoudre alternativement les systèmes d'équations (4.1) et (4.2) peut être obtenu au moyen de l'algorithme suivant:

1. déterminer des valeurs initiales pour les probabilités de réponse (ou pour les paramètres α_0 et α_1), par exemple poser $p(Y_i)^{(0)} = 1$ pour toutes les unités répondantes;
2. poser $j = 1$, où j indique le nombre d'itérations;
3. résoudre le système d'équations (4.1) au moyen des estimations courantes des probabilités de réponse, $p(Y_i)^{(j-1)}$, en utilisant une procédure de régression pondérée pour ainsi obtenir $\beta_0^{(j)}$ et $\beta_1^{(j)}$;
4. imputer les valeurs manquantes par $Y_i^{(j)} = \beta_0^{(j)} + \beta_1^{(j)} X_i$, pour $i \in O$;
5. résoudre le système d'équations (4.2) en utilisant une procédure de régression logistique pour obtenir $p(Y_i)^{(j)}$;
6. arrêter si la convergence est atteinte, sinon poser $j = j + 1$ et retourner à l'étape 3.

Il suffit donc uniquement de disposer d'une procédure de régression linéaire et d'une procédure de régression logistique pour obtenir les estimations désirées. En pratique, cet algorithme est très efficace pour trouver la solution quoique, dans certains cas, il puisse prendre beaucoup d'itérations avant de converger. En fait, il a convergé dans tous les cas où il a été utilisé. Il est également à noter que cet algorithme possède certaines ressemblances avec l'algorithme EM de Dempster, Laird et Rubin (1977), à l'exception qu'ici on ne maximise pas une fonction de vraisemblance.

Dans les simulations de la prochaine section, on a plutôt opté pour l'algorithme Newton-Raphson qui est plus rapide à converger. Cependant, on a dû avoir recours à l'algorithme ci-haut pour les quelques cas pour lequel l'algorithme Newton-Raphson a éprouvé des problèmes de convergence.

L'algorithme proposé pourrait s'avérer très utile pour fournir des valeurs initiales à un algorithme plus rapide tel que l'algorithme Newton-Raphson. Il suffirait d'utiliser l'algorithme proposé avec un critère de convergence peu sévère de telle sorte qu'après quelques itérations seulement, il puisse fournir des valeurs initiales suffisamment bonnes pour assurer la convergence de l'algorithme Newton-Raphson. Dans un autre contexte, Beaumont et Demnati (1998) utilisent une approche similaire qui consiste à commencer le processus itératif avec un algorithme de type EM pour ensuite fournir des valeurs initiales à un algorithme plus rapide de type Newton-Raphson. Ils montrent empiriquement que la combinaison de ces deux algorithmes fournit un bon compromis entre le temps d'exécution et l'efficacité pour trouver la solution.

Comme à la section (3.3), une fois que les paramètres des modèles (3.2) et (3.3) sont estimés, on peut choisir l'estimateur de la moyenne (2.1) ou (2.2). L'estimateur (2.1) est obtenu en remplaçant $w_{R,i}^*$ par $1/p^*(Y_i)$, où $p^*(Y_i)$ est la probabilité de réponse estimée. Cet estimateur sera noté $\mu_{P,Y,ROB}^*$. L'estimateur (2.2) est également obtenu comme à la section (3.3) en déterminant les valeurs imputées Y_i^* de telle sorte qu'on minimise $\sum_{i\in S} e_i^2$ et que les contraintes $\sum_{i\in S} e_i = 0$ et $\sum_{i\in S} e_i X_i = 0$ soient respectées, où $e_i = Y_i - B_0^* - B_1^* X_i$, pour $i\in R$, et $e_i = Y_i^* - B_0^* - B_1^* X_i$, pour $i\in O$. Cet estimateur sera noté $\mu_{I,Y,ROB}^*$. La qualité de ces deux estimateurs de la moyenne dépendra en grande partie de la validité des modèles (3.2) et (3.3) et de la qualité de l'approximation (4.3).

Une modification de l'étape (5) de l'algorithme présenté dans cette section a été proposée dans Beaumont (1999). Les résultats d'une étude de simulation montrent que cette modification donne des résultats légèrement supérieurs à ceux obtenus au moyen de la méthode proposée dans cet article. Cependant, cela ne revient plus à utiliser la méthode du maximum de vraisemblance pour estimer les paramètres du modèle (3.3) étant donné que $f(Y_i|X_i)$ est connue et on ne peut plus utiliser une procédure de régression logistique à l'étape (5). Il faut toutefois mentionner qu'il n'est pas absolument obligatoire d'avoir recours à la méthode du maximum de vraisemblance pour estimer α_0 et α_1, même si c'est la méthode qui a été privilégiée dans cet article.

4.2 Vérification de l'hypothèse de normalité des erreurs

Dans le but d'utiliser la méthode du maximum de vraisemblance, on peut être intéressé à vérifier l'hypothèse de normalité des erreurs (ou plutôt l'hypothèse de normalité des résidus puisque les erreurs ne sont pas observées). Lorsqu'il n'y a pas de non-réponse, une méthode classique (D'Agostino 1986, p. 25, équation 2.11) consiste à faire le graphique de $\Phi^{-1}[F_n(e_i)]$ en fonction des résidus e_i, pour $i\in S$, où $\Phi(.)$ est la fonction de répartition d'une variable aléatoire suivant la loi normale standard et $F_n(.)$ est la fonction de répartition empirique. Lorsque les erreurs sont normalement distribuées, les points de ce graphique devraient à peu près s'aligner le long d'une droite de pente $1/\sigma$ passant par l'origine.

En présence de non-réponse, on peut utiliser la même technique qu'au paragraphe précédent mais on doit estimer la fonction de répartion empirique au moyen des unités répondantes. Puisque les unités de l'échantillon répondent avec probabilités inégales, la fonction de répartition empirique estimée peut être donnée par (Särndal, Swensson et Wretman 1992, p. 199):

$$F_n^*(e_i) = \frac{\sum_{j:j\in R \text{ et } e_j \le e_i} 1/p^*(Y_j)}{\sum_{j\in R} 1/p^*(Y_j)}.$$

Il est à noter que, dans cette dernière équation, les probabilités de réponse sont estimées contrairement à la formule de Särndal, Swensson et Wretman dans laquelle les probabilités de sélection sont connues. Ainsi, on peut vérifier l'hypothèse de normalité des erreurs en faisant le graphique de $\Phi^{-1}[F_n^*(e_i)]$ en fonction des résidus e_i, pour $i\in R$. Cette méthode sera valide à condition que $F_n^*(e_i)$ estime correctement $F_n(e_i)$, ce qui est le cas lorsque les probabilités de réponse sont correctement estimées. Lorsque la non-réponse est non-ignorable et qu'on se sert de la méthode d'estimation proposée dans cet article, les probabilités de réponse devraient être bien estimées si les modèles (3.2) et (3.3) sont appropriés de même que l'approximation (4.3).

5. ÉTUDE DE SIMULATION

Afin de comparer les estimateurs de la moyenne présentés aux deux sections précédentes, on a effectué une étude de simulation. On a simulé 4 populations de taille 1 000 selon le modèle (3.2) avec $\beta_0 = 2$ et $\beta_1 = 3$. Les variables aléatoires X_i sont indépendantes entre elles et suivent la loi exponentielle de moyenne 1. Les erreurs ε_i sont indépendantes entre elles, ne sont pas corrélées avec les X_i et sont de moyenne nulle et de variance σ^2. Deux populations ont leurs erreurs suivant la loi normale ($\varepsilon_i \sim \text{Nor}(0,\sigma^2)$) et les deux autres ont leurs erreurs suivant la loi exponentielle de moyenne σ recentrée à $0(\varepsilon_i \sim \text{Exp}(\sigma) - \sigma)$. Pour chacune de ces distributions, une population a l'écart-type σ égal à 1,5 correspondant à un coefficient de corrélation (entre X et Y) au carré de 80% ($R^2 = 80\%$) et l'autre a l'écart-type égal à 3 correspondant à un coefficient de corrélation au carré de 50% ($R^2 = 50\%$).

Pour chaque population, on a simulé 1 000 échantillons de répondants selon le modèle (3.3) avec $\alpha_1 = 0{,}5$. Le paramètre α_0 est déterminé séparément pour chacune des 4 populations de telle sorte que le taux de réponse moyen soit de 70%. Ce paramètre varie entre -1,185 et -0,958. Il est à noter qu'on effectue ici un recensement ($n = N = 1\,000$). Ceci a l'avantage de se concentrer seulement sur l'erreur due à la non-réponse puisqu'il n'y a aucune erreur due à l'échantillonnage. De plus, le fait de générer des populations de taille relativement grande (1 000) permet de mettre surtout l'emphase sur le biais des estimateurs plutôt que sur leur variance puisque la variance devrait diminuer à mesure que la taille de population augmente (pour un taux de réponse moyen fixe).

Pour chacun des 1 000 échantillons de répondants, on a calculé les 7 estimations de la moyenne décrites aux deux sections précédentes. On a ensuite calculé, pour chaque population, la moyenne et la variance de ces 1 000 estimations, notées $\bar{\mu}^*$ et $S_{\mu^*}^2$, respectivement. Finalement, on a calculé une estimation du biais relatif (exprimé en pourcentage), $\text{BR}^* = [(\bar{\mu}^* - \mu)/\mu] \times 100\%$, une estimation de l'erreur-type associée à ce biais relatif, $\text{ET}^* = (100/\mu)\sqrt{S_{\mu^*}^2/1\,000}$, et une estimation de la racine carrée de l'erreur quadratique moyenne, $\text{REQM}^* = \sqrt{S_{\mu^*}^2 + (\bar{\mu}^* - \mu)^2}$.

Le tableau 1 présente les résultats de l'étude de simulation. L'analyse de ce tableau indique que, peu importe la distribution des erreurs, le biais relatif et l'erreur quadratique moyenne de tous les estimateurs est plus faible lorsque la corrélation entre X et Y est plus élevée, ce qui n'est pas surprenant.

Tableau 1
Résultats des simulations permettant de comparer 7 estimateurs de la moyenne μ

Estimateur	R^2 = 80%			R^2 = 50%		
	BR*(%)	ET*	REQM*	BR*(%)	ET*	REQM*
Population avec les erreurs normalement distribuées						
$\hat{\mu}_{P,U}^{*}$	16,90	0,03	0,84	26,68	0,04	1,33
$\hat{\mu}_{P,X}^{*}$	5,65	0,02	0,28	18,02	0,03	0,90
$\hat{\mu}_{P,Y,ML}^{*}$	-0,14	0,03	0,05	1,27	0,10	0,17
$\hat{\mu}_{P,Y,ROB}^{*}$	1,14	0,03	0,08	10,12	0,06	0,51
$\hat{\mu}_{I,X}^{*}$	5,50	0,02	0,27	17,74	0,03	0,89
$\hat{\mu}_{I,Y,ML}^{*}$	0,13	0,03	0,04	1,03	0,07	0,13
$\hat{\mu}_{I,Y,ROB}^{*}$	0,64	0,03	0,05	7,53	0,06	0,39
Population avec les erreurs exponentiellement distribuées						
$\hat{\mu}_{P,U}^{*}$	17,83	0,04	0,86	26,60	0,05	1,29
$\hat{\mu}_{P,X}^{*}$	5,44	0,02	0,26	16,06	0,04	0,78
$\hat{\mu}_{P,Y,ML}^{*}$	-0,54	0,02	0,04	5,18	0,05	0,26
$\hat{\mu}_{P,Y,ROB}^{*}$	1,31	0,02	0,07	7,43	0,03	0,36
$\hat{\mu}_{I,X}^{*}$	5,19	0,02	0,25	15,41	0,03	0,75
$\hat{\mu}_{I,Y,ML}^{*}$	-3,42	0,03	0,17	-25,47	0,05	1,23
$\hat{\mu}_{I,Y,ROB}^{*}$	0,49	0,02	0,04	4,07	0,03	0,20

L'analyse du biais relatif indique que la méthode du maximum de vraisemblance est la meilleure lorsque les erreurs sont normalement distribuées, suivie par la méthode d'estimation robuste décrite à la section (4.1). Les estimateurs qui supposent un mécanisme de réponse non-ignorable ont un biais relatif plus faible que ceux qui supposent à tort un mécanisme de réponse ignorable. Parmi ces derniers estimateurs, l'estimateur $\hat{\mu}_{P,U}^{*}$ est le plus biaisé. Pour une méthode donnée, il existe, en général, peu de différences entre l'estimateur pondéré (2.1) et l'estimateur comprenant des valeurs imputées (2.2). Il faut toutefois donner un léger avantage à ce dernier.

Les conclusions du paragraphe précédent s'appliquent toujours lorsque les erreurs sont exponentiellement distribuées à l'exception que la méthode d'estimation robuste devient la meilleure. Cette constatation n'est pas surprenante puisque la méthode du maximum de vraisemblance est basée sur l'hypothèse de normalité des erreurs. Toutefois, l'estimateur pondéré $\hat{\mu}_{P,Y,ML}^{*}$ reste faiblement biaisé, ce qui est plus difficilement explicable.

Les conclusions tirées à partir de l'analyse du biais relatif s'appliquent toujours lorsqu'on analyse l'erreur quadratique moyenne. En effet, on remarque que les estimateurs qui sont très biaisés ont une forte tendance à avoir une erreur quadratique moyenne élevée et vice-versa.

6. DISCUSSION

Lorsque l'hypothèse d'un mécanisme de réponse non-ignorable est réaliste et que l'hypothèse de normalité des erreurs du modèle de régression linéaire (3.2) est justifiée, l'utilisation de la méthode du maximum de vraisemblance peut être appropriée. Par contre, lorsque cette dernière hypothèse n'est pas justifiée, les résultats de l'étude de simulation décrite à la cinquième section montrent que la méthode d'estimation robuste présentée dans cet article lui est préférable.

De plus, Beaumont (1999) présente les résultats d'une autre étude de simulation qui montrent que la méthode d'estimation proposée dans cet article est robuste autant par rapport à l'hypothèse de normalité des erreurs que par rapport au modèle (3.2). Quant à la méthode du maximum de vraisemblance, elle s'est avérée encore plus sensible à la validité du modèle (3.2) qu'à l'hypothèse de normalité des erreurs. Cette dernière méthode ne devrait donc être utilisée que lorsque toutes les hypothèses associées au modèle (3.2) et (3.3) sont raisonnables.

Évidemment, tous les estimateurs sont peu biaisés en présence d'un très faible taux de non-réponse. De même, lorsque le coefficient de corrélation entre X et Y est très élevé, tous les estimateurs sont faiblement biaisés, à l'exception de l'estimateur qui suppose un mécanisme de réponse uniforme $\hat{\mu}_{P,U}^{*}$. Dans l'un ou l'autre de ces cas, le choix d'un estimateur devrait être basé sur le critère de simplicité, qui favorise les estimateurs de la section (3.2) et plus particulièrement l'estimateur $\hat{\mu}_{I,X}^{*}$.

Il est à noter que les modèles (3.2) et (3.3) pourraient être complexifiés selon la nature du problème. Par exemple, on pourrait inclure d'autres variables indépendantes dans ces modèles. On pourrait également catégoriser la variable Y en se servant de variables indicatrices et utiliser ces variables indicatrices dans le modèle (3.3) au lieu de la variable Y elle-même.

Dans cet article, on s'est attardé uniquement au problème de l'estimation d'une moyenne lorsque le mécanisme de réponse n'est pas ignorable. Les méthodes présentées à la troisième et à la quatrième section sont toutefois applicables à d'autres types d'estimation. Par exemple, on pourrait utiliser les poids ou les valeurs imputées pour l'estimation des paramètres d'une régression quelconque.

L'objectif de cet article a consisté à présenter une méthode d'estimation robuste par rapport à l'hypothèse de normalité des erreurs du modèle (3.2), qui permet de réduire le biais dû à un mécanisme de réponse qui n'est pas ignorable. Dans un travail futur, il serait intéressant d'évaluer des méthodes simples d'estimation de la variance en présence de données imputées et lorsqu'on utilise cette méthode d'estimation robuste.

REMERCIEMENTS

L'auteur tient à remercier la division des données régionales et administratives de Statistique Canada qui a rendu possible ce travail. Il tient aussi à remercier Eric Rancourt, les deux arbitres et le rédacteur associé pour leurs commentaires utiles qui ont permis d'améliorer la qualité de cet article.

BIBLIOGRAPHIE

BEAUMONT, J.-F., et DEMNATI, A. (1998). Estimation des paramètres d'un mélange fini de distributions pour des données longitudinales dichotomiques; une comparaison d'algorithmes. *Recueil: Symposium 98, Analyse longitudinale pour les enquêtes complexes, Statistique Canada*, 207-213.

BEAUMONT, J.-F. (1999). A robust estimation method in the presence of nonignorable nonresponse. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. (À paraître).

D'AGOSTINO, R.B. (1986). Graphical analysis. *Goodness-of-fit Techniques*, (Éd. R.B. D'Agostino et M.A. Stephens), 7-62. New York: Marcel Dekker.

DEMPSTER, A.P., LAIRD, N.M., et RUBIN, R.B. (1977). Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society B*, 39, 1-38.

GAGNON, F., LEE, H., RANCOURT, E., et SÄRNDAL, C.-E. (1996). Estimating the variance of the generalized regression estimator in the presence of imputation for the Generalized Estimation System. *Recueil 1996 de la section des méthodes d'enquête, Société statistique du Canada*, 151-156.

GREENLEES, J.S., REECE, W.S., et ZIESCHANG, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.

LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

RANCOURT, E., LEE, H., et SÄRNDAL, C.-E. (1994). Corrections du biais pour des estimations d'enquête tirées de données comprenant des valeurs imputées par quotient par suite d'une non-réponse selon un mécanisme confondu. *Techniques d'enquête*, 20, 143-153.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SAS INSTITUTE INC. (1990). *SAS/STAT User's Guide*, 2, Version 6, 4e édition, Cary, NC: SAS Institute Inc.

ZEGER, S.L., LIANG, K., et ALBERT, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44, 1049-1060.

Techniques d'enquête, décembre 2000
Vol. 26, n° 2, pp. 153-155
Statistique Canada

Un effet de plan de sondage approximatif pour une pondération inégale en cas de corrélation possible entre les mesures et les probabilités de sélection

BRUCE D. SPENCER[1]

RÉSUMÉ

On estime couramment l'effet de plan de sondage attribuable à la pondération par 1 plus la variance relative des poids de l'échantillon. Cette formule a été justifiée en l'absence de corrélation entre les probabilités de sélection et la variable d'intérêt. Une approximation de l'effet de plan de sondage est présentée pour tenir compte de la présence d'une corrélation.

MOTS CLÉS: Pondération; effet de plan de sondage; variance d'échantillonnage; échantillons complexes.

1. INTRODUCTION

Une pratique courante consiste à pondérer les observations dans un échantillon à probabilités inégales par les valeurs inverses des probabilités de sélection. On le fait parce que l'absence des poids entraîne une erreur systématique s'il y a une corrélation entre les poids d'échantillonnage et la variable d'intérêt. Un inconvénient du processus de pondération est l'augmentation de la variance d'échantillonnage si les poids varient de façon excessive dans l'échantillon. Cette augmentation peut être quantifiée à l'aide de l'effet de plan de sondage. Celui-ci est le rapport entre la variance de la statistique d'intérêt pour le plan de sondage en question et la variance de la statistique pour un échantillonnage aléatoire simple et la même taille d'échantillon (Kish 1965). Les effets de plan de sondage sont importants tant pour l'approximation des erreurs types une fois que l'échantillon est disponible que pour la prédiction préalable des erreurs types, étape critique d'un plan de sondage efficace.

Kish (1965, 1992) a décrit une approximation de l'effet de plan de sondage pour des estimations pondérées tirées d'échantillons à probabilités inégales: $1 + rvw$, où rvw est défini comme la variance relative des poids de l'échantillon. Si donc w_i est le poids de l'unité i de l'échantillon, et si $\bar{w}$ est la moyenne de l'échantillon, $rvw = n^{-1}\sum_{i=1}^{n}(w_i - \bar{w})^2/\bar{w}^2$. Gabler, Haeder et Lahiri (1999) ont utilisé un modèle de superpopulation pour obtenir un effet de plan de sondage en présence d'une mise en grappes également. Leur formule, qui correspond à des résultats fondés sur un plan de sondage figurant dans Kish (1965), se réduit à $1 + rvw$ en présence d'une corrélation intraclasse nulle. L'approximation $1 + rvw$ pour l'effet de plan de sondage se fonde sur un modèle ou un plan dans lequel il n'y a pas de corrélation entre les poids et la variable d'intérêt (de sorte qu'une estimation non pondérée serait tout aussi bonne ou meilleure que l'estimation pondérée). Nous présentons ici une approximation de l'effet de plan de sondage pour un modèle dans lequel il peut y avoir une corrélation. Ce faisant, nous ne supposons pas que la population est tirée d'une superpopulation. L'exactitude de l'approximation dépend uniquement des caractéristiques du plan de sondage et de la population d'intérêt.

Par souci de simplicité, nous avons recours à l'échantillonnage à un degré à probabilités inégales avec remise. L'élargissement heuristique des résultats à l'échantillonnage sans remise est abordé à la section 4.

2. REPRÉSENTATION DE RÉGRESSION DU PLAN DE SONDAGE ET DE LA POPULATION

Soit y_i, la mesure d'intérêt, P_i, la probabilité de sélection (d'un tirage à l'autre) pour un échantillon de taille n, et $w_i = 1/(nP_i)$, le poids d'échantillonnage pour l'unité i d'une population de taille N, $1 \le i \le N$. À noter que $\bar{P} = \sum_{i=1}^{N} P_i/N = N^{-1}$. Considérons la droite de régression des moindres carrés

$$y_i = \alpha + \beta P_i + \varepsilon_i, \qquad (1)$$

où $\alpha = \bar{Y} - \beta/N$, $\beta = \sum_{i=1}^{N}(y_i - \bar{Y})(P_i - \bar{P})/\sum_{i=1}^{N}(P_i - \bar{P})^2$ et $\bar{Y} = \sum_{i=1}^{N} y_i/N$. Nous notons les variances de population des y, des ε, des ε^2 et des w par $\sigma_y^2, \sigma_\varepsilon^2, \sigma_{\varepsilon^2}^2$, et σ_w^2 avec, par exemple, $\sigma_y^2 = \sum_{i=1}^{N}(y_i - \bar{Y})^2/N$. Nous notons la corrélation de population entre y et P par $\rho_{y,P}$, entre ε et w par $\rho_{\varepsilon,w}$ et entre ε^2 et w par $\rho_{\varepsilon^2,w}$. Il s'ensuit des propriétés des moindres carrés, ou de façon équivalente des définitions de α et de β, que $\sum_{i=1}^{N}\varepsilon_i P_i = \sum_{i=1}^{N}\varepsilon_i/N = 0$ et que $\sigma_\varepsilon^2 = (1 - \rho_{y,P}^2)\sigma_y^2$. Si des données sont disponibles, nous pouvons ajuster la représentation de régression (1) et estimer $\alpha, \beta, \sigma,$ et ρ, par exemple, par $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}$ et $\hat{\rho}$.

Soit $\hat{Y} = \sum_{i=1}^{n} w_i y_i$, l'estimateur pondéré habituel du total de la population, Y. La variance de $\hat{Y}$ est bien connue (Cochran 1977, 253):

[1] Bruce D. Spencer, Department of Statistics and Institute for Policy Research, 2006 Sheridan Road, Northwestern University, Evanston, IL 60208, U.S.A.

$$V(\hat{Y}) = n^{-1}\sum_{i=1}^{N} P_i(y_i/P_i - Y)^2. \quad (2)$$

À l'aide de la formulation de régression (1), nous pouvons réexprimer la variance comme suit

$$V(\hat{Y}) = \alpha^2 N(\bar{W} - N/n) + (1 - \rho_{y,P}^2)\sigma_y^2 N\bar{W} + N\rho_{\varepsilon^2,w}\sigma_{\varepsilon^2}\sigma_w + 2\alpha N\rho_{\varepsilon,w}\sigma_\varepsilon\sigma_w, \quad (3)$$

où $\bar{W} = \sum_{i=1}^{N} w_i/N$.

Cette expression ne se fonde sur aucune hypothèse au sujet de l'ajustement du modèle de régression (voir le calcul à la section 5).

Si le modèle de régression est assez bien ajusté pour que $\rho_{\varepsilon^2,w}$ et $\rho_{\varepsilon,w}$ soient nuls, la variance en (3) est simplifiée de façon à donner $V(\hat{Y}) = \alpha^2 N(\bar{W} - N/n) + (1 - \rho_{y,P}^2)\sigma_y^2 N\bar{W}$. Si l'on avait eu recours à un échantillonnage aléatoire simple avec remise, la variance aurait été $n^{-1}N^2\sigma_y^2$. Par conséquent, si $\rho_{\varepsilon^2,w}$ et $\rho_{\varepsilon,w}$ sont négligeables, l'effet de plan de sondage est approximativement:

$$\text{deff} = (1 - \rho_{y,P}^2)n\bar{W}/N + (\alpha/\sigma_y)^2(n\bar{W}/N - 1). \quad (4)$$

Cette approximation n'exige pas que les résidus de la régression soient négligeables, et elle reste valable lorsque σ_ε est grand. Un examinateur a fait remarquer que la condition voulant que $\rho_{\varepsilon^2,w}$ et $\rho_{\varepsilon,w}$ soient négligeables peut sembler anormale dans un modèle qui régresse y sur P plutôt que sur $w \propto 1/P$. À noter, toutefois, qu'en présence non seulement d'une corrélation nulle entre ε et P mais également d'indépendance, nous aurions une corrélation nulle entre les fonctions de ε et les fonctions de P, de sorte que $\rho_{\varepsilon^2,w}$ et $\rho_{\varepsilon,w}$ seraient nuls eux aussi.

3. ESTIMATION DE L'EFFET DE PLAN DE SONDAGE

Afin d'estimer l'effet de plan de sondage lorsque l'échantillon est disponible, nous pouvons utiliser $1 + rvw$ pour estimer $n\bar{W}/N$. Afin d'en comprendre le bien-fondé, notons d'abord que

$$1 + rvw = \frac{n^{-1}\sum_{i=1}^{n} w_i^2}{\bar{w}^2}. \quad (5)$$

L'espérance du numérateur est $N\bar{W}/n$. L'espérance de $\bar{w}$ est N/n, et donc le dénominateur de (5) peut être considéré comme un estimateur de $(N/n)^2$. Le fait de diviser l'espérance du numérateur par $(N/n)^2$ nous permet d'obtenir $n\bar{W}/N$. Ainsi, l'effet de plan de sondage peut être estimé à partir de l'échantillon comme suit:

$$(1 - \hat{\rho}_{y,P}^2)(1 + rvw) + (\hat{\alpha}/\hat{\sigma}_y)^2(rvw). \quad (6)$$

Notons, à titre de cas spécial, que si nous fixons $\hat{\rho}_{y,P} = 0$, qui est le cas de la «pondération au hasard» (Kish 1992), l'estimation de l'effet de plan de sondage se réduit à

$$1 + rvw + rvw(\hat{\alpha}^2/\hat{\sigma}_y^2). \quad (7)$$

Cette estimation se rapproche de celle de Kish pour $\hat{\alpha}/\hat{\sigma}_y$ presque nul.

4. ÉCHANTILLONNAGE SANS REMISE

Le calcul de l'effet de plan de sondage exact pour un échantillonnage sans remise serait plus complexe, car il supposerait que l'on tienne compte des probabilités de sélection composées pour des paires d'unités. Un élargissement heuristique des résultats, par contre, est facile. Rappelons que le rapport entre la variance d'une moyenne de l'échantillon pour un échantillonnage aléatoire simple sans remise et la variance pour un échantillonnage avec remise est approximativement $(1 - n/N)$.

Les résultats que nous avons calculés pour l'effet de plan de sondage s'appliquent à des échantillons à un degré à probabilités inégales de n unités sans remise si la variance de l'estimateur de Horvitz-Thompson du total est approximativement $(1 - n/N)$ fois la variance en (2), P_i étant considéré comme n^{-1} fois la probabilité de sélection globale pour l'unité i (Särndal, Swensson et Wretman 1992, 154).

5. CALCUL DE LA FORMULE DE LA VARIANCE (3)

À partir de (2) nous avons $V(\hat{Y}) = n^{-1}(\sum_{i=1}^{N} y_i^2/P_i - Y^2)$. Ensuite, nous notons que (1) suppose que

$$Y^2 = (N\alpha + \beta)^2 = N^2\alpha^2 + 2N\alpha\beta + \beta^2 \quad (8)$$

et que

$$\sum_{i=1}^{N} y_i^2/P_i = \sum_{i=1}^{N} [\alpha^2/P_i + \beta^2 P_i + \varepsilon_i^2/P_i + 2\alpha\beta + 2\alpha\varepsilon_i/P_i + 2\beta\varepsilon_i]$$

$$= \alpha^2\sum_{i=1}^{N} P_i^{-1} + \beta^2 + \sum_{i=1}^{N} \varepsilon_i^2/P_i + 2N\alpha\beta + 2\alpha\sum_{i=1}^{N} \varepsilon_i/P_i$$

$$= \alpha^2 n\sum_{i=1}^{N} w_i + \beta^2 + n\sum_{i=1}^{N} \varepsilon_i^2 w_i + 2N\alpha\beta + 2\alpha n\sum_{i=1}^{N} \varepsilon_i w_i. \quad (9)$$

Le fait de soustraire (8) de (9) et de diviser par n donne

$$V(\hat{Y}) = \alpha^2\left(\sum_{i=1}^{N} w_i - N^2/n\right) + \sum_{i=1}^{N} \varepsilon_i^2 w_i + 2\alpha\sum_{i=1}^{N} \varepsilon_i w_i.$$

Afin d'obtenir (3), nous notons que

$$\sum_{i=1}^{N} \varepsilon_i^2 w_i = N\rho_{\varepsilon^2,w}\sigma_{\varepsilon^2}\sigma_w + N\bar{W}\sigma_\varepsilon^2$$

$$= N\rho_{\varepsilon^2,w}\sigma_{\varepsilon^2}\sigma_w + (1-\rho_{y,p}^2)\sigma_y^2 N\bar{W}$$

et

$$\sum_{i=1}^{N} \varepsilon_i w_i = N\rho_{\varepsilon,w}\sigma_\varepsilon\sigma_w.$$

BIBLIOGRAPHIE

COCHRAN, W. G. (1977). *Sampling Techniques*. 3e éd. New York: Wiley.

GABLER, S., HAEDER, S., et LAHIRI, P. (1999). Justification à base de modèle de la formule de Kish pour les effets de plan de sondage liés à la pondération et à l'effet de grappe. *Techniques d'enquête*, 25, 1, 119-120.

KISH, L. (1965). *Survey Sampling*. New York: Wiley.

KISH, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, 8, 2, 183-200.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Techniques d'enquête, décembre 2000
Vol. 26, n° 2, pp. 157-171
Statistique Canada

Validité de l'analyse markovienne de structure latente pour l'estimation de l'erreur de classification des données sur la population active

PAUL P. BIEMER et JOHN M. BUSHERY[1]

RÉSUMÉ

Les auteurs ont surtout cherché à vérifier la validité des estimations MLCA (analyse markovienne de structure latente) de l'erreur de classification de la population active et à évaluer la possibilité que ce type d'analyse remplace les méthodes traditionnelles d'évaluation de la qualité des données. Les auteurs ont analysé des données d'interview de la Current Population Survey (CPS, enquête sur la population active) pour les trois premiers mois de chacune de trois années (1993, 1995, 1996) et ils ont mené une analyse supplémentaire des données de réinterview non rapprochées de la CPS pour à peu près les mêmes périodes de temps. Les données de réinterview représentent une autre stratégie d'estimation de l'erreur de classification de la CPS qui, si on la compare aux estimations de type MLCA, aide à examiner la validité de la stratégie MLCA. Les auteurs abordent cinq dimensions de la validité de la stratégie MLCA: a) diagnostics modélisés, b) validité de l'ajustement du modèle pour trois années de la CPS, c) accord entre les estimations de type modèle et de type réinterview test-retest des probabilités de réponse, d) accord entre les estimations de type modèle et de type réinterview test-retest du manque de cohérence et e) plausibilité des profils d'erreur de classification. De plus, les auteurs examinent la robustesse des estimations MLCA à l'égard d'infractions à l'hypothèse markovienne. La présente analyse ne fournit aucune raison de mettre en doute la validité de la stratégie MLCA. La méthode a donné de bons résultats pour les cinq épreuves de validité.

MOTS CLÉS: Enquêtes par panel; erreur non due à l'échantillonnage; chômage; qualité des données.

1. INTRODUCTION

La Current Population Survey (CPS, enquête sur la population active) est une enquête sur échantillon menée mensuellement auprès des ménages par le Bureau of the Census des États-Unis en vue de la préparation d'estimations de l'emploi, du chômage et d'autres caractéristiques de la population active des États-Unis. Des estimations nationales tirées de la CPS pour ce qui est de la taille, de la composition et de l'évolution de la composition de la population active sont publiées chaque mois par le Bureau of Labor Statistics in Employment and Earnings des États-Unis. Les estimations de la population active tirées de la CPS sont l'un des indicateurs économiques clés du pays; depuis 1942, l'administration fédérale fait appel aux séries de données de la CPS pour suivre l'évolution mensuelle et annuelle de la population active.

Compte tenu de l'importance des séries de données de la CPS pour la politique gouvernementale, l'exactitude des données a été évaluée à plusieurs reprises. Ainsi, depuis le début des années 1950, le Bureau of the Census a dirigé le programme de réinterview de la CPS afin d'évaluer la qualité des données sur la population active. Dans le cadre de ce programme, on tire un petit sous-échantillon (moins de 5%) des répondants de la CPS et on pose de nouveau certaines questions de l'interview originale, notamment des questions sur la population active. Jusqu'en 1994, on a soumis un quart environ de l'échantillon à une réinterview non rapprochée et les trois quarts restants à une réinterview rapprochée. Le volet réinterview rapprochée, utilisé surtout en vue du contrôle de la qualité des interviews, a été éliminé en 1994, à cause de préoccupations au sujet de la qualité des données. Toutefois, la réinterview non rapprochée se poursuit encore aujourd'hui et sert à l'estimation de la stabilité (ou de la cohérence des réponses). On trouvera dans Forsman et Schreiner (1991) une description détaillée du programme de réinterview de la CPS.

Plusieurs documents préparés par des chercheurs à l'extérieur du Bureau of the Census ont analysé les données du programme de réinterview de la CPS afin d'en estimer l'erreur de classification (voir Sinclair et Gastwirth 1996, 1998; Biemer et Forsman 1992; Chua et Fuller 1987; Poterba et Summers 1986; Abowd et Zellner 1985). Récemment, Poterba et Summers (1995) ont utilisé des données du programme de réinterview de la CPS afin d'estimer les taux d'erreur de classification de la CPS et d'évaluer l'effet de l'erreur de classification sur les taux de changement de situation vis-à-vis de l'activité. Comme dans le document de 1986, l'analyse plus récente de ces auteurs se fonde sur l'hypothèse que le processus de rapprochement des réinterviews de la CPS fournit des données que l'on peut considérer comme véridiques. Abowd et Zellner (1985) ont adopté une stratégie semblable.

Plusieurs auteurs (par exemple Sinclair et Gastwirth 1996, 1998; Biemer et Forsman 1992; Forsman et Schreiner 1991; Schreiner 1980) ont mis en doute l'hypothèse selon

[1] Paul P. Biemer, Research Triangle Institute, Research Triangle Park, NC 27709 et John M. Bushery, Bureau of Transportation Statistics, Washington, DC 20590 U.S.A.

laquelle les réinterviews rapprochées fournissent des valeurs véridiques. Ces auteurs présentent d'importantes indications que les données de réinterview se prêtent à des erreurs de classification appréciables. En réalité, cette prise de conscience a été responsable de l'élimination par le Bureau of the Census du volet réinterview rapprochée du programme de réinterview de la CPS en 1994.

Comme substitut de l'hypothèse d'infaillibilité, Chua et Fuller (1987) et Fuller et Chua (1985) ont appliqué un type de modèle de structure latente aux données de réinterview rapprochée de la CPS afin d'estimer les probabilités de réponse à la CPS. Par souci d'identifiabilité du modèle, ces auteurs imposent de rigoureuses restrictions aux probabilités de réponse, forçant le biais dû à l'erreur de classification à être nul tant pour l'interview que pour la réinterview. De plus, ils supposent des erreurs de classification indépendantes pour l'interview et la réinterview (on parle d'hypothèse ICE dans la documentation) et d'un mois à l'autre dans l'échantillon. L'hypothèse ICE représente une limite de leur analyse puisque certains documents indiquent que l'hypothèse n'est peut-être pas valable pour la CPS (voir par exemple O'Muircheartaigh 1991, Singh et Rao 1995). Par conséquent, il se peut que les probabilités de réponse estimées à l'aide de la stratégie de Chua et Fuller soient biaisées.

Sinclair et Gastwirth (1996) et Sinclair et Gastwirth (1998) ont appliqué une stratégie de modélisation de classe latente aux données d'interview-réinterview de la CPS en fonction de restrictions de modèle proposées originalement par Hui et Walter (1980) pour des épreuves diagnostiques médicales. À l'aide de données d'interview-réinterview recoupées selon le sexe, Sinclair et Gastwirth ont supposé que les probabilités d'erreur de classification sont égales pour les hommes et les femmes, tandis que les taux d'activité de la population active sont différents pour ces groupes. Puisque les paramètres de modèle absorbent tous les degrés de liberté disponibles pour l'estimation des paramètres, aucun degré de liberté résiduel n'est disponible pour la vérification du manque d'ajustement du modèle. Par conséquent, leur analyse n'indique pas directement si ces hypothèses au sujet du modèle sont valables pour les données de la CPS.

Dans une étude des déterminants du biais dû aux groupes de renouvellement, Shockey (1988) a également appliqué une analyse de structure latente à la CPS. Son analyse a indiqué que le biais dû aux groupes de renouvellement signalé d'abord par Bailar (1975) est peut-être causé par l'erreur de réponse suscitée dans le cadre de l'interview. Shockey n'a pas utilisé de données de réinterview, se fondant plutôt sur des méthodes confirmatoires d'analyse par facteurs pour appuyer ses affirmations. Ses taux d'erreur étaient beaucoup plus élevés que les taux signalés par d'autres auteurs, ce qui indique peut-être un biais dû au modèle. Malheureusement, comme pour Sinclair et Gastwirth, les données de Shockey ne permettent pas une vérification complète des hypothèses du modèle utilisé.

La méthode de l'analyse markovienne de structure latente, une stratégie prometteuse d'estimation de l'erreur de classification des données d'enquête par panel, n'avait pas encore été appliquée à la CPS. Cette méthode tire avantage de la nature répétitive des enquêtes par panel afin d'extraire de l'information sur l'erreur de classification directement des données d'interview. Le modèle MLCA est en réalité une combinaison de deux modèles: un modèle de chaîne markovienne latente qui représente les changements d'un mois à l'autre de la véritable situation vis-à-vis de l'activité, et un modèle d'erreur de classification qui représente les écarts par rapport à la situation véritable et observée vis-à-vis de l'activité.

Puisque la stratégie MLCA tire avantage de la nature répétitive des enquêtes par panel en vue de l'extraction d'informations sur l'erreur de classification directement des données d'interview, elle n'exige aucune mesure infaillible externe ni aucune mesure obtenue à l'aide de méthodes de réinterview. À cet égard, la méthode offre certains avantages par rapport aux méthodes traditionnelles du Bureau of the Census et aux méthodes de Chua et Fuller, Abowd et Zellner, Poterba et Summers, et Sinclair et Gastwirth pour ce qui est de l'évaluation de la qualité des données d'enquête. Dans de nombreuses enquêtes par panel, les réinterviews ne sont pas possibles à cause de restrictions budgétaires, de la complexité des travaux sur place et du fardeau de réponse. La stratégie MLCA est peut-être la seule façon d'évaluer l'erreur de mesure de ces enquêtes. Pour ce qui est des enquêtes par panel comme la CPS, pour lesquelles des données de réinterview sont disponibles, les méthodes de réinterview et MLCA offrent une autre stratégie analytique d'évaluation de l'erreur de classification. Ainsi, comme dans la présente analyse, la méthode MLCA peut servir à modéliser et à vérifier les hypothèses traditionnelles d'analyse des réinterviews. De plus, la stratégie MLCA fournit un schéma statistique pour la combinaison des données recueillies au moyen d'un panel et des données de réinterview en vue de l'obtention d'une information encore plus riche au sujet de l'erreur de classification (van de Pol et Langeheine 1997).

Un autre avantage de la stratégie MLCA est la possibilité d'intégrer l'ensemble des données recueillies au moyen d'un panel aux estimations de l'erreur de classification au lieu de se limiter à l'échantillon relativement petit retenu pour la réinterview. Par conséquent, certains aspects de la qualité des données des enquêtes par panel qu'il n'était pas possible d'examiner jusqu'à présent à cause d'un manque de données sont peut-être abordables maintenant.

Le présent exposé décrit nos résultats pour ce qui est de l'utilité de la stratégie de modélisation MLCA pour l'évaluation de l'erreur de classification de la population active dans le cadre de la CPS. Les logiciels servant à ajuster tout un choix de modèles de structure latente de type MLCA ou autre sont accessibles de différentes sources. Le logiciel utilisé dans notre analyse est ℓEM (Vermunt 1997),

et il permet d'ajuster une grande classe de modèles linéaires logarithmiques en présence ou en l'absence de variables latentes. Grâce à sa souplesse et à sa généralité, ce logiciel permet à l'analyste de vérifier une gamme considérable de modèles d'erreur de classification et d'examiner les hypothèses au sujet de la cause et des corrélations de l'erreur de classification.

À la section suivante, nous décrivons le modèle MLCA ainsi que la méthode d'estimation et ses fondements théoriques. À la section 3, il est question de la méthode MLCA appliquée à la CPS, et nous ajustons une série de modèles à la CPS tout en examinant l'ajustement de ces modèles. Dans la présente section, nous préparons également des estimations de l'erreur de classification en fonction du meilleur modèle MLCA. À la section 4, nous vérifions de différentes façons la validité des estimations MLCA, notamment en comparant celles-ci aux estimations d'une nouvelle analyse interview-réinterview. Enfin, à la section 5, nous résumons nos résultats et formulons des recommandations au sujet de l'utilité de la méthode MLCA pour de futures évaluations de l'erreur de classification de la situation vis-à-vis de l'activité.

2. ANALYSE MARKOVIENNE DE STRUCTURE LATENTE POUR TROIS PÉRIODES DE TEMPS

Les modèles markoviens de structure latente ont d'abord été proposés par Wiggins (1973) et perfectionnés par Poulsen (1982). Van de Pol et de Leeuw (1986) ont établi les conditions dans lesquelles le modèle est identifiable, et ils ont décrit d'autres conditions de l'estimabilité des paramètres de modèle. Dans la présente section, nous élaborons le modèle MLCA dans le contexte de la CPS, et nous en suggérons d'autres applications de même que la généralisation.

Nous divisons la population cible de la CPS en groupes L (âge, race, sexe, par exemple). Soit G, une variable qui désigne la composition du groupe. Ainsi, $G_i = 1$ si le i-ième membre de la population se trouve dans le groupe 1, $G_i = 2$ pour le groupe 2 et ainsi de suite. Soit X_{gi}, Y_{gi}, et Z_{gk}, les véritables classifications de la situation vis-à-vis de l'activité pour la i-ième personne du groupe $G = g$ (pour $g = 1, \ldots, L$ et $i = 1, \ldots, n_g$) où X_{gi} se définit comme

$$X_{gi} = \begin{cases} 1 & \text{si la personne } (g, i) \text{ est employée au cours de la période 1} \\ 2 & \text{si la personne } (g, i) \text{ est sans emploi au cours de la période 1} \\ 3 & \text{si la personne } (g, i) \text{ est inactive au cours de la période 1} \end{cases}$$

avec des définitions semblables pour Y_{gi} et Z_{gi} au cours des périodes 2 et 3 respectivement. Soit $\pi_{x,y,z|g}$, $\Pr(X=x, Y=y, Z=z|G=g)$, $\pi_{y|g,x}$, $\Pr(Y=y|X=x, G=g)$ et $\pi_{z|g,y,x}$, $\Pr(Z=z|Y=y, X=x, G=g)$. La probabilité qu'un membre du groupe g se trouve dans une situation d'activité x au cours de la période 1, y au cours de la période 2 et z au cours de la période 3 est

$$\pi_{x,y,z|g} = \pi_{x|g}\,\pi_{y|g,x}\,\pi_{z|g,x,y}. \qquad (1)$$

Enfin, en vertu de l'hypothèse markovienne d'ordre un, une condition nécessaire de l'identifiabilité du modèle (voir van de Pol et de Leeuw 1986), nous supposons

$$\pi_{z|g,x,y} = \pi_{z|g,y} \qquad (2)$$

c'est-à-dire qu'au cours de la période 3, la véritable situation d'une personne ne dépend pas de la situation au cours de la période 1, une fois la situation au cours de la période 2 connue. Une autre interprétation est que la situation courante, compte tenu de la situation au cours de la période précédente, ne dépend pas du changement de la période précédente.

On peut concevoir un certain nombre de scénarios dans lesquels l'hypothèse markovienne n'est peut-être pas valable pour la situation d'activité mensuelle. L'hypothèse serait enfreinte, par exemple, si des personnes sans emploi au cours de la période 2 avaient plus de chances d'être sans emploi au cours de la période 3, étant donné qu'elles étaient également sans emploi au cours de la période 1. Le groupe de personnes sans emploi au cours de la période 2 et de la période 1 englobe probablement une plus forte proportion de chômeurs chroniques que le groupe de personnes sans emploi au cours de la période 2 mais non au cours de la période 1. Ce groupe (sans emploi au cours de la période 2 mais non pas de la période 1) compte probablement une plus forte proportion de chômeurs occasionnels en train de changer d'emploi.

Toutefois, la validité de cette hypothèse ne peut pas être examinée convenablement à l'aide des données observées parce que les données comportent une distorsion dont l'ampleur est inconnue à cause de la présence d'erreurs de classification. Il existe au moins deux méthodes d'évaluation de la validité de l'hypothèse markovienne pour des données recueillies au moyen d'un panel. Van de Pol et de Leeuw (1986) ont suggéré une méthode fondée sur quatre cycles de données recueillies au moyen d'un panel qui substitue une restriction markovienne d'ordre 2 à la restriction d'ordre un en (2). Une autre méthode, suggérée par van de Pol et Langeheine (1997), fait appel à une combinaison de données de type panel sur la situation d'activité et de données de réinterview pour chaque période de temps. Ni l'une ni l'autre de ces méthodes n'a été utilisée dans le présent exposé pour vérifier directement l'hypothèse de la MLCA. Nous avons plutôt évalué la validité globale des estimations MLCA à l'aide des méthodes décrites à la section 3.2 ci-dessous. À la section 3.6, nous présentons des résultats d'une étude de simulation afin d'illustrer la robustesse des estimations MLCA de l'erreur de classification vis-à-vis des infractions à l'hypothèse markovienne.

Nous considérons maintenant les classifications observées de la situation vis-à-vis de l'activité relativement à la CPS notées A_{gi}, B_{gi}, et C_{gi} pour les périodes 1, 2 et 3, respectivement, où

$$A_{gi} = \begin{cases} 1 & \text{si la personne } (g, i) \text{ est classée comme employée au cours de la période 1} \\ 2 & \text{si la personne } (g, i) \text{ est classée comme sans emploi au cours de la période 1} \\ 3 & \text{si la personne } (g, i) \text{ est classée comme inactive au cours de la période 1} \end{cases}$$

avec des définitions semblables pour les indicateurs de réponse B_{gi}, et C_{gi} pour les périodes 2 et 3, respectivement. À l'aide d'une extension de la notation établie ci-dessus, nous désignons les probabilités de réponse pour chacune de ces classifications par $\pi_{a|g,x} = \Pr(A = a \mid G = g, X = x)$, avec des définitions semblables pour $\pi_{b|g,y}$ et $\pi_{c|g,z}$. Ainsi, $\pi_{a=1|g,x=2}$ est la probabilité qu'une personne dans le groupe g soit classée par la CPS comme employée ($A = 1$) lorsqu'elle est en réalité sans emploi ($X = 2$). De même, $\pi_{a=2|g,x=2}$ est la probabilité qu'une personne dans le groupe g soit classée correctement comme sans emploi.

Enfin, nous supposons

$$\pi_{a,b,c|g,x,y,z} = \pi_{a|gx}\pi_{b|gy}\pi_{c|gz} \tag{3}$$

ou que l'erreur de classification de la situation observée vis-à-vis de l'activité est indépendante de l'un à l'autre des trois mois. Cette hypothèse, appelée l'hypothèse de l'indépendance locale, a été examinée relativement à la CPS par Meyers (1988) dans son étude de la stratégie d'estimation utilisée par Abowd et Zellner (1985). Meyers a conclu que l'hypothèse semble représenter une approximation raisonnable. Singh et Rao (1995), qui ont étudié la robustesse de l'hypothèse en fonction d'un certain nombre de scénarios de la population active, ont tiré une conclusion semblable. van de Pol et Langeheine (1997) ont modélisé la distribution combinée des données recueillies au moyen d'un panel et des données de type réinterview à l'aide de modèles de structure latente afin de vérifier l'indépendance locale de divers types de changements de situation vis-à-vis de l'activité. Ils ont trouvé certaines indications que les personnes qui changent de situation vis-à-vis de l'activité manifestent une fiabilité inférieure à celle des personnes qui ne changent pas de situation, l'effet étant toutefois assez faible. Par conséquent, nous allons également supposer (3) sans tenter d'en approfondir davantage la validité dans le présent exposé.

Les classifications de la CPS pour ce qui est de la situation vis-à-vis de l'activité au cours de chaque mois du premier trimestre de l'année sont les variables résultantes de notre analyse. Soit A, B et C, les classifications observées, et X, Y et Z, les véritables classifications (non observées) pour janvier, février et mars, respectivement. Soit G, une variable de groupement (ou stratification) quelconque à définir ultérieurement au cours de l'analyse. Compte tenu de ces hypothèses, nous pouvons formuler la probabilité qu'un membre de l'échantillon de la CPS soit classé dans la cellule (g, a, b, c) du tableau $GABC$ comme suit:

$$\pi_{g,a,b,c} = \sum_{x,y,z} \pi_g \pi_{x|g} \pi_{a|g,x} \pi_{y|x,g} \pi_{b|y,g} \pi_{z|g,y} \pi_{c|g,z}. \tag{4}$$

Les extensions à plus d'une variable de groupement se font directement.

Dans le cadre d'un échantillonnage multinomial, la fonction de vraisemblance pour le tableau $GABC$ est

$$\Pr(GABC) = k \prod_{g,a,b,c} \pi_{g,a,b,c}^{n_{gabc}} \tag{5}$$

où k est la constante multinomiale et Π désigne le produit des termes pour les indices g, a, b et c. En vertu des hypothèses énoncées antérieurement, les paramètres de modèle peuvent être estimés à l'aide de méthodes d'estimation du maximum de vraisemblance. Van de Pol et de Leeuw (1986) ont fourni la formule permettant d'appliquer l'algorithme E-M afin d'estimer les paramètres de ce modèle, et ils ont décrit les conditions de leur estimabilité. Le logiciel ℓEM, appliqué aux ensembles de données de la CPS à la section suivante, assure la mise en œuvre de ces méthodes.

3. APPLICATION À LA CPS

3.1 Notation

Une partie de notre évaluation de la stratégie MLCA est une comparaison des estimations MLCA de l'erreur de classification aux estimations tirées de l'analyse des données d'interview-réinterview. Nous utilisons la notation décrite à la section précédente; soit A et A', la classification de la situation vis-à-vis de l'activité pour l'original et pour la réinterview, respectivement; nous définissons $\pi_a = \Pr(A = a)$ et $\pi_{a'} = \Pr(A' = a')$. Soit AA', le tableau de tri croisé des interviews-réinterviews observées $K \times K$; soit $\boldsymbol{\pi}^{AA'|X}$, la matrice $K \times K$ des probabilités de cellule, $\Pr(A = a, A' = a' \mid X = x)$. Si nous supposons que $\pi_{aa'} = \Pr(A = a, A' = a' \mid X = x) = \pi^2_{a|x}$, appelée dans la documentation l'hypothèse des mesures parallèles (Bohrnstedt 1983), nous avons

$$\boldsymbol{\pi}^{AA'|X} = \boldsymbol{\pi}^{A|X}(\boldsymbol{\pi}^{A|X})^T \tag{6}$$

où $(\boldsymbol{\pi}^{A|X})^T$ désigne la transposée d'un vecteur de probabilités conditionnelles, $\boldsymbol{\pi}^{A|X}$.

Soit $\boldsymbol{\pi}^X$, le vecteur K des véritables probabilités de classification. Nous avons alors

$$\boldsymbol{\pi}^{AA'} = \boldsymbol{\pi}^{AA'|X}\boldsymbol{\pi}^X \tag{7}$$

c'est-à-dire que la probabilité du tableau de classification des interviews-réinterviews observées, $\boldsymbol{\pi}^{AA'}$, est égale au produit de la matrice des probabilités de réponse conditionnelles, $\boldsymbol{\pi}^{AA'|X}$, et du vecteur des véritables probabilités de classification, $\boldsymbol{\pi}^X$.

Comme nous l'avons expliqué à la section précédente, la MLCA des données longitudinales de la CPS fournira des estimations du maximum de vraisemblance de $\pi^{A|X}$ et de π^{X}, permettant l'estimation de $\pi^{AA'}$ à l'aide de (6) et de (7). Nous pouvons estimer la fiabilité test-retest, R, pour toute situation vis-à-vis de l'activité en appliquant les méthodes d'estimation habituelles (voir par exemple Bohrnstedt 1983) à cette estimation de $\pi^{AA'}$. Pour notre analyse, nous calculons l'indice d'incohérence, $I = 1 - R$, qui représente la mesure traditionnelle de la fiabilité des données sur la population active de la CPS (voir U.S. Bureau of the Census 1985). Soit I_a, l'indice d'incohérence pour la catégorie $A = a$. Un estimateur de I_a est

$$\frac{gdr}{2\hat{\pi}_a(1-\hat{\pi}_a)} \tag{8}$$

où *gdr* représente le taux de différence brut défini par

$$gdr_a = 2\sum_{a \neq a'} \hat{\pi}_{a,a'} \tag{9}$$

et où $\hat{\pi}_a$ et $\hat{\pi}_{a,a'}$ désignent des estimations de structure latente de π_a et de $\pi_{a,a'}$, respectivement.

Le Bureau of the Census des États-Unis (1985, 88-91) a fourni les formules des erreurs-types de même qu'une mesure globale de l'incohérence pour toutes les catégories K combinées, appelée l'indice global d'incohérence, I_{AG}. L'indice global est une mesure de la non-fiabilité au niveau des questions égale à $1 - \kappa$ (Hess, Singer et Bushery 2000) où κ est la mesure de la fiabilité kappa de Cohen (Cohen 1960) et une moyenne pondérée des indices au niveau des catégories.

Enfin, pour une estimation de π^X donnée, nous pouvons estimer le vecteur K des biais de mesure, notés β_A, associés aux catégories K de A à l'aide de l'identité

$$\beta_A = \pi^A - \pi^X. \tag{10}$$

3.2 Évaluation de la validité de la méthode MLCA

Le présent exposé a comme principal objectif d'évaluer la validité de la stratégie MLCA. Les études précédentes de la mesure de l'erreur de classification de la CPS n'ont pas complètement abordé la validité des stratégies d'estimation utilisées (Meyers 1988). Nous comptons déterminer si la stratégie MLCA est informative et utile pour l'étude de l'erreur de classification de la CPS. Plus précisément, nous espérons déterminer si les estimations modélisées des probabilités d'erreur, $\pi^{A|X}$, reflètent les véritables niveaux d'erreur des classifications de la situation vis-à-vis de l'activité que l'on trouve dans la CPS. Malheureusement, pour des raisons mentionnées antérieurement, il n'existe aucun étalon généralement accepté pour l'évaluation de l'exactitude de la CPS (voir par exemple Sinclair et Gastwirth 1996, 1998, Biemer et Forsman 1992, ainsi que Schreiner 1980). Par conséquent, il n'est pas possible d'estimer le biais des estimations MLCA.

Dans la suite du texte, nous examinons la validité des estimations MLCA de l'erreur de classification de la CPS à l'aide de cinq critères:

1. **Diagnostics modélisés.** Une condition nécessaire de la validité d'un modèle est que celui-ci soit plausible (que les hypothèses soient raisonnables et concordent avec la réalité) et qu'il ajuste les données de façon adéquate. Nous utilisons le critère traditionnel chi-carré de la validité de l'ajustement ainsi que d'autres mesures diagnostiques de l'ajustement du modèle afin d'évaluer le bien-fondé de celui-ci et la mesure dans laquelle les données concordent avec lui.

2. **Validité de l'ajustement du modèle d'une année à l'autre de la CPS.** Une méthode répandue de validation d'un modèle consiste à évaluer l'ajustement du modèle pour des données qui sont indépendantes des données utilisées pour la modélisation (voir par exemple Kleinbaum, Kupper et Muller 1988, 330). Cette méthode permet d'éviter un surparamétrage du modèle et une sélection de modèle guidée par les données (plutôt que par la théorie). Dans la présente étude, l'ajustement du même modèle aux données de chaque année séparément est un aspect de cette technique de vérification indépendante du modèle. L'accord du modèle d'une année à l'autre aurait tendance à confirmer la validité de sa structure. Cette méthode comporte une difficulté pour la présente application. Après 1993, le questionnaire imprimé de la CPS a été remanié en fonction de l'IPAO (interview sur place assistée par ordinateur), de sorte que l'ampleur des erreurs de réponse a pu changer après 1993. Toutefois, si les principales sources d'erreur de réponse de la CPS n'ont pas changé sous l'effet du remaniement, une structure modélisée qui décrit convenablement l'erreur pour 1993 devrait également décrire l'erreur pour 1995 et 1996.

3. **Accord entre les estimations MLCA et les estimations test-retest de Hui-Walter pour les probabilités de réponse.** La méthode de Hui-Walter (H-W) (Hui et Walter 1980) d'estimation des probabilités de réponse de la CPS fait appel à des données de réinterview non rapprochée (Sinclair et Gastwirth 1996; 1998). Même si les méthodes MLCA et H-W font toutes deux appel à des modèles de structure latente, les hypothèses de modélisation sont très différentes. Par exemple, la méthode H-W n'exige pas d'hypothèse markovienne pour l'identifiabilité du modèle. De plus, dans la présente recherche, les entrées de données pour la méthode H-W sont indépendantes de celles utilisées pour la méthode MLCA. Le bon accord entre les deux ensembles d'estimations confirme la validité des deux méthodes, tandis qu'un faible accord indique qu'une des stratégies au moins n'est pas valable. Une forte concordance entre les estimations MLCA et H-W fournit également une certaine assurance que les estimations MLCA des probabilités de réponse sont relativement robustes vis-à-vis d'infractions éventuelles à l'hypothèse markovienne.

4. **Accord entre les estimations modélisées et test-retest de l'indice d'incohérence.** Ce critère est semblable au troisième critère car il compare des estimations tirées de la MLCA à des estimations fondées sur des données de réinterview non rapprochée. Toutefois, cette analyse ne fait pas appel à la validité de la méthode d'estimation de Hui-Walter pour l'évaluation de la validité des estimations MLCA. Nous utilisons plutôt les estimations MLCA de l'erreur de classification pour obtenir des estimations de l'indice d'incohérence à l'aide de (7) à (9). Nous comparons ces estimations de la fiabilité directement aux estimations de la fiabilité tirées du programme de réinterview de la CPS, obtenues de données de réinterviews non rapprochées. Un bon accord entre les estimations de réinterview et de type MLCA confirme la validité des deux méthodes, tandis qu'un faible accord indique qu'une des stratégies au moins n'est pas valable.
5. **Plausibilité des profils d'erreur de classification.** Enfin, la plausibilité (ou la validité apparente) des estimations des probabilités de réponse peut également servir d'épreuve de la validité. Par exemple, il semble peu plausible que des réponses par procuration à des questions sur la population active soient plus exactes que des réponses sans procuration. D'autres profils d'erreur de classification peuvent être examinés et évalués en fonction de la plausibilité. Pour autant que les estimations modélisées semblent plausibles, leur validité apparente est confirmée.

À la section suivante, nous discutons de nos résultats de modélisation MLCA dans le contexte de ces critères de validité. Nous commençons par décrire les ensembles de données de la CPS et les résultats du processus de sélection du modèle.

3.3 Ensembles de données de la CPS

En 1994, dans le cadre de la mise en œuvre de l'IPAO (interview sur place assistée par ordinateur), la CPS a été remaniée de fond en comble, avec restructuration des questions servant à déterminer la situation vis-à-vis de l'activité. Rothgeb (1994) a décrit le remaniement de la CPS. Comme conséquence de ces améliorations, nous nous attendons à observer une différence (plus précisément une réduction) de l'erreur de classification de la CPS après 1994 comparativement à 1993. Bien qu'il ne s'agisse pas d'un objectif principal de la présente recherche, nous avons comparé l'erreur de la CPS avant et après le remaniement. Nous avons vérifié la stratégie MLCA pour trois années de la CPS: 1993, 1995 et 1996, puisque des données de réinterview non rapprochée de la CPS étaient facilement accessibles pour cette période de temps.

Les ménages de la CPS sont interviewés au cours de quatre mois consécutifs, ne font pas partie de l'enquête pendant huit mois, puis sont de nouveau interviewés au cours d'une deuxième série de quatre mois consécutifs. La MLCA exige au moins trois interviews consécutives pour l'identifiabilité des paramètres de modèle. Nous avions un choix d'ensembles de données comprenant toutes les personnes interviewées au cours de trois ou quatre mois consécutifs de la CPS. Puisque le recours à quatre mois de données réduirait la taille de l'échantillon pour l'analyse de moitié, nous avons choisi de faire porter l'analyse sur trois mois consécutifs (janvier, février et mars) pour les trois années de données. Les cas de non-réponse et les cas de changement du ménage tout entier pour un ou plusieurs des trois mois ont été exclus de l'analyse.

Le modèle MLCA le plus simple précise que les probabilités de réponse $\pi_{a|x}$, $\pi_{b|y}$, et $\pi_{c|z}$, ainsi que les probabilités de changement, $\pi_{y|x}$, $\pi_{z|y}$ sont identiques pour toutes les personnes de la population cible (c'est ce qu'on appelle l'homogénéité). Toutefois, notre analyse préliminaire (Biemer, Bushery et Flanagan 1997) a indiqué que les probabilités de réponse et de changement ne sont pas homogènes. Afin de rendre compte de cette hétérogénéité, nous avons examiné un certain nombre de covariables et de variables de stratification à inclure dans les modèles, y compris: le sexe, la scolarité, le mode d'interview, le type de réponse (avec ou sans procuration) et la race. Parmi celles-ci, c'est la variable tirée de l'indicateur de réponse avec ou sans procuration de la CPS qui a le mieux tenu compte de l'hétérogénéité de la population. Cette variable, notée P, se définit comme suit:

$$P = \begin{cases} 1 & \text{si les trois interviews sont toutes menées sans procuration (SELF)} \\ 2 & \text{si deux des trois interviews sont menées sans procuration (MOSTLY SELF)} \\ 3 & \text{si deux des trois interviews sont menées par procuration (MOSTLY PROXY)} \\ 4 & \text{si les trois interviews sont toutes menées par procuration (PROXY)} \end{cases}$$

À noter que nous utilisons maintenant P pour représenter la variable de groupement, au lieu de G, que nous avons utilisé à la section 2. Compte tenu de recherches antérieures (par exemple O'Muircheartaigh 1991), nous nous attendons à ce que le groupe Self (P=1) comporte moins d'erreurs de classification que le groupe Proxy (P=4). Nous vérifions cette hypothèse dans le cadre du critère de plausibilité des estimations (critère 4 ci-dessus).

La taille des échantillons pour les trois ensembles de données utilisés dans notre analyse est la suivante:

1993: 45 291 personnes
1995: 49 347 personnes
1996: 41 751 personnes

Pour 1993, un tiers environ de l'échantillon se trouve dans le groupe Self, un quart environ dans le groupe Proxy, les autres membres de l'échantillon étant répartis à peu près également entre les groupes Mostly Self et Mostly Proxy. Pour 1995 et 1996, le nombre de membres de l'échantillon

dans le groupe Proxy est légèrement plus élevé (un tiers au lieu d'un quart).

3.4 Ajustement des modèles MLCA

Pour l'ajustement d'un modèle MLCA comportant une seule variable de groupement, P, l'ensemble de données d'entrée était un tableau 4×3×3×3 à dénombrement de cellules défini par le tri croisé de P×A×B×C, où A, B et C représentent la classification de la situation vis-à-vis de l'activité pour janvier, février et mars, respectivement.

Le logiciel ℓEM et d'autres progiciels d'ajustement de modèle MLCA supposent un échantillonnage aléatoire simple, de sorte que le plan de sondage complexe de la CPS ne peut pas être modélisé exactement. Il est possible de rendre compte de la structure d'échantillonnage à probabilités inégales de la CPS à l'aide de dénombrements de cellules pondérés et remis à l'échelle plutôt que de totaux de cellules bruts (Clogg et Eliason 1985). Toutefois, le recours à des données non pondérées pour la stratégie MLCA offre deux avantages importants. Premièrement, nous pouvons comparer les estimations MLCA aux estimations tirées des études citées antérieurement de l'erreur de classification de la CPS, toutes fondées sur des données non pondérées. Deuxièmement, les données de réinterview de la CPS servant à évaluer les critères 3 et 4 ne sont pas pondérées et aucun poids n'est disponible. Par conséquent, une partie au moins de l'analyse exige des données non pondérées; l'utilisation de données pondérées pour les autres critères risquerait de produire de fausses incohérences dans les résultats.

Afin d'examiner la validité d'inférences pour la population totale fondées sur une analyse non pondérée, nous avons estimé les erreurs de classification pour des données à la fois pondérées et non pondérées et nous avons observé que les estimations d'erreur de classification exprimées sous forme de proportions étaient pratiquement identiques, la différence n'étant manifeste qu'à la troisième décimale. Les résultats présentés ci-dessous pour des dénombrements de cellules non pondérés sont donc appropriés pour une inférence au-delà de l'échantillon de la CPS à la population totale.

Un autre aspect du recours à une analyse non pondérée est l'estimation des erreurs-types. Puisqu'elles sont calculées en fonction d'hypothèses d'échantillonnage aléatoire simple, les estimations ℓEM de l'erreur-type risquent d'être sous-estimées par manque de connaissance de l'effet de groupement dans l'échantillon de la CPS. Pour tenir compte approximativement de ce phénomène, nous pouvons multiplier les variances ℓEM par un effet de plan de sondage calculé à partir des estimations de la population active de la CPS. Le Bureau of the Census des États-Unis (2000, 14-9) a indiqué que les effets de plan de sondage pour les estimations de la population active de la CPS ne dépassent pas 1,3; le fait de multiplier les erreurs-types ℓEM par (1,3)½ devrait entraîner une inflation suffisante des erreurs -types pour tenir compte du groupement. Une stratégie équivalente consiste à utiliser un niveau de signification de 3% plutôt que de 5% au moment de déclarer la différence entre deux estimations. Cette dernière stratégie sera considérée comme appropriée dans l'analyse à venir. Nous croyons que cela représente un test prudent puisque l'effet du plan de sondage de la CPS reflète l'augmentation de la variance attribuable tant au groupement de l'échantillon qu'à la pondération inégale, tandis que seuls des effets de groupement figurent dans nos estimations non pondérées.

Le tableau 1 indique les résultats de l'ajustement d'une série de modèles MLCA de plus en plus complexes pour chacun des trois ensembles de données. Le modèle de base est le modèle MLCA le plus simple; il précise que les probabilités de changement et les probabilités de réponse sont homogènes (c'est-à-dire qu'elles ne diffèrent pas selon le groupe P) et stationnaires (c'est-à-dire qu'elles sont identiques pour les trois mois). Ce modèle peut s'écrire comme suit

$$\pi_{p,a,b,c} = \sum_{x,y,z} \pi_p \pi_{x|g} \pi^3_{a|x} \pi^2_{y|x,} \qquad (11)$$

que l'on obtient de (4) en imposant les contraintes

$$\pi_{z|yp} = \pi_{y|xp} = \pi_{y|x} \qquad (12)$$

et

$$\pi_{a|xp} = \pi_{b|yp} = \pi_{c|zp} = \pi_{a|x} \qquad (13)$$

pour tous les p.

Pour le modèle 1 nous réduisons la contrainte (12) à

$$\pi_{z|yp} = \pi_{y|xp} \text{ pour } p = 1,...,4 \qquad (14)$$

ce qui permet aux changements survenus de janvier à février et de février à mars de varier selon le groupe Self/Proxy, P. Pour le modèle 2, nous réduisons la contrainte (12) davantage à

$$\pi_{y|xp} = \pi_{y|x} \text{ et } \pi_{z|yp} = \pi_{z|y} \qquad (15)$$

pour tous les p. Le modèle 3 réduit les contraintes tant d'homogénéité que de stationnarité pour les probabilités de changement de sorte que $\pi_{y|xp} \neq \pi_{z|yp}$. Ce modèle permet donc aux probabilités de changement de varier selon le groupe et selon le mois. Toutefois, les probabilités de réponse doivent toujours être égales d'un groupe à l'autre et d'un mois à l'autre.

Le modèle 4 est le plus général des modèles identifiables que nous avons considérés. Le modèle 4 permet aux probabilités de changement de janvier à février et de février à mars de varier indépendamment de l'un à l'autre des quatre groupes Proxy/Self. Ce modèle précise également que les probabilités de réponse sont les mêmes pour janvier, février et mars, mais qu'elles peuvent varier de l'un à l'autre des quatre groupes Proxy/Self. Nous avons obtenu ce modèle à partir du modèle 3 en réduisant les contraintes exigeant des

$$\pi_{a|xp} = \pi_{b|yp} = \pi_{c|zp} \qquad (16)$$

Tableau 1
Diagnostics modélisés pour d'autres modèles MLCA selon l'année

Données de 1993	*df*	npar[1]	L^2	valeur *p*	BIC	*d*
Modèle de base: Changements homogènes et stationnaires et probabilités de réponse	90	17	645	0,000	-320	0,048
Modèle 1: Changements non homogènes	84	23	632	0,000	-269	0,047
Modèle 2: Changements non stationnaires	66	41	99	0,006	-609	0,007
Modèle 3: Changements non homogènes et non stationnaires	42	65	64	0,016	-386	0,005
Modèle 4: Changements non homogènes et non stationnaires et probabilités de réponse non homogènes	24	83	23	0,501	-234	0,002
Données de 1995	*df*	npar[1]	L^2	valeur *p*	BIC	*d*
Modèle de base: Changements homogènes et stationnaires et probabilités de réponse	90	17	697	0,000	-275	0,044
Modèle 1: Changements non homogènes	84	23	668	0,000	-240	0,043
Modèle 2: Changements non stationnaires	66	41	146	0,000	-567	0,008
Modèle 3: Changements non homogènes et non stationnaires	42	65	82	0,000	-372	0,005
Modèle 4: Changements non homogènes et non stationnaires et probabilités de réponse non homogènes	24	83	25	0,410	-234	0,002
Données de 1996	*df*	npar[1]	L^2	valeur *p*	BIC	*d*
Modèle de base: Changements homogènes et stationnaires et probabilités de réponse	90	17	632	0,000	-325	0,045
Modèle 1: Changements non homogènes	84	23	585	0,000	-308	0,044
Modèle 2: Changements non stationnaires	66	41	159	0,000	-543	0,010
Modèle 3: Changements non homogènes et non stationnaires	42	65	82,6	0,000	-364	0,005
Modèle 4: Changements non homogènes et non stationnaires et probabilités de réponse non homogènes	24	83	39,3	0,026	-216	0,003

[1] À noter que «npar» représente le nombre de paramètres du modèle.

pour tous les *p*. En vertu de ces contraintes, (4) peut s'écrire comme suit

$$\pi_{p,a,b,c} = \sum_{x,y,z} \pi_p \pi_{x|p} \pi_{y|xp} \pi_{z|py} (\pi_{a|p,x})^3.$$

Dans le tableau 1, nous indiquons la statistique d'ajustement de base pour l'ensemble des cinq modèles et des trois années. La colonne 4 du tableau indique la statistique chi-carré habituelle du rapport des vraisemblances L^2 (voir Agresti 1990, 48), tandis que la colonne 5 indique la valeur *p* correspondante. Une valeur *p* égale ou supérieure à 0,05 est le critère habituel d'un ajustement de modèle adéquat. Toutefois, à cause de la taille élevée des échantillons de notre analyse, le fait d'exiger une valeur *p* aussi grande pourrait entraîner un surajustement du modèle. Nous considérons une valeur *p* aussi petite que 0,01 comme étant acceptable. La mesure BIC du tableau se définit comme suit:

$$\text{BIC} = L^2 - (\log N)df$$

où *N* est la taille de l'échantillon total et *df* représente les degrés de liberté du modèle. Le BIC résume essentiellement le compromis entre l'ajustement du modèle (L2) et la parcimonie du modèle (*df*). Puisque de petites valeurs du BIC sont favorables, nous considérons le modèle comportant le plus faible BIC comme le meilleur relativement à la validité de l'ajustement et à la parcimonie. Liu et Dayton (1997) ont discuté de cette stratégie pour des modèles de structure latente.

Enfin, l'indice de dissimilitude (*d*) est la proportion des observations qui auraient à changer de cellule pour que le modèle soit parfaitement ajusté. En général, on considère des modèles comportant $d \leq 0{,}05$ (erreur de modèle de 5%) comme étant bien ajustés aux données (Vermunt 1997).

Pour chaque année de données, le modèle 4 est le seul à fournir un ajustement acceptable lorsque le critère des valeurs *p* est considéré. Le modèle 4 est également plausible du point de vue de la théorie des réponses puisqu'il postule que l'erreur de classification varie selon le groupe Self/Proxy. Cela, comme nous l'avons dit, correspond à la documentation publiée sur les méthodes d'enquête (voir par exemple O'Muircheartaigh 1991 et Moore 1988). L'indice de dissimilitude *d*, pour le modèle est de 0,3%, ce qui indique un ajustement de modèle très valable. Nous utilisons donc le modèle 4 pour produire les estimations de l'erreur de classification de la population active.

3.5 Estimation de l'erreur de classification

Le tableau 2 montre les estimations des probabilités de réponse tirées du modèle 4 pour les catégories employé, sans emploi et inactif. Pour les personnes réellement employées et celles réellement inactives, la probabilité d'une réponse correcte est assez élevée: au moins 98% pour les employés et 97% pour les inactifs. Toutefois, pour les personnes réellement sans emploi, la probabilité d'une réponse correcte varie d'une année à l'autre et d'un groupe à l'autre entre 68% environ et 86% environ. Comme prévu,

le groupe Self comporte la plus forte probabilité d'une réponse correcte (statistiquement significative pour 1993 et 1996). Le groupe Mostly Self manifeste également une tendance à une plus forte probabilité de réponse correcte que le groupe Mostly Proxy, mais la différence n'est pas statistiquement significative.

Un résultat surprenant du tableau 2 est la direction que prend la différence d'exactitude de déclaration pour les personnes réellement sans emploi d'une des trois années à l'autre. On se souviendra que le questionnaire d'interview de la CPS a été remanié en 1994 afin d'améliorer l'exactitude de déclaration. Toutefois, ces résultats indiquent que l'exactitude de déclaration est plus élevée pour 1993 (année antérieure au remaniement majeur) que pour 1995 et 1996 (les années suivant le remaniement). En 1993, la probabilité d'une réponse correcte était de 81,8%, comparativement à 76,1% et à 74,4% pour 1995 et 1996, respectivement (valeur $p < 0{,}001$). Cet effet est peut-être une conséquence du remaniement, ou reflète peut-être des changements réels de la population, ou les deux à la fois. Des recherches se poursuivent actuellement au Bureau of the Census pour comprendre les causes qui sous-tendent ces résultats.

Le tableau indique que la classification erronée des personnes sans emploi comme des personnes inactives est un problème plus grave que la classification erronée de ces personnes comme des personnes employées. En moyenne, pour les trois années, deux tiers environ de l'erreur de classification des sans-emploi est une classification erronée de ces personnes comme inactives, mais le taux est élevé pour les deux types d'erreur.

Nous comparons ensuite nos estimations des probabilités de classification de la CPS à des estimations semblables tirées de la documentation. Au tableau 3, les estimations MLCA pour chacune des trois années sont comparées aux estimations de Chua et Fuller (1987), de Poterba et Summers (1995) et du programme de réinterviews non rapprochées de la CPS. Encore une fois, les trois derniers ensembles d'estimations se fondent sur des données de réinterview, tandis que les estimations MLCA sont préparées directement à partir des données d'interview de la CPS. En général, la valeur relative des estimations MLCA d'une catégorie d'activité à l'autre concorde avec les estimations antérieures. La plus grande différence est observée pour la population des personnes réellement sans emploi. Pour ce groupe, plus élevées que les estimations MLCA correspondantes pour 1993, la période de temps qui correspond le plus étroitement aux estimations comparatives.

Tableau 2
Probabilités de classification estimatives de la situation vis-à-vis de l'activité selon le groupe et l'année
(Les erreurs-types sont indiquées entre parenthèses)

Observé		Réellement employé			Réellement sans emploi			Réellement inactif		
	Groupe	1993	1995	1996	1993	1995	1996	1993	1995	1996
Employé	Sans procuration	98,90	98,98	99.03	4,24	5,66	5,78	1,05	0,67	0,92
		(0,12)	(0,12)	(0,13)	(0,96)	(1,28)	(1,52)	(0,13)	(0,10)	(0,13)
	Sans procuration surtout	98,91	98,61	98,86	8,11	7,66	12,20	1,73	1,59	0,93
		(0,18)	(0,21)	(0,22)	(1,89)	(2,50)	(2,96)	(0,27)	(0,30)	(0,24)
	Par procuration surtout	98,62	98,56	98,76	11,19	12,86	10,18	1,68	0,95	1,50
		(0,18)	(0,22)	(0,20)	(1,96)	(2,97)	(2,73)	(0,27)	(0,25)	(0,30)
	Par procuration	98,66	98,63	98,66	6,60	7,77	8,30	1,56	1,50	1,30
		(0,16)	(0,13)	(0,15)	(1,43)	(1,44)	(1,62)	(0,20)	(0,17)	(0,18)
	Total	98,77	98,73	98,83	7,06	7,86	8,57	1,41	1,11	1,13
		(0,11)	(0,11)	(0,11)	(0,70)	(0,90)	(1,00)	(0,11)	(0,11)	(0,11)
Sans emploi	Sans procuration	0,33	0,39	0,41	85,92	80,52	79,66	0,48	0,35	0,40
		(0,08)	(0,08)	(0,09)	(1,62)	(2,04)	(2,44)	(0,11)	(0,10)	(0,12)
	Sans procuration surtout	0,28	0,50	0,28	82,68	71,71	73,60	1,03	0,91	0,79
		(0,12)	(0,13)	(0,15)	(2,48)	(3,40)	(3,71)	(0,23)	(0,21)	(0,28)
	Par procuration surtout	0,37	0,54	0,37	76,54	68,09	72,74	0,81	0,90	1,53
		(0,11)	(0,15)	(0,13)	(2,55)	(3,37)	(3,70)	(0,22)	(0,28)	(0,26)
	Par procuration	0,34	0,55	0,39	80,09	77,12	71,63	0,95	0,94	1,18
		(0,10)	(0,09)	(0,09)	(2,19)	(2,18)	(2,49)	(0,19)	(0,17)	(0,20)
	Total	0,34	0,49	0,37	81,81	76,09	74,42	0,75	0,69	0,87
		(0,11)	(0,11)	(0,11)	(0,90)	(1,21)	(1,21)	(0,11)	(0,11)	(0,11)
Inactif	Sans procuration	0,77	0,63	0,55	9,84	13,82	14,56	98,47	98,98	98,68
		(0,10)	(0,08)	(0,09)	(1,39)	(1,76)	(2,11)	(0,17)	(0,14)	(0,18)
	Sans procuration surtout	0,81	0,89	0,86	9,21	20,63	14,20	97,24	97,50	98,28
		(0,13)	(0,15)	(0,16)	(1,83)	(2,94)	(2,94)	(0,35)	(0,36)	(0,37)
	Par procuration surtout	1,01	0,90	0,87	12,27	19,05	17,08	97,52	98,15	96,96
		(0,14)	(0,15)	(0,15)	(1,97)	(2,58)	(3,16)	(0,35)	(0,38)	(0,40)
	Par procuration	0,10	0,82	0,95	13,31	15,11	20,07	97,49	97,56	97,52
		(0,13)	(0,10)	(0,12)	(1,83)	(1,86)	(2,22)	(0,27)	(0,24)	(0,27)
	Total	0,89	0,78	0,79	11,13	16,04	17,00	97,84	98,20	98,00
		(0,11)	(0,11)	(0,11)	(0,90)	(1,21)	(1,21)	(0,11)	(0,11)	(0,11)

Tableau 3
Comparaison des estimations MLCA aux estimations publiées antérieurement

Classification		MLCA		Chua et Fuller (données de 1982)	Poterba et Summers (données de 1981)	CPS Réinterview rapprochée (1977-1982)
Réelle	Observée					
Employé	Employé	98,77	(1993)	98,66 (1 mois)	97,74	98,78
		98,73	(1995)	98,65 (2 mois)		
		98,73	(1996)			
	Sans emploi	0,34	(1993)	0,32 (1 mois)	0,54	0,19
		0,49	(1995)	0,34 (2 mois)		
		0,37	(1996)			
	Inactif	0,89	(1993)	1,02 (1 mois)	1,72	1,03
		0,78	(1995)	1,01 (2 mois)		
		0,79	(1996)			
Sans emploi	Employé	7,06	(1993)	3,52 (1 mois)	3,78	1,91
		7,86	(1995)	3,51 (2 mois)		
		8,57	(1996)			
	Sans emploi	81,81	(1993)	88,27 (1 mois)	84,76	88,57
		76,09	(1995)	88,23 (2 mois)		
		74,42	(1996)			
	Inactif	11,13	(1993)	8,21 (1 mois)	11,46	9,53
		16,04	(1995)	8,16 (2 mois)		
		17,00	(1996)			
Inactif	Employé	1,41	(1993)	1,60 (1 mois)	1,16	0,5
		1,11	(1995)	1,61 (2 mois)		
		1,13	(1996)			
	Sans emploi	0,75	(1993)	1,19 (1 mois)	0,64	0,29
		0,69	(1995)	1,24 (2 mois)		
		0,87	(1996)			
	Inactif	97,84	(1993)	97,21 (1 mois)	98,2	99,21
		98,20	(1995)	97,15 (2 mois)		
		98,00	(1996)			

Une explication de cette différence est que les estimations comparatives manifestent un biais vers le haut à cause de corrélations entre les erreurs d'interview et de réinterview. Une autre explication est que les estimations MLCA comportent un biais vers le bas à cause du manque de validité de l'hypothèse markovienne. Nous estimons que les deux explications sont valables dans une certaine mesure. Toutefois, nous indiquons à la section suivante que l'échec de l'hypothèse markovienne exerce probablement une faible influence sur les estimations de l'erreur de classification.

3.6 Robustesse de la stratégie MLCA à l'égard des changements non markoviens de la situation vis-à-vis de l'activité

Plusieurs auteurs ont examiné l'effet de la situation actuelle et antérieure vis-à-vis de l'emploi sur la situation future vis-à-vis de l'emploi (voir par exemple Akerlof et Main 1980; Heckman et Borjas 1980; Lynch 1989, Corak 1993). Heckman et Borjas ont indiqué que l'étude de cette question est assez difficile à cause des biais de sélection, des erreurs de réponse et de l'hétérogénéité non observée. Ces sources de confusion rendent peut-être compte des résultats peu cohérents de la documentation. Ainsi, à l'aide de données de la CPS, Akerlof et Main (1980) ont pu indiquer que la probabilité d'un futur chômage dépend du nombre de situations antérieures de chômage de même que de la durée de ces situations. Toutefois, dans une étude d'hommes ayant terminé leurs études secondaires, Heckman et Borjas (1980) n'ont trouvé aucune indication que les situations antérieures de chômage ou leur durée ont un effet sur le futur comportement vis-à-vis de l'activité lorsque l'on tient compte du biais de tirage de l'échantillon et du biais dû à l'hétérogénéité. Les résultats tirés de la documentation sont également incohérents et ambigus relativement à la mesure dans laquelle l'hypothèse markovienne exprimée en (2) risque d'être enfreinte pour la CPS et d'autres enquêtes sur la population active. Néanmoins, dans la présente section, nous cherchons à fournir une réponse au moins partielle à la question de savoir comment les changements non markoviens de la situation vis-à-vis de l'activité influencent les estimations MLCA de l'erreur de classification.

Afin de pouvoir étudier l'effet des infractions à l'hypothèse markovienne en (2) pour la présente application, nous avons mené une étude de simulation limitée. Afin d'orienter la recherche tout en simplifiant le cadre de la simulation,

nous avons considéré des structures latentes comportant uniquement deux classes ou situations à un moment donné: chômage, noté X, Y ou $Z = 1$, et autre (employé ou inactif), noté X, Y ou $Z = 2$ avec des définitions semblables pour les situations observées A, B et C. Afin d'établir une population en vue de la simulation, les probabilités latentes π_x, $\pi_{y|x}$, et $\pi_{z|xy}$ et les probabilités de réponse $\pi_{a|x} = \pi_{b|y} = \pi_{c|z}$ ont été précisées comme concordant avec les ensembles combinés de données de 1993, 1995 et 1996.

Nous avons alors défini deux paramètres, λ_1 et λ_2, variant dans la simulation, où

$$\lambda_1 = \frac{\pi_{z=1|x=2,y=1}}{\pi_{z=1|x=1,y=1}} \tag{17}$$

et

$$\lambda_2 = \frac{\pi_{z=1|x=2,y=2}}{\pi_{z=1|x=1,y=2}}. \tag{18}$$

Ainsi, λ_1 est la probabilité qu'une personne soit «sans emploi» en mars, après avoir été «sans emploi» en février et «autre» en janvier relativement à la probabilité d'être «sans emploi» en mars après avoir été «sans emploi» au cours des deux mois précédents. En conformité avec les résultats obtenus par Akerlof et Main (1980), qui ont montré que la probabilité qu'une personne demeure sans emploi augmente à mesure que le nombre de situations sans emploi augmente, nous supposons que $0 \leq \lambda_1 \leq 1$. De même, λ_2 est la probabilité qu'une personne soit «sans emploi» en mars, après avoir été «autre» au cours des deux mois précédents, relativement à la probabilité d'être «sans emploi», après avoir été «autre» en février et «sans emploi» en janvier. Encore une fois, suivant Akerlof et Main, nous supposons que $0 \leq \lambda_2 \leq 1$. À noter que, lorsque $\lambda_1 = \lambda_2 = 1$, les changements de chômage entre février et mars sont markoviens.

Les données simulées ont été produites de façon à correspondre complètement à un modèle MLCA à probabilités de changement non stationnaires lorsque $\lambda_1 = \lambda_2 = 1$. Nous avons simulé l'échec de l'hypothèse markovienne en variant λ_1 et λ_2 entre 0 et 1. Par souci d'uniformité avec les données 1993-1996, nous avons fixé la probabilité d'une réponse «sans emploi» correcte, $\pi_{a=1|x=1}$, à 0,80 et la probabilité d'une réponse «autre» correcte, $\pi_{a=2|x=2}$, à 0,99 dans toutes les simulations. De plus, les dénominateurs de λ_1 et de λ_2 ont été fixés à leurs valeurs déterminées à partir des données combinées 1993-1996, tandis que les numérateurs ont été calculés à partir de (17) et de (18) à l'aide de valeurs de λ_1 et de λ_2 précisées pour chaque exécution d'une simulation.

Le tableau 4 résume les résultats de la simulation pour $\lambda_1 = \lambda_2 = \lambda$ où l'on fait varier λ entre 0,2 et 1,0 par paliers de 0,2. À noter que pour $\lambda_1 = \lambda_2 = 1{,}0$, qui correspond à un modèle markovien, les probabilités estimatives de réponse correcte sont exactement comme prévu. Pour des valeurs plus petites de λ_1 et de λ_2, les estimations manifestent un biais négatif, le biais maximal correspondant à la valeur la plus faible envisagée, 0,2. Néanmoins, le biais absolu attribuable à des probabilités de changement non markovien ne dépasse jamais 3 points de pourcentage. Les résultats du tableau 4 correspondent à ceux de Bushery et Kindelberger (1999), qui ont utilisé une stratégie un peu différente afin d'illustrer le même caractère de robustesse des modèles MLCA pour des données de la CPS. Les deux études indiquent que le manque de validité de l'hypothèse markovienne ne semble pas être une source importante de biais dans l'estimation des probabilités d'erreur de classification de la CPS.

Tableau 4
Estimations de classification correcte pour des changements non markoviens
(les entrées de cellules sont des pourcentages)

Pr (correct)	$\lambda_1 = \lambda_2 = \lambda$				
	$\lambda = 0.2$	$\lambda = 0.4$	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 1.0$ (Markov)
Pr («chôm.»| «chôm.» réel) $= \pi_{a=1\|x=1}$	776	781	787	793	800
Pr («autre»| «autre» réel) $= \pi_{a=2\|x=2}$	986	987	988	989	990

4. COMPARAISON DES ESTIMATIONS DE TYPE MLCA ET DE RÉINTERVIEW NON RAPPROCHÉE

4.1 Estimation de Hui-Walter

On peut obtenir un autre ensemble d'estimations des probabilités de réponse des données de réinterview de la CPS à l'aide d'un type de modèle de structure latente d'abord proposé par Hui et Walter (1980). Nous utilisons la notation mentionnée ci-dessus; soit X, la classification réelle de la situation vis-à-vis de l'activité à un moment donné; soit A et A', les classifications d'interview et de réinterview, respectivement. Soit G, une variable de groupement définie comme en (4). Considérons la vraisemblance du tableau groupe× interview× réinterview noté GAA'. Soit $\pi_{gaa'}$ la probabilité qu'une personne relevant du groupe g soit classée dans la cellule (a, a') du tableau. Le modèle pour $\pi_{gaa'}$ proposé par Hui et Walter est le suivant:

$$\pi_{gaa'} = \sum_x \pi_g \pi_{x|g} \pi_{a|x} \pi_{a'|x}. \tag{19}$$

Dans ce modèle, l'hypothèse des mesures parallèles pour les réponses d'interview et de réinterview est réduite et les probabilités de réponse pour les deux mesures, c'est-à-dire $\pi_{a|x}$ et $\pi_{a'|x}$, sont estimées séparément. L'hypothèse ICE est adoptée comme condition de l'identifiabilité. Il est également supposé que $\pi_{a|x}$ et $\pi_{a'|x}$, ne dépendent pas de la variable de groupe, G, tandis que la fréquence des trois catégories employé, sans emploi et inactif, c'est-à-dire $\pi_{x|g}$, dépend toujours de G.

Sinclair et Gastwirth (1996), dans leur analyse de l'erreur de classification de la population active de la CPS, ont utilisé le sexe comme variable de groupement, et nous utilisons également cette variable de groupement dans notre analyse. Sinclair et Gastwirth ont limité leur analyse à des

hommes blancs et des femmes blanches et à deux catégories d'activité: inactif et actif. Cette dernière catégorie est la somme de nos catégories employé et sans emploi. Dans notre analyse, nous considérons des membres d'échantillon de toutes les races et nous analysons la classification à trois catégories de la situation vis-à-vis de l'activité utilisée dans la stratégie MLCA. Ainsi, l'analyse H-W estime 16 paramètres pour chaque année, ce qui correspond au nombre de degrés de liberté disponibles du tableau $G \times A \times A'$, ne laissant aucun degré de liberté pour vérifier l'ajustement du modèle.

Le logiciel ℓEM a été utilisé pour l'ajustement du modèle H-W aux données d'interview et de réinterview non rapprochée pour trois périodes de temps coïncidant avec les trois périodes de notre stratégie MLCA: avant 1994, 1995, 1996. Nous avons tenté de limiter l'analyse au premier trimestre seulement de ces périodes de temps. Malheureusement, à cause de la faible taille des échantillons, les estimations ont été plutôt instables. Il a donc fallu utiliser les données de réinterview de l'ensemble des quatre trimestres de ces périodes de temps. Les données antérieures à 1994 ont été recueillies de 1985 à 1988 grâce à un échantillon de réinterviews non rapprochées.

Les résultats de cette comparaison des estimations MLCA et H-W sont résumés dans le tableau 5. Les estimations MLCA sont les mêmes que celles des lignes du tableau 2 étiquetées «Total». Les estimations H-W sont les probabilités de classification associées à l'interview originale, c'est-à-dire la mesure A en (19). Le tableau montre la comparaison pour les trois années. Puisque le taux d'erreur le plus élevé de la stratégie MLCA est survenu pour la catégorie sans emploi, celle-ci revêt un intérêt particulier dans la comparaison MLCA/H-W.

Tableau 5
Comparaison d'estimations modélisées MLCA et H-W des probabilités de réponse de la CPS selon l'année (écarts-types entre parenthèses)

Classification		1993		1995		1996	
Réelle	Observée	H-W	MLCA	H-W	MLCA	H-W	MLCA
Employé	Employé	99,3 (0,3)	98,8 (0,1)	99,5 (0,7)	98,7 (0,1)	99,6 (0,1)	98,8 (0,1)
	Sans emploi	0,0 (0,0)	0,3 (0,1)	0,0 (n/a)	0,5 (0,1)	0,4 (0,1)	0,4 (0,1)
	Inactif	0,7 (0,3)	0,9 (0,1)	0,5 (0,7)	0,8 (0,1)	0,0 (n/a)	0,8 (0,1)
Sans emploi	Employé	11,1 (1,0)	7,1 (0,7)	11,5 (2,3)	7,9 (0,9)	4,6 (15,2)	8,6 (1,0)
	Sans emploi	74,3 (2,7)	81,8 (1,1)	67,9 (6,1)	76,1 (1,3)	67,6 (11,1)	74,4 (1,4)
	Inactif	14,7 (2,9)	11,1 (0,9)	20,6 (6,5)	16,0 (1,2)	27,9 (5,3)	17,0 (1,2)
Inactif	Employé	2,0 (0,5)	1,4 (0,1)	2,5 (1,5)	1,1 (0,1)	2,6 (1,5)	1,1 (0,1)
	Sans emploi	1,2 (0,3)	0,8 (0,1)	0,5 (0,6)	0,7 (0,1)	0,0 (n/a)	0,9 (0,1)
	Inactif	96,8 (0,6)	97,8 (0,1)	97,0 (1,6)	98,2 (0,1)	97,4 (1,1)	98,0 (0,1)

En général, les deux ensembles d'estimations sont en assez bon accord. Les années 1995 et 1996 ne manifestent aucune différence statistiquement significative (au niveau de 5%) entre les estimations MLCA et H-W pour ce qui est de la population des sans-emploi. Les estimations antérieures à 1994 affichent des différences appréciables; toutefois, celles-ci s'expliquent peut-être par le fait que les données de réinterview antérieures à 1994 ont été recueillies de 1985 à 1988, plutôt qu'en 1993. Ces différences sont étudiées plus en profondeur dans la section qui suit.

4.2 Comparaison des indices d'incohérence

Comme il a été décrit à la section 3.1, nous estimons l'indice d'incohérence pour chaque période de temps à l'aide des estimations MLCA modélisées des probabilités de réponse. Essentiellement, nous estimons le tableau de tri croisé prévu des interviews-réinterviews à partir des estimations MLCA des probabilités de réponse, et nous appliquons ensuite la formule pour l'indice à ce tableau comme si le tableau était observé. Un deuxième tableau de la classification prévue des interviews-réinterviews peut être estimé à l'aide des estimations H-W des probabilités de réponse. Nous comparons ensuite ces deux ensembles d'estimations à l'estimation de l'indice obtenue directement des données de réinterview de la CPS à l'aide de méthodes traditionnelles (U.S. Bureau of the Census 1985). L'accord entre les trois estimations confirme la validité des trois méthodes.

Le tableau 6 montre les trois méthodes d'estimation de l'indice d'incohérence pour les trois périodes de temps. Comme auparavant, la catégorie sans emploi revêt un intérêt particulier à cause de son taux d'erreur élevé. Puisque les erreurs-types ne sont pas disponibles pour les estimations MLCA ou H-W de l'indice, il n'est pas possible de vérifier l'hypothèse formellement. Toutefois, il existe des erreurs-types pour les estimations traditionnelles que l'on peut utiliser comme approximation des erreurs-types pour les estimations H-W.

Dans l'ensemble, le profil général des estimations MLCA aussi bien que l'ordre de grandeur des estimations MLCA sont en assez bon accord généralement avec les estimations H-W et traditionnelles pour les trois années. Toutefois, pour ce qui est de la catégorie inactif en 1995 et en 1996, les estimations traditionnelles de I sont un peu plus grandes que l'une ou l'autre des estimations modélisées de structure latente. Une analyse plus poussée indique que cette différence est attribuable à un biais de la stratégie d'estimation traditionnelle résultant de l'invalidité de l'hypothèse des mesures parallèles.

Le Bureau of the Census des États-Unis (1985) a montré que si les processus d'interview et de réinterview comportent des fiabilités différentes, l'estimation traditionnelle de l'indice est biaisée. Si, par exemple, la fiabilité des données de réinterview est inférieure à celle des données d'interview, l'estimateur traditionnel de la fiabilité test-retest

sous-estime la fiabilité réelle des données de la CPS; l'indice d'incohérence de la CPS est donc trop élevé.

L'interview et la réinterview de la CPS comporteront des fiabilités différentes si la répartition des erreurs n'est pas égale pour les deux interviews. Il est possible de le vérifier en comparant l'ajustement d'un modèle de type H-W en présence et en l'absence de la restriction $\pi_{a|x} = \pi_{a'|x}$. L'hypothèse d'une fiabilité égale est rejetée si la différence entre les chi-carrés du rapport des vraisemblances pour les deux modèles est supérieure à un chi-carré ayant 6 degrés de liberté. Ce test a été rejeté pour 1995 et 1996 au niveau de signification de 10%. La différence entre les estimations de la catégorie inactif pour 1995 et 1996 semble donc attribuable en partie au biais des estimations traditionnelles de *I*.

Tableau 6
Comparaison des estimations MLCA, H-W et traditionnelles de l'indice d'incohérence selon l'année et la classification de la population active

Méthode d'estimation	Classification de la population active			Indice global
	Employé	Sans emploi	Inactif	
1993				
Estimation traditionnelle	8,16 (0,24)	33,49 (1,16)	9,96 (0,27)	11,05 (0,26)
H-W	7.37	34.93	10.07	10.78
MLCA	6.35	28.04	7.63	8.73
1995				
Estimation traditionnelle	6,69 (0,44)	36,28 (2,85)	10,80 (0,56)	10,42 (0,53)
H-W	6.82	37	8.98	9.7
MLCA	6.06	36.19	7.2	8.72
1996				
Estimation traditionnelle	5,93 (0,39)	35,97 (2,68)	11,95 (0,56)	10,61 (0,51)
H-W	5.67	39.46	7.55	8.56
MLCA	5.99	37.39	7.76	9.06

À noter également que les indices H-W et MLCA concordent assez bien pour 1995 et 1996, même s'ils diffèrent quelque peu en 1993. Toutefois, comme il a été noté à la discussion du tableau 5, les comparaisons entre les estimations MLCA et H-W pour cette année-là sont obscurcies par les différentes périodes de temps utilisées pour la construction de l'ensemble des données d'interview-réinterview antérieur à 1994. Cela pourrait expliquer au moins partiellement l'écart entre les estimations pour cette année-là.

5. SOMMAIRE ET CONCLUSIONS

La présente recherche a surtout tenté d'examiner la validité des estimations MLCA de l'erreur de classification de la population active de la CPS, et de déterminer l'efficacité de la stratégie MLCA comme substitut des méthodes traditionnelles d'évaluation de la qualité des données de la CPS. Nous avons analysé des données d'interview de la CPS pour le premier trimestre des trois années 1993, 1995 et 1996, et nous avons mené une analyse supplémentaire des données de réinterviews non rapprochées de la CPS pour les mêmes périodes de temps à peu près. Les données de réinterview ont représenté une autre stratégie d'estimation de l'erreur de classification de la CPS; comparées aux estimations MLCA, elles ont aidé à examiner la validité de la stratégie MLCA.

Nous avons abordé cinq dimensions de la validité de la stratégie MLCA comme suit:

1. **Diagnostics modélisés.** Nous avons examiné tout un choix de modèles MLCA comportant des variables de groupement définies selon l'âge, la race, le sexe, la scolarité, le mode d'interview et le type de réponse (avec ou sans procuration). Le modèle manifestant la plus grande parcimonie et le meilleur ajustement pour les trois années comporte une variable de groupement définie par le type de réponse (avec ou sans procuration) selon quatre catégories: trois cycles tous menés sans procuration, deux cycles seulement menés sans procuration, deux cycles seulement menés par procuration et trois cycles tous menés par procuration. Pour cette catégorie de modèles, le meilleur a été le modèle 4 (voir le tableau 1), qui précisait des probabilités de changement non homogènes et non stationnaires et des probabilités de réponse non homogènes. Ce modèle a fourni un ajustement adéquat aux données pour les trois années.
2. **Validité de l'ajustement du modèle d'une année à l'autre de la CPS.** Un autre indicateur de la validité du modèle est son ajustement d'un échantillon indépendant à l'autre pour la même population. À supposer que la dynamique de la population active et que la structure des probabilités de réponse pour la CPS soient stables pendant quatre ans, le même modèle général devrait manifester un ajustement adéquat pour les trois années. Le modèle 4 affiche une validité de l'ajustement sur plusieurs années (voir le tableau 1). De plus, d'autres variables de groupement ont été vérifiées dans l'étude, mais le modèle comportant la variable avec ou sans procuration a été le meilleur pour les trois années.
3. **Accord entre les estimations modélisées et de type test-retest des probabilités de réponse.** À l'aide des données d'interview-réinterview non rapprochée tirées de la CPS pour trois périodes (avant 1994, 1995, 1996), nous avons appliqué la méthode H-W à l'estimation des probabilités de réponse et nous avons comparé celles-ci aux estimations MLCA. Un bon accord a été observé pour 1995 et 1996, deux années pour lesquelles les périodes de temps pour les données de réinterview et les données de la CPS manifestaient une étroite concordance (voir le tableau 5). Pour

1993, nous avons observé des différences faibles mais significatives entre les estimations MLCA et les estimations H-W correspondantes. Ces différences pourraient s'expliquer par des différences de période de temps, puisque les données de réinterview ont précédé de quelques années les données d'interview de la CPS.

4. **Accord entre les estimations modélisées et de type test-retest du manque de cohérence.** Nous avons comparé des estimations MLCA modélisées de l'indice d'incohérence aux estimations directes correspondantes tirées du programme de réinterview de la CPS. Les deux ensembles d'estimations concordent assez bien pour les trois années, exception faite pour la catégorie inactif (voir le tableau 6). Pour 1995 et 1996, les différences s'expliquent en partie par le biais de l'estimateur traditionnel résultant de l'échec de l'hypothèse des mesures parallèles. La méthode H-W, qui n'exige pas l'hypothèse des mesures parallèles, permet de produire des estimations de l'indice qui concordent bien avec les estimations MLCA pour 1995 et 1996. Pour 1993, la différence entre les estimations MLCA et H-W est peut-être attribuable à des différences de période de temps pour les ensembles de données de réinterview et de la CPS.

5. **Plausibilité des profils d'erreur de classification.** Les estimations MLCA des probabilités d'erreur de classification semblent plausibles. Les estimations d'un groupe avec/sans procuration à l'autre concordaient avec les attentes antérieures voulant que des taux d'erreur moins élevés soient observés pour les répondants sans procuration que pour les répondants par procuration. De plus, les taux d'erreur les plus élevés ont été observés pour la population des sans-emploi et l'ordre de grandeur de ces estimations concordait avec celui d'études antérieures (voir par exemple Fuller et Chua 1985, Abowd et Zellner 1985, Poterba et Summers 1986, Sinclair et Gastwirth 1996; voir le tableau 3).

En résumé, nous n'avons trouvé dans ces analyses aucune raison de mettre en doute la validité de la stratégie MLCA. La méthode a donné de bons résultats pour les cinq épreuves de validité. Nous recommandons donc que la méthode MLCA soit considérée comme une autre méthode d'évaluation de l'exactitude des estimations de la population active de la CPS. La forte concordance entre les estimations MLCA et H-W confirme également la validité de la méthode H-W. Nous recommandons que les deux méthodes soient considérées dans de futures études de la qualité des données de la CPS.

Même si la stratégie MLCA a donné de bons résultats dans nos épreuves, nous recommandons que la méthode soit appliquée avec prudence à d'autres situations. Dans notre analyse, les données de réinterview ont permis d'évaluer la validité des estimations MLCA. Toutefois, les données de réinterview ne sont typiquement pas disponibles dans les enquêtes par panel et, par conséquent, les analystes ne pourront peut-être appliquer que les critères (1), (2) et (5) ci-dessus pour vérifier la validité du modèle. L'hypothèse markovienne est la clé de la stratégie MLCA. Il se peut que des données recueillies au moyen d'un panel portent sérieusement atteinte à cette hypothèse. Heureusement, l'échec de l'hypothèse markovienne ne semble pas être un facteur important de la validité des estimations MLCA de l'erreur de classification de la population active de la CPS (voir le tableau 4).

BIBLIOGRAPHIE

ABOWD, J., et ZELLNER, A. (1985). Estimating gross labor-force flows. *Journal of Business and Economic Statistics*, 3, 3, 254-283.

AGRESTI, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.

AKERLOF, G.A., et MAIN, G.M. (1980). Unemployment spells and unemployment experience. *The American Economic Review*, 70, 3, 885-893.

BAILAR, B. A. (1975). The effect of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

BIEMER, P., BUSHERY, J., et FLANAGAN, P. (1997). An Application of Latent Markov Models to the CPS. Internal U.S. Bureau of the Census Technical Report.

BIEMER, P., et FORSMAN, G. (1992). On the quality of reinterview data with applications to the Current Population Survey. *Journal of the American Statistical Association*, 87, 420, 915-923.

BOHRNSTEDT, G.W. (1983). Measurement. *Handbook of Survey Research*, (P.H. Rossi, R.A. Wright, et A.B. Anderson, Éds.). New York: Academic Press.

BUSHERY, J., et KINDELBERGER, K. (1999). Simulation Examples for MLC Analysis. Internal U.S. Bureau of the Census Memorandum, Washington, DC, 70-122.

CHUA, T.C., et FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 397, 46-51.

CLOGG, C., et ELIASON, S. (1985). Some common problems in log-linear analysis. *Sociological Methods and Research*, 16, 8-14.

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 210, 37-46.

CORAK, M. (1993). Is unemployment insurance addictive? Evidence from the benefit durations of repeat users. *Industrial and Labor Relations Review*, 47, 1, 62-72.

FORSMAN, G., et SCHREINER, I. (1991). The design and analysis of reinterview: an overview. *Measurement Errors in Surveys*, (P.P. Biemer, et al., Éds.). New York: John Wiley & Sons. 279-302.

FULLER, W., et CHUA, T.C. (1985). Gross change estimation in the presence of response error. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*. Washington, D.C., U.S. Bureau of the Census and U.S. Bureau of Labor Statistics, 65-77.

HECKMAN, J.J., et BORJAS, G.J. (1980). Does unemployment cause future unemployment? Definitions, questions, and answers from a continuous time model of heterogeneous and state dependence. *Economica*, 47, 247-283.

HESS, J., SINGER, E., et BUSHERY, J. (2000). Predicting test-retest reliability from behavior coding. *International Journal of Public Opinion Research*, II, 4, 346-360.

HUI, S.L., et WALTER, S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, 36, 167-171.

KLEINBAUM, D.G., KUPPER, L.L., et MULLER, K.E. (1988). *Applied Regression Analysis and Other Multivariate Methods*, Boston: PWS-KENT Publishing Co.

LIU, T.H., et DAYTON, C.M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics*, 22, 249 - 264.

LYNCH, L.M. (1989). The youth labor market in the eighties: determinants of re-employment probabilities for young men and women. *The Review of Economics and Statistics*, 37-45.

MEYERS, B. D. (1988). Classification-error models and labor-market dynamics. *Journal of Business and Economic Statistics*, 6, 3, 385-390.

MOORE, J.C. (1988). Self/proxy response status and survey response quality. *Journal of Official Statistics*, 4, 2, 155-122.

O'MUIRCHEARTAIGH, C. (1991). Simple response Variance: Estimation and Determinants. *Measurement Errors in Surveys*, (P. Biemer, *et al.*, Éds.). New York: John Wiley & Sons, 551-574.

POTERBA, J., et SUMMERS, L. (1986). Reporting errors and labor market dynamics. *Econometrics*, 54, 6, 1319-1338.

POTERBA, J., et SUMMERS, L. (1995). Unemployment benefits and labor market transitions: a multinomial logit model with errors in classification. *The Review of Economics and Statistics*, 77, 207-216.

POULSEN, C.S. (1982). Latent Structure Analysis with Choice Modeling Applications. Thèse de doctorat, Wharton School, University of Pennsylvania.

ROTHGEB, J. (1994). Revisions to the CPS Questionnaire: Effects on Data Quality, U.S. Bureau of the Census. CPS Overlap Analysis Team Technical Report 2, April 6.

SCHREINER, I. (1980). Reinterview Results from the CPS Independent Reconciliation Experiment (Second Quarter 1978 through Third Quarter 1979). Internal U.S. Bureau of the Census Report.

SHOCKEY, J. (1988). Adjusting for response error in panel surveys, a latent class approach. *Sociological Methods and Research*, 17, 1, 65-92.

SINCLAIR, M., et GASTWIRTH, J. (1996). On procedures for evaluating the effectiveness of reinterview survey methods: application to labor force data. *Journal of the American Statistical Association*, 91, 961-969.

SINCLAIR, M., et GASTWIRTH, J. (1998). Estimations des erreurs de classification dans l'enquête sur la population active et analyse de leur incidence sur les taux de chômage publiés. *Techniques d'enquête*, 24, 2, 171-183.

SINGH, A.C., et RAO, J.N.K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the canadian labour force survey. *Journal of the American Statistical Association*, 90, 430, 478-488.

U.S. BUREAU OF THE CENSUS (1985). Evaluating Censuses of Population and Housing, STD-ISP-TR-5. Washington, D.C.: U.S. Government Printing Office.

U.S. BUREAU OF THE CENSUS (2000). Current Population Survey: Design and Methodology. U.S. Bureau of the Census Technical Paper 63, Washington, D.C.: Government Printing Office.

VAN DE POL, F. et DE LEEUW, J. (1986). A latent markov model to correct for measurement error. *Sociological Methods and Research*, 15, 1-2, 118-141.

VAN DE POL, F., et LANGEHEINE, R. (1997). Separating change and measurement error in panel surveys with an application to labor market data. *Survey Measurement and Process Quality*, (L. Lyberg, *et al*, Éds.). New York: John Wiley & Sons.

VERMUNT, J. (1997). (EM: A General Program for the Analysis of Categorical Data. Tilburg University.

WIGGINS, L.M. (1973). *Panel Analysis, Latent Probability Models for Attitude and Behavior Processing*, Amsterdam: Elsevier S.P.C.

Techniques d'enquête, décembre 2000
Vol. 26, n° 2, pp. 173-183
Statistique Canada

L'estimation et l'estimation de la variance par répliques des prix de vente médians des maisons vendues

KATHERINE J. THOMPSON et RICHARD S. SIGMAN[1]

RÉSUMÉ

Le Bureau of the Census des États-Unis publie des estimations de la médiane pour plusieurs caractéristiques des maisons neuves, une estimation clé étant le prix de vente des maisons vendues. Ces estimations sont calculées à partir de données acquises par voie d'interview des constructeurs de maisons dans le cadre de l'enquête SOC (Survey of Construction). La SOC est une enquête probabiliste à plusieurs degrés dont le plan d'échantillonnage se prête bien à la méthode MHS (répétition modifiée avec demi-échantillon) d'estimation de la variance. Des documents publiés appuient l'application de la méthode MHS à la médiane d'échantillons répétés en vue de l'estimation de la variance d'échantillonnage d'une médiane. Il existe cependant plusieurs avantages, du point de vue des calculs, à utiliser des données groupées en vue de l'estimation des médianes, une interpolation linéaire étant utilisée dans l'intervalle des données groupées comportant la médiane. À l'aide de données d'enquête et de populations finies simulées, nous avons comparé l'effet de l'absence de groupement (donc la médiane de l'échantillon), le groupement avec intervalles de taille fixe et le groupement avec intervalles dont la taille dépend des données sur les médianes et sur les estimations MHS connexes de la variance. Nous avons examiné les erreurs quadratiques moyennes et les erreurs absolues moyennes des estimations de la médiane de même que le biais relatif et la stabilité des estimations de la variance et la couverture des intervalles de confiance connexes. Nous avons observé que les intervalles dont la taille dépend des données donnent des estimations de la variance comportant le biais le moins élevé, la meilleure stabilité et les meilleurs intervalles de confiance.

MOTS CLÉS: Médiane; répétition modifiée avec demi-échantillon; Survey of Construction.

1. INTRODUCTION

Le Bureau of the Census des États-Unis publie des estimations de la médiane pour plusieurs caractéristiques des maisons neuves, une estimation clé étant le prix de vente des maisons vendues. Ces estimations sont calculées à partir de données acquises par voie d'interview des constructeurs de maisons dans le cadre de l'enquête SOC (Survey of Construction). La SOC est une enquête probabiliste à plusieurs degrés dont le plan d'échantillonnage se prête bien à la méthode MHS (répétition modifiée avec demi-échantillon) pour des raisons qui sont expliquées à la section 3.B. Le Bureau of the Census fera bientôt passer la SOC des systèmes actuels d'estimation et d'estimation de la variance à son système remanié, le Standardized Economic Processing System (StEPS). À ce moment-là, la procédure actuelle d'estimation non répétée de la variance pour la SOC sera remplacée par la méthode MHS d'estimation de la variance avec répétition (Thompson 1998). Puisque la méthodologie d'estimation de la variance pour la SOC évolue, nous avons décidé de réexaminer la méthodologie d'estimation de la médiane pour des données continues. Il s'agissait de trouver une méthode d'estimation de la médiane ayant de bonnes propriétés d'estimation et d'estimation de la variance, compte tenu de la répétition de type MHS.

Nous avons considéré deux méthodes d'estimation de la médiane. La première fait appel aux poids d'échantillon pour estimer la médiane à l'aide de fonctions empiriques de distribution cumulative. La deuxième méthode fait appel à une interpolation linéaire de données continues groupées pour obtenir une approximation de la médiane. Cette dernière méthode est mis en œuvre en VPLX (Variances from ComPLeX Survey, Fay 1995), qui est le progiciel d'estimation de la variance avec répétition élaboré au Bureau of the Census.

Le calcul direct des médianes d'échantillon peut exiger un travail de calcul intense parce qu'il exige des tris distincts pour chaque valeur d'une variable de classement donnée. Une autre méthode d'estimation consiste à grouper les données continues en intervalles discrets (appelés casiers) et à utiliser une interpolation linéaire pour l'intervalle comportant la médiane. Pourvu que la distribution des données soit à peu près uniforme pour l'intervalle comportant la médiane, l'interpolation fournit une bonne approximation et exige beaucoup moins de ressources informatiques. Toutefois, la largeur et l'emplacement optimaux des casiers peuvent différer selon le domaine et risquent d'évoluer dans le temps à mesure que la distribution des échantillons change.

Dans le présent exposé, nous comparons six méthodes d'estimation de la médiane, compte tenu de la méthode de répétition MHS: la médiane de l'échantillon et cinq variations axées sur une interpolation linéaire. On trouvera à la section 2 un bref aperçu du plan de sondage de la SOC. La méthodologie générale est expliquée à la section 3. La section 4 décrit les résultats empiriques des quatre mois de données de la SOC qui ont motivé l'étude de simulation

[1] Katherine J. Thompson et Richard S. Sigman, Economic Statistical Methods and Programming Division, U.S. Bureau of the Census, Washington DC, 20233 U.S.A.

présentée à la section 5. Nos conclusions et nos recommandations se trouvent à la section 6.

2. PLAN DE SONDAGE DE LA SOC

L'univers de la SOC contient deux sous-populations: des zones locales qui exigent des permis de construction et des zones locales qui n'en exigent pas. Les unités d'échantillonnage de la SOC tirées de la première sous-population comprennent l'enquête intitulée Survey of the Use of Permits (SUP), et celles tirées de la deuxième sous-population, l'enquête Nonpermit Survey (NP). L'échantillon de la SUP constitue la plus grande partie de l'estimation de la SOC. Les deux sont des échantillons probabilistes à plusieurs degrés stratifiés selon des variables ayant une corrélation prévue élevée avec les statistiques clés de l'enquête: les mises en chantier, les travaux achevés et les ventes.

Le premier degré du tirage des échantillons SUP et NP est un sous-échantillon des unités primaires d'échantillonnage (UPE) du plan de 1980 de la CPS (Current Population Survey), qui sont des parcelles de terrain adjacentes ayant des limites bien définies. Ainsi, les deux enquêtes comportent les mêmes UPE, mais pour le reste les échantillons sont indépendants. Les UPE ont été stratifiées à l'intérieur de la région selon la population pondérée de 1980 des 16 ans et plus, les permis résidentiels pondérés de 1982 et le pourcentage de logements dans des zones sans permis. Dans la mesure du possible, les strates sont constituées d'UPE du même État ayant le même caractère métropolitain. On a tiré une UPE par strate. Des UPE autoreprésentatives (AR) ont été incluses dans l'échantillon avec certitude (la strate est constituée d'une UPE). Des UPE non autoreprésentatives (NAR) ont été tirées moyennant une probabilité proportionnelle à la taille (PPT) à partir de strates comportant plus d'une UPE.

Le deuxième degré du tirage de l'échantillon SUP est un échantillon systématique stratifié d'endroits émettant des permis à l'intérieur des UPE de l'échantillon (tirage une fois par décennie). Ces endroits ont été stratifiés en fonction d'une moyenne pondérée du rapport entre l'émission de permis pour l'année $\underline{i}$ et le total des permis émis aux États-Unis pour l'année $\underline{i}$ ($\underline{i}$ = 78, 81, 82). Dans de nombreux cas, on a tiré une seule unité pour le deuxième degré. Le troisième degré du tirage de l'échantillon SUP est exécuté mensuellement: chaque mois, des représentants locaux (RL) sélectionnent un échantillon systématique de permis de construction provenant des bureaux de permis de chaque endroit échantillonné qui émet des permis. On sélectionne systématiquement des permis de construction comportant d'une à quatre unités de façon à obtenir un échantillon global de type un pour quarante; des permis de construction de cinq unités ou plus sont inclus avec certitude. Les échantillons du troisième degré sont indépendants selon le mois; ceux des premier et deuxième degrés ne le sont pas.

Le deuxième degré du tirage de l'échantillon NP est un échantillon systématique stratifié de petites régions (districts de dénombrement du Recensement de 1980 = DD), stratifié selon la taille de la population du Recensement de 1980. Pour le troisième degré du tirage de l'échantillon NP, des représentants locaux parcourent l'ensemble des routes et des chemins des DD échantillonnés (appelés segments). Afin de réduire le parcours, on a subdivisé certains grands DD et on a choisi un sous-segment, ou on a tiré un sous-échantillon sur deux pour les grands DD. Il existe actuellement un total de soixante-et-onze segments actifs sans permis. Toutes les unités de logement neuves sont incluses dans l'échantillon NP avec certitude.

Les estimations de médiane sont obtenues des échantillons SUP et NP groupés, et elles sont calculées à l'aide d'un poids stratifié a posteriori pour le volet SUP et d'un poids non biaisé pour le volet NP.

3. MÉTHODOLOGIE

A. Procédures d'estimation de la médiane

1. Médiane de l'échantillon

   Une façon d'estimer la médiane d'une population consiste à calculer la médiane de l'échantillon à partir de données non groupées, le poids de l'échantillon permettant de situer la médiane. Cette stratégie est recommandée dans Kovar, Rao et Wu (1988) et dans Rao et Shao (1996). Cette procédure se fonde sur les étapes ci-dessous:

   - trier les observations de l'échantillon dans l'ordre ascendant;
   - accumuler la somme des poids d'enquête associés;
   - sélectionner la première observation pour laquelle la somme associée des poids dépasse la moitié du poids total.

2. Interpolation linéaire

   Une autre façon d'estimer la médiane d'une population consiste à grouper les données d'échantillon et à y interpoler la médiane de l'échantillon. Woodruff (1952) fournit la formule ci-dessous pour l'interpolation linéaire de la médiane d'un échantillon:

$$\hat{M} = F^{-1}(\frac{1}{2}\hat{N}) \approx ll + \left(\frac{\frac{1}{2}\hat{N} - \text{cf}}{f_i} \right) * (i) \qquad (3.1)$$

où

F = la fréquence cumulative de la caractéristique axée sur des poids d'échantillon

ll = la limite inférieure du casier contenant la médiane

$\hat{N}$ = le nombre total estimatif des éléments de la population

cf = la fréquence cumulative pour tous les intervalles précédant le casier contenant la médiane

f_i = la fréquence de la catégorie médiane (nombre total estimatif des éléments de la population de l'intervalle contenant la médiane)

i = la largeur du casier contenant la médiane.

C'est là la méthode utilisée pour le système actuel d'estimation de la variance pour la production de la SOC pour les estimations mensuelles, et c'est également la méthode d'interpolation linéaire utilisée pour le progiciel VPLX.

Nous avons considéré deux options pour le réglage de la taille de catégorie (largeurs de casier) pour l'interpolation. La première option consiste à élaborer des casiers en fonction de la caractéristique particulière à l'étude selon les données originales. La deuxième option comporte une transformation linéaire des données selon une échelle standard, un ensemble standard de casiers étant alors utilisé pour chaque caractéristique. Nous avons fait appel à la transformation linéaire ci-dessous:

$$X_i' = X_i * \frac{1{,}000}{Q_3} \qquad (3.2)$$

où Q_3 est le troisième quartile de la distribution de l'échantillon (estimé à l'aide des observations ordonnées et du poids d'échantillon décrits à la section 2.A.1). La médiane interpolée du X′ est multipliée par (Q_3/1 000) de façon à donner une médiane estimative selon l'échelle originale. [Si la distribution comporte des valeurs négatives (par exemple, une distribution du revenu net), on utilise alors $X_i'' = (X_i - X_{(1)}) * 1\ 000/Q_3(X_i - X_{(1)})$, où $X_{(1)}$ est la statistique d'ordre un et $Q_3(X_i - X_{(1)})$ est calculé à partir de la distribution de $(X_i - X_{(1)})$. Pour obtenir une médiane estimative selon l'échelle originale, on multiplie la médiane interpolée par $(Q_3(X_i - X_{(1)})/1\ 000)$ et on ajoute $X_{(1)}$.] Cette procédure correspond à la simple division de l'échantillon original de 0 à Q_3 en x casiers de largeur égale, le reste des données étant placé dans un casier qui, de par sa conception, est beaucoup plus grand que les autres.

Cette procédure est conçue en fonction de distributions symétriques ou positivement asymétriques (comme c'est habituellement le cas des données économiques). Les données qui se trouvent dans le dernier casier ne servent pas à estimer la médiane parce qu'elles sont plus grandes que Q_3, jugé éloigné de la médiane. Si l'on fondait la transformation linéaire sur Q_1 (le premier quartile), le casier qui contient la médiane pourrait être très rapproché du casier inférieur de la distribution. Dans un tel cas, la différence de variabilité entre une médiane interpolée et la médiane de l'échantillon serait petite.

Le fait d'utiliser les données originales afin d'élaborer les médianes permet de préparer des estimations et des ET prêtes pour la production. Il est difficile, toutefois, de déterminer la largeur de casier fixe appropriée. Plus la largeur des casiers est petite (largeur 1), et plus les estimations de la variance sont instables. L'augmentation de la largeur des casiers fait augmenter le biais de l'estimation attribuable à l'interpolation. La taille «optimale» du casier établit un équilibre entre le biais et la stabilité de l'estimation de la variance. Malheureusement, la largeur optimale du casier ne reste pas nécessairement constante d'un échantillon à l'autre. Souvent, la distribution évolue au fil du temps, et la largeur/emplacement des casiers dans l'échantillon devrait refléter ce changement d'échelle. De plus, la largeur optimale du casier peut différer pour diverses valeurs d'une variable de classement: par exemple, la largeur optimale du casier pour le prix de vente dans le Midwest diffère probablement de la largeur optimale du casier pour le prix de vente dans le Sud.

La transformation linéaire a été motivée par l'intérêt pour une largeur de casier dépendant de l'échantillon. Les largeurs de casier «standard» utilisées pour les données transformées inférieures à Q_3 ne sont pas standard pour l'échelle non transformée: la largeur du casier dépend des données. L'utilisation des données transformées linéairement suppose un travail comptable accru pour ce qui est des constantes d'échelle, mais permet facilement de changer l'échelle et la forme de la distribution.

Les figures 1 à 4 montrent l'effet de la transformation linéaire sur la largeur des casiers et leur emplacement pour deux distributions. Les figures 1 et 2 illustrent une distribution comportant un large éventail de valeurs de données, y compris quelques observations très grandes. Les figures 3 et 4 illustrent une distribution constituée surtout de petites valeurs de données.

La figure 1 présente un histogramme de la distribution originale de maisons vendues moyennant un financement conventionnel, avec une largeur de casier de 25 000$ [Nota: la taille du casier a été choisie purement pour faciliter la présentation, puisqu'il s'agit d'une distribution allongée.] La médiane de cette distribution est de 167 130$, avec un Q_3 de 225 000$. La figure 2 présente l'histogramme de la distribution transformée linéairement avec une largeur de casier de 50. Dans cet exemple, les casiers transformés de largeur 50 correspondent à des casiers de largeur 11 250$ dans l'échelle originale (225 000$/1 000$)*50). On se souviendra que la taille des casiers de données originales à l'étude est de 1 000$ et de 2 000$. Ainsi, les casiers de données transformées de largeur 4 auraient une largeur de 900$ dans l'échelle originale non transformée. On remarquera l'importante «pointe» du dernier casier, qui contient toutes les valeurs de l'échantillon supérieures à Q_3.

Ces figures illustrent également les différences de distribution des tailles d'échantillon d'un casier à l'autre pour les deux méthodes. L'utilisation de largeurs de casier fixes avec les données originales entraîne des tailles assez variables d'échantillon de casier (voir la Figure 1). Par contre, de par leur conception, les tailles d'échantillon à l'intérieur des casiers qui dépendent des données sont beaucoup plus uniformes pour tous les casiers sauf le dernier (voir la figure 2).

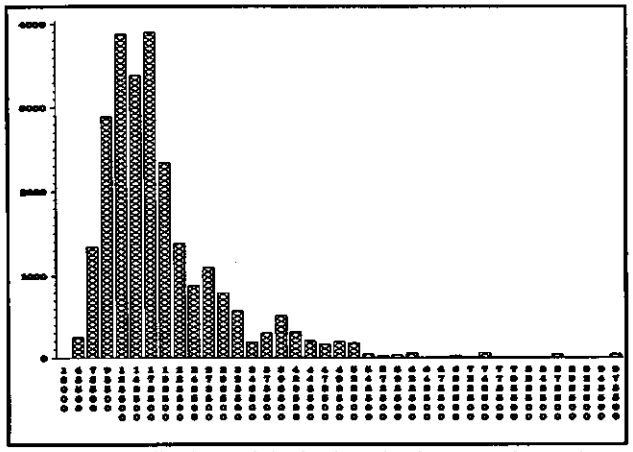


Figure 1: Distribution originale du prix de vente des maisons vendues ayant une largeur conventionnelle de casier de financement (largeur de casier = 25 000$).

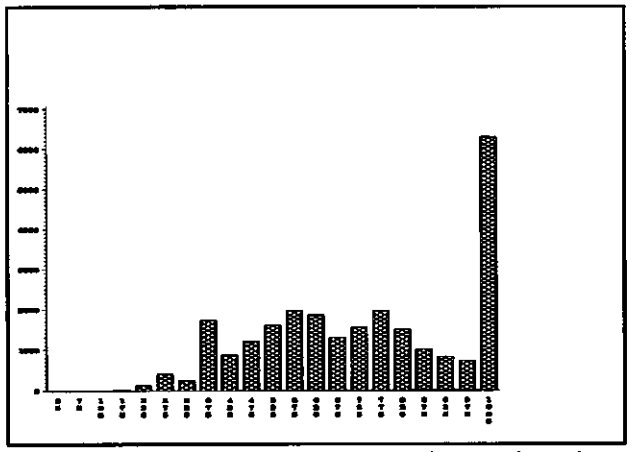


Figure 2: Distribution transformée du prix de vente des maisons vendues moyennant un financement conventionnel pour une largeur de casier correspondant à une largeur de casier = 50 de l'échelle non transformée = 11 250$.

La figure 3 présente un histogramme de la distribution originale des maisons vendues moyennant des prêts de la FHA, avec une largeur de casier de 4 000$ (encore une fois, la largeur de casier est choisie pour des raisons de présentation). La médiane de cette distribution est de 108 280$, et on a 124 990$ pour Q_3. La figure 4 présente l'histogramme de la distribution transformée linéairement avec une largeur de casier de 50. Dans cet exemple, les casiers transformés de largeur 50 correspondent à des casiers de largeur 6 250$ dans l'échelle originale, et les casiers des données transformées de largeur 4 auraient une largeur approximative de 500$ dans l'échelle non transformée originale.

Les figures 1 à 4 montrent la souplesse des casiers élaborés pour des données transformées linéairement. La taille de casier dans l'échelle non transformée peut être plus grande ou plus petite suivant l'écart des données. De plus, la taille des échantillons de casier dépendant des données est moins variable comparativement à celle qui est associée à des casiers fixes.

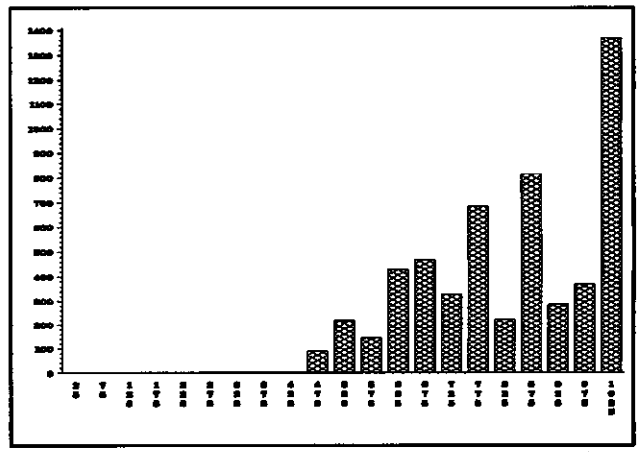


Figure 4: Distribution transformée du prix de vente des maisons vendues moyennant des prêts de la FHA (largeur de casier = 50). Largeur de bande dans l'échelle non transformée = 6 250$.

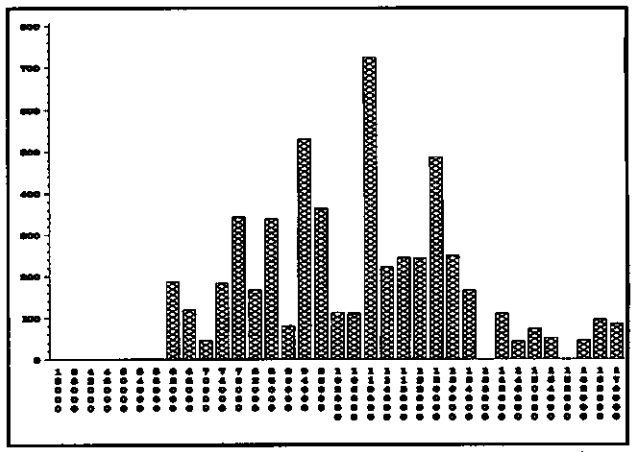


Figure 3: Distribution originale du prix de vente des maisons vendues moyennant des prêts de la FHA (largeur de casier = 4 000$).

Afin d'évaluer la première option d'interpolation (médianes interpolées selon des données originales), nous avons utilisé deux ensembles de largeurs de casier (tailles de classement): casiers de taille 2 000$ (la même largeur de casier utilisée pour le système actuel d'estimation de la variance pour la production) et casiers de taille 1 000$. [Nota: le logiciel VPLX d'estimation de la variance ne permet aucune taille de casier inférieure à 1 000$ puisque le nombre de catégories dépassait l'éventail admissible.] Après avoir examiné plusieurs mois d'estimations du prix de vente pour l'ensemble des États-Unis, nous avons supposé que le prix de vente médian serait toujours supérieur à 36 000$ et inférieur à 550 000$; le premier classement de données originales est donc toujours de (minimum à 35 999), le dernier classement de données originales étant toujours de (550 000 à maximun): cela donne 257 casiers de taille 2 000$ ou 514 casiers de taille 1 000$, plus un casier de taille 36 000$ et un casier dont la largeur dépend de la plus grande observation de l'échantillon. Un problème évident lié à l'emplacement de ces casiers est l'effet éventuel de l'inflation. Il est

concevable que, pour des catégories de financement spécial ou certaines régions, le prix de vente médian des maisons vendues se rapproche de 550 000$, ce qui entraînerait l'échec de l'interpolation.

Afin d'évaluer la deuxième option d'interpolation (médianes interpolées selon des données transformées), nous avons eu recours à trois ensembles de largeurs de casier: casiers de taille 4, 25 et 50. Les casiers de taille 4 ont été choisis semblables aux casiers de taille 2 000 pour ce qui est du nombre de casiers. Il y a 250 casiers de taille 4 pour les données transformées inférieures à Q_3, et un casier plus grand contenant toutes les données supérieures à Q_3. La sélection des largeurs 25 et 50 a été relativement arbitraire: nous avons choisi le casier de taille 50 afin d'avoir un total de 20 casiers pour les données inférieures à Q_3; nous avons choisi le casier de taille 25 pour examiner l'effet du doublement du nombre de casiers et de la réduction de moitié de la largeur des casiers pour les données inférieures à Q_3. La médiane des données transformées est toujours inférieure à 1 000, de sorte que le dernier classement de données transformées est toujours de (1 000 à maximum). Ainsi, par définition, le dernier casier contient jusqu'à 25% des données, et il est considéré comme plus large que les autres casiers.

B. Estimation de la variance

Nous avons utilisé la méthode de répétition MHS (modifiée avec demi-échantillon) (Fay 1989 et Judkins 1990) pour estimer la variance d'une médiane tel que mentionné dans la documentation publiée (par exemple, Rao, Wu et Yue (1992); Rao et Shao (1996); Kovacevic et Yung (1997) pour la répétition équilibrée; Judkins (1990) pour la répétition de type MHS). La répétition de type MHS est une variation de l'estimation équilibrée «traditionnelle» de la variance avec demi-échantillon décrite par Wolter (1985, 110-152). La répétition BRR (équilibrée avec demi- échantillon) est une méthode d'estimation de la variance conçue pour un plan de sondage de deux UPE par strate. Selon la méthode BRR, une répétition avec demi- échantillon est effectuée par sélection d'une unité de chaque paire et par pondération de l'unité sélectionnée par 2 (de façon qu'elle représente les deux unités). Ainsi, des estimations pour chaque UPE sont comprises dans chaque répétition même si la moitié d'entre elles ont une pondération zéro. Les répétitions (demi- échantillons) sont précisées à l'aide d'une matrice de Hadamard. On trouvera dans Wolter (1985, 114-115) une description détaillée de la procédure de répétition axée sur des matrices de Hadamard. La répétition de type MHS fait appel à des poids de répétition de 1,5 et de 0,5 au lieu de 2 et 0. L'erreur type d'une estimation de la médiane axée sur une répétition de type MHS est donnée par

$$\hat{SE}(\hat{Med}) = \sqrt{\frac{4}{R} * \sum_{r=1}^{R} (\hat{Med}_r - \hat{Med}_0)^2} \qquad (2.3)$$

où l'indice r désigne l'estimation de la médiane r répétée ($r=1,2,...,R$) et l'indice 0 désigne l'estimation de la médiane de l'échantillon complet. Cette expression comporte le chiffre quatre (4) dans le numérateur parce que l'EQM des estimations répétées est trop faible d'un facteur de $1/(1-0,5)^2$. Voir Judkins (1990).

Ni le plan SUP ni le plan NP n'est un plan de sondage à deux unités d'échantillonnage par strate. Au premier degré, on sélectionne une UPE par strate. Les deuxième et troisième degrés sont des échantillons systématiques, et souvent une seule unité par strate a été sélectionnée au deuxième degré. Une façon courante d'aborder le problème suscité par une unité d'échantillonnage par strate consiste à

- «diviser» les unités d'échantillonnage autoreprésentatives en deux panels par unité d'échantillonnage à l'aide de la méthode d'échantillonnage originale;
- former des strates groupées par appariement de deux (ou trois) unités d'échantillonnage non autoreprésentatives «semblables»; et
- appliquer la stratégie du demi-échantillon de façon que les éléments des demi-échantillons soient des panels à l'intérieur d'unités d'échantillonnage pour les unités d'échantillonnage autoreprésentatives, et qu'ils soient les unités d'échantillonnage de premier degré (UPE) à l'intérieur de strates groupées pour les unités d'échantillonnage non autoreprésentatives.

Le système actuel de variance pour la production de la SOC fait appel à un estimateur de Keyfitz (estimateur de différence apparié) pour un échantillon non autoreprésentatif et à un estimateur de formule d'échantillonnage approximatif pour un échantillon autoreprésentatif afin de produire des estimations uniformes de la variance (Luery 1990). Puisque les méthodologistes de la SOC avaient déjà groupé des strates non autoreprésentatives pour leur estimateur de différence apparié, une application de type BRR était un prolongement logique de la structure préexistante d'estimation de la variance. Pour une répétition de type MHS, nous trions les permis à l'intérieur de groupes d'unités d'échantillonnage prédéterminés en unités autoreprésentatives selon le code géographique et la date d'autorisation, et nous divisons systématiquement l'échantillon ordonné en deux panels comme le suggère Wolter (1985, 131). Bien que ce soit essentiellement la seule stratégie disponible pour le plan de sondage de la SOC, cette méthode risque de ne pas fournir des estimations correctes de la variance puisque des unités des deux panels sont corrélées (dans la méthode originale avec demi-échantillon, les deux UPE de la strate sont supposées indépendantes). On trouvera des détails supplémentaires sur l'attribution des répétitions dans Thompson (1998).

Le système de production de la SOC fait appel à la méthode de Woodruff (Woodruff 1952) afin d'évaluer l'erreur type d'une médiane. La méthode de Woodruff fait appel à l'erreur type estimative d'une proportion p ($p = 0,50$ pour une estimation de la médiane) avec projection de

l'intervalle ($\underline{p} \pm ET(\underline{p})$) sur la distribution cumulative des fréquences de façon à fournir la limite inférieure d'un intervalle de confiance de 62,86% pour la médiane (l'ET($\underline{p}$) peut être estimée à l'aide de méthodes de répétition). L'ET de la médiane est alors estimée par soustraction. Cette méthodologie a été plus ou moins fructueuse dans le passé d'après les analystes de l'enquête SOC.

4. RÉSULTATS DES DONNÉES EMPIRIQUES

Initialement, nous avons utilisé quatre mois des données de l'échantillon de la SOC afin d'examiner la variance des méthodes d'estimation de la médiane pour le prix de vente des maisons vendues: mars 1997, mai 1997, juin 1997 et juillet 1997. Nous avons préparé des médianes selon la région et selon le type de financement. Nous avons utilisé les mêmes poids que pour les systèmes de variance et d'estimation de la production de la SOC (stratifiés a posteriori pour l'échantillon SUP et non biaisés pour l'échantillon NP), les données des deux enquêtes étant groupées en vue de l'obtention de médianes. Chaque série d'estimations de la variance a été produite à l'aide de 200 répétitions.

Nous avons observé que les six méthodes d'estimation de la médiane ont produit des estimations très semblables, donnant lieu toutefois à trois séries distinctes d'ET: une série pour la médiane de l'échantillon, une série pour les médianes interpolées selon les données originales (largeur de casier fixe) et une série pour les médianes interpolées selon les données transformées (largeur de casier qui dépend des données). Il n'y avait pas de lien net entre la largeur du casier et les estimations de l'ET pour les deux séries de médianes interpolées. En effet, pour un même type de données (originales ou transformées), les ET étaient toutes très rapprochées. De toute évidence, il y avait une transformation linéaire et un effet d'interpolation. Aucune des méthodes d'estimation de la médiane n'a donné des erreurs types ressemblant aux erreurs types publiées, de sorte que l'on ne pouvait pas conclure qu'il y avait uniformité des données publiées.

De plus, il semblerait que les ET publiées de la méthode de Woodruff sont sous-estimées ou du moins inappropriées pour le plan d'échantillonnage utilisé. Kovar, Rao et Wu (1988) ont comparé les ET de Woodruff et celles de la méthode BRR, et ils ont trouvé que les deux méthodes avaient des propriétés semblables, sauf dans le cas des échantillons stratifiés, où les strates se fondent sur des variables séparées fortement corrélées (comme le plan de la SOC). Dans ce cas, l'ET de Woodruff est souvent trop petite, et ils ont conclu que les méthodes BRR (sic) sont plus robustes vis-à-vis de diverses structures démographiques, puisque l'erreur est extraite directement des répétitions. Lorsque les ET de Woodruff du système de production faisaient appel à l'ET($\underline{p}$) calculée directement, les ET de Woodruff étaient généralement plus petites que les ET répétées.

Les résultats empiriques nous ont laissés perplexes. Nous avions trois séries distinctes d'estimations de la variance et aucun «étalon-or» permettant de les mesurer. Puisque nos résultats empiriques n'étaient pas concluants, nous avons mené une étude de simulation de Monte Carlo afin d'évaluer les propriétés des estimations de la variance de type MHS préparées à l'aide des différents estimateurs de la médiane.

5. COMPARAISON DES ÉTUDES DE SIMULATION

A. Procédure pour les études de simulation

Nous avons créé quatre populations artificielles finies fondées sur une analyse des données de quatre populations de l'échantillon de la SOC: une population du type de financement (financement conventionnel) et trois populations régionales (Midwest = région 2, Sud = région 3 et Ouest = région 4). Ces populations représentaient divers types des populations de la SOC servant à produire des estimations. À noter que la population du type de financement de la SOC n'est pas indépendante des populations des régions de la SOC.

Afin d'obtenir une approximation de la population finie du prix de vente des maisons vendues, nous avons créé w_i enregistrements pour chaque unité d'échantillonnage $\underline{i}$, où w_i est le poids d'échantillon associé à l'unité $\underline{i}$. Des distributions normales logarithmiques ont permis d'obtenir une approximation de la distribution du prix de vente des maisons vendues à une seule unité. La distribution normale logarithmique comporte la fonction de densité

$$f(y) = \frac{1}{y-\theta}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{(\log(y-\theta)-\zeta)}{\sigma}\right)^2\right)$$

$$\text{pour } \theta < y < \infty$$

où θ est le paramètre de seuil, ζ est le paramètre d'échelle et σ est le paramètre de forme.

À l'aide de nos modèles, nous avons créé quatre populations simulées finies à deux variables ayant une corrélation prévue de $\rho = 0{,}6$ à l'aide de la méthode décrite dans Naylor, Balintfy, Burdick et Chu (1968, 99). La première des deux variables de chaque population représentait le prix de vente des maisons vendues, et elle a été obtenue par la création d'une variable normale aléatoire avec moyenne ζ et variance σ^2 à l'aide des paramètres déterminés ci-dessus, suivie d'une élévation à la puissance et d'un décalage à l'aide de paramètres d'emplacement appropriés (θ). La deuxième variable a servi à former des strates et des grappes de premier degré. Cette variable comportait une distribution normale standard marginale; elle a été obtenue par la création indépendante d'une deuxième valeur normale aléatoire standard, multipliée par 0,8, ce terme étant additionné à 0,6 × la variable aléatoire normale

standard servant à créer la variable du prix de vente. Les percentiles, l'asymétrie de l'échantillon et l'aplatissement de l'échantillon de la variable du prix de vente de chaque population simulée se rapprochaient beaucoup de la statistique correspondante de la population originale, surtout lorsque les valeurs aberrantes étaient supprimées à l'aide de la règle robuste des clôtures extérieures décrite dans Hoaglin et Iglewicz (1987). La taille de chaque population était le $\hat{N}$ estimé à partir des populations de l'échantillon. Les paramètres de modèle, les corrélations de l'échantillon (entre les prix de vente simulés et la variable stratifiante), la taille de la population (N) et la taille des échantillons (n) sont présentés dans le tableau 1.

Après avoir créé les populations finies, nous avons formé 50 strates de taille égale dans chaque population, puis nous avons tiré deux séries d'échantillons pour deux plans d'enquête différents:

- Le premier s'inspire de l'échantillon SUP des permis pour quatre unités de logement ou moins des bureaux de permis dans des UPE autoreprésentatives (28% à peu près de l'échantillon de la SOC). Dans cette étude, nous avons tiré 5 000 échantillons stratifiés aléatoires sans remise pour chaque population simulée à l'aide du même taux d'échantillonnage dans chaque strate. Afin d'effectuer une répétition de type MHS, nous avons trié l'échantillon à l'intérieur de chaque strate selon la variable stratifiante et nous avons divisé systématiquement l'échantillon en deux panels.
- Le deuxième plan s'inspire de l'échantillon SUP des permis pour quatre unités de logement ou moins des bureaux de permis non autoreprésentatifs dans des UPE autoreprésentatives et des bureaux de permis autoreprésentatifs dans des UPE non autoreprésentatives (40% à peu près de l'échantillon de la SOC). Dans cette étude, nous avons tiré 5 000 échantillons à deux degrés pour chaque population simulée. Le premier degré est un échantillon stratifié aléatoire sans remise de deux UPE par strate ($N_h = 5$). Le deuxième degré est un échantillon systématique d'unités à l'intérieur des UPE. Puisque toutes les UPE ont la même taille, cette étude ne tient pas compte de l'échantillonnage PPT de la SOC et ne comprend pas le groupement des unités du premier degré. La répétition de type MHS fait appel aux unités d'échantillonnage (UPE) du premier degré à l'intérieur de la même strate. Les poids répétés ne rendent pas compte d'importantes fractions d'échantillonnage au premier degré de la sélection comme le recommande Wolter (1985, 122), de sorte que les estimations de la variance comportent probablement toutes un biais vers le haut.

Nous n'avons pas tenté de simuler l'échantillon SUP des permis pour quatre unités de logement ou moins dans des UPE non autoreprésentatives et des bureaux de permis non autoreprésentatifs (échantillon à trois degrés, 25% à peu près de l'échantillon de la SOC); l'échantillon SUP des permis pour cinq unités de logement ou plus (2% environ de l'échantillon de la SOC); ou l'échantillon NP des DD (5% environ de l'échantillon de la SOC). L'échantillon à trois degrés, bien que non négligeable dans la SOC, est rarement utilisé dans d'autres enquêtes du Bureau of the Census, et les deux autres secteurs du plan de la SOC ne comptent pas suffisamment dans les estimations pour justifier une étude distincte.

Afin d'examiner la précision de chaque procédure d'estimation de la médiane pour des échantillons répétés, nous avons estimé les erreurs quadratiques moyennes (EQM) et les erreurs absolues moyennes (EAM) empiriques des 5 000 échantillons en fonction de ce qui suit:

ME: la médiane de l'échantillon pour chaque demi-échantillon

IO2000: les médianes interpolées selon les données originales, casiers de taille 2 000 (largeur de casier fixe)

IO1000: les médianes interpolées selon les données originales, casiers de taille 1 000 (largeur de casier fixe)

IT4: les médianes interpolées selon les données transformées linéairement, casiers de taille 4 (largeur de casier qui dépend des données)

Tableau 1
Caractéristiques des populations simulées et taille des échantillons stratifiés

		Paramètres de prix de vente			Corrélation (stratificateur, prix de vente)	Taille de la population	Taille de l'échantillon
Population	Distribution	θ	σ	ζ	ρ	N	n
Financement conventionnel	normale logarithmique	27 578	0,4895	11,84	0,57030	25 150	500
Midwest	normale logarithmique	31 801	0,5957	11,69	0,55835	6 500	150
Sud	normale logarithmique	29 414	0,5549	11,55	0,55929	14 550	300
Ouest	normale logarithmique	53 781	0,5822	11,59	0,55525	11 550	250

IT25: les médianes interpolées selon les données transformées linéairement, casiers de taille 25 (largeur de casier qui dépend des données)

IT50: les médianes interpolées selon les données transformées linéairement, casiers de taille 50 (largeur de casier qui dépend des données)

La transformation linéaire a eu lieu une fois pour les procédures IT4, IT25 et IT50. Les données originales ont été transformées à l'aide de l'échantillon Q_3 au complet, et ces données transformées ont été attribuées aux demi- échantillons (y compris la répétition zéro, l'échantillon au complet). Le tableau 2 présente la médiane et le troisième quartile de chaque population finie, de même que les largeurs de casier de l'échelle originale pour les données transformées.

Nous avons calculé $M(\zeta_i)$, l'erreur quadratique moyenne empirique de la procédure d'estimation de la médiane i sous la forme

$$M(\zeta_i) = \frac{\sum_r (\zeta_{ri} - \bar{\zeta}_i)^2}{5\ 000} + (\bar{\zeta}_i - \zeta_p)^2$$

$$= \hat{\sigma}^2(\zeta_i) + \hat{\text{bias}}^2(\zeta_i) \qquad (5.1)$$

où ζ_{ri} est la médiane estimée pour l'échantillon r et l'estimateur i, $\bar{\zeta}_i$ est la moyenne du ζ_{ri}, et ζ_p est la médiane de la population. C'est là l'erreur quadratique moyenne empirique décrite dans Judkins (1990).

Nous avons calculé l'erreur absolue moyenne (MAE) de chaque procédure d'estimation de la médiane i sous la forme

$$\text{MAE}(\zeta_i) = \frac{\sum_r |\zeta_{ri} - \zeta_p|}{5\ 000} \qquad (5.2)$$

suivant la définition de DeGroot (1986, 209-211).

Afin de comparer les propriétés d'estimation de la variance des différentes méthodes d'estimation de la médiane, nous avons calculé une estimation de la variance de type MHS (v_{ij}) correspondant à chaque procédure d'estimation de la médiane i pour 1 000 des 5 000 échantillons. Ces estimations de la variance ont été comparées relativement à ce qui suit:

Biais relatif $(\sum_{j=1} v_{ij}/1\ 000)/M(\zeta_i) - 1$

Stabilité relative $[(\sum_{j=1} (v_{ij} - M(\zeta_i))^2/1\ 000)]^{1/2}/M(\zeta_i)$

Taux d'erreur Nombre d'échantillons où ($\zeta_p < \theta_{Li}$ ou $\zeta_p > \theta_{Ui}$)/1 000 où

θ_{Li} est l'extrémité inférieure d'un intervalle de confiance de 90% et

θ_{Ui} est l'extrémité supérieure d'un intervalle de confiance de 90%.

Ces critères ont été utilisés dans Kovar, Rao et Wu (1988) et dans Rao et Shao (1996). Le biais relatif est une mesure du biais de l'estimation de la variance comme proportion de la véritable erreur quadratique moyenne. La stabilité est une mesure de la variance des estimations de la variance; elle est une approximation d'un c.v. de l'estimation de la variance v_i. À noter que la stabilité relative n'est pas l'erreur quadratique moyenne relative définie dans Wolter (1985, 297) qui utilise le carré de l'EQM dans le dénominateur. En présence d'un estimateur «optimal» de la variance, le biais relatif aussi bien que la stabilité relative se rapprocheront de zéro, et le taux d'erreur sera de 10%.

B. Résultats

1. Comparaison des procédures d'estimation de la médiane

   Le tableau 3 présente la racine de l'EQM empirique, l'erreur type (ET), le biais et l'erreur absolue moyenne (EAM) de chaque procédure d'estimation de la médiane pour les deux études de simulation. Chacune de ces statistiques a été calculée à partir de 5 000 échantillons.

   Ces résultats ont confirmé nos soupçons relativement à l'analyse des données empiriques décrite ci-dessus. Au moins pour ce qui est du prix de vente, les six procédures d'estimation de la médiane donnent des résultats à peu près identiques, la racine des EQM et les erreurs absolues moyennes étant à peu près égales d'une procédure à l'autre pour chaque population.

2. Comparaison des propriétés d'estimation de la variance avec répétition de type MHS des procédures d'estimation de la médiane.

Tableau 2
La médiane, le troisième quartile et la largeur des casiers pour l'échelle originale des données simulées transformées

Population	Médiane	Q_3	Largeur de casier		
			4	25	50
Financement conventionnel	167 173	222 263	889	5 557	11 113
Midwest (région 2)	151 312	210 647	843	5 266	10 532
Sud (région 3)	133 745	180 868	723	4 522	9 043
Ouest (région 4)	162 130	214 320	857	5 358	10 716

Tableau 3
La racine de l'EQM empirique, l'erreur type (ET), le biais et l'erreur absolue moyenne (EAM) pour les procédures d'estimation de la médiane

Population	Procédure d'estimation de la médiane	Échantillon à un degré non en grappes				Échantillon à deux degrés en grappes			
		Racine de l'EQM	ET	Biais	EAM	Racine de l'EQM	ET	Biais	EAM
Financement conventionnel	SM	3 345	3 345	-12	2 671	3 389	3 374	324	2 733
	IO2000	3 320	3 316	161	2 698	3 346	3 341	189	2 685
	IO1000	3 387	3 368	-354	2 642	3 431	3 420	-278	2 774
	IT4	3 351	3 340	273	2 673	3 378	3 364	311	2 719
	IT25	3 304	3 293	276	2 617	3 337	3 321	322	2 664
	IT50	3 282	3 265	329	2 606	3 305	3 283	375	2 636
Région 2 Midwest	SM	6 316	6 287	-598	4 966	6 273	6 228	-753	4 959
	IO2000	6 276	6 275	-127	4 992	6 335	6 207	-1 271	5 029
	IO1000	6 343	6 297	-767	4 939	6 526	6 280	-1 774	5 204
	IT4	6 372	6 363	328	5 004	6 294	6 228	-908	4 979
	IT25	6 273	6 272	127	4 937	6 270	6 154	-1 199	4 971
	IT50	6 220	6 218	160	4 936	6 224	6 114	-1 164	4 966
Région 3 Sud	SM	3 670	3 658	301	2 931	3 835	3 752	796	3 054
	IO2000	3 708	3 669	539	2 998	3 796	3 739	656	3 011
	IO1000	3 742	3 740	101	2 941	3 806	3 804	212	3 066
	IT4	3 718	3 662	639	2 951	3 814	3 736	766	3 028
	IT25	3 699	3 638	669	2 924	3 793	3 711	787	2 992
	IT50	3 692	3 616	745	2 912	3 778	3 680	856	2 970
Région 4 Ouest	SM	4 385	4 382	-140	3 509	4 394	4 351	616	3 506
	IO2000	4 425	4 421	185	3 578	4 362	4 339	449	3 487
	IO1000	4 477	4 469	-258	3 530	4 411	4 410	-57	3 535
	IT4	4 414	4 403	318	3 514	4 383	4 342	599	3 494
	IT25	4 376	4 364	315	3 460	4 334	4 296	573	3 439
	IT50	4 367	4 350	391	3 455	4 320	4 271	644	3 436

Lorsque nous avons examiné les propriétés d'estimation de la variance pour chaque procédure, les résultats ont été bien différents. Comme pour notre analyse des données empiriques, nous avions trois séries de résultats très distinctes. Le tableau 4 résume les trois mesures de comparaison pour les estimations de la variance des quatre populations. Les numérateurs pour le biais relatif et la stabilité et les taux de couverture se fondent sur 1 000 échantillons. Le dénominateur pour le biais relatif et la stabilité («vérité») se fondent sur 5 000 échantillons. L'astérisque (*) de la dernière colonne du tableau 4 indique que le taux d'erreur est appréciablement différent du taux d'erreur nominal de 0,10 en présence de l'approximation normale de la distribution binominale pour le niveau de confiance de 90%.

Dans les deux études, les estimations de la variance des médianes interpolées selon les données transformées donnent les meilleurs résultats pour ce qui est du biais relatif et de la stabilité. En particulier:

- Les estimations de la variance des médianes interpolées selon les données transformées (IT4, IT25, IT50) comportent le biais relatif le plus petit. La différence liée à la méthode d'estimation est bien marquée dans trois des quatre populations, où le biais relatif le plus important des médianes interpolées selon les données transformées représente moins de la moitié de la taille du plus petit biais relatif des médianes interpolées selon les données originales et des médianes de l'échantillon. Ces résultats sont remarquablement solides pour le plan à deux degrés en grappes, puisque les estimations de la variance sont supposées comporter un biais vers le haut (voir la section 5.A).

- Les estimations de la variance des médianes interpolées offraient la meilleure stabilité. Les estimations de la variance de la médiane de l'échantillon offraient la stabilité la plus médiocre pour les quatre populations. Ce résultat était prévu à cause de l'effet de lissage de l'interpolation. Encore une fois, les médianes interpolées selon les données transformées ont donné, de façon générale, de meilleurs résultats que les médianes interpolées selon les données originales, bien que la différence ne soit pas aussi marquée que dans le cas du biais relatif. En général, la stabilité est proche avec les trois largeurs de casier pour les médianes interpolées selon les données transformées.

Tableau 4
Le biais relatif et la stabilité relative pour des estimations de la variance et les taux d'erreur pour des intervalles de confiance de 90%

Population	Procédure d'estimation de la médiane	Plan à un degré sans grappes			Plan à deux degrés avec grappes		
		Biais relatif	Stabilité relative	Taux d'erreur	Biais relatif	Stabilité relative	Taux d'erreur
Financement conventionnel	SM	0,19	0,69	11,0%	0,11	0,58	15,1%*
	IO2000	0,25	0,35	6,9%*	0,25	0,37	9,0%
	IO1000	0,21	0,32	7,0%*	0,19	0,33	9,3%
	IT4	0,06	0,25	10,0%	0,06	0,27	11,3%
	IT25	0,07	0,25	10,9%	0,06	0,27	11,8%*
	IT50	0,05	0,26	9,5%	0,05	0,28	12,1%*
Région 2 Midwest	SM	0,57	1,24	7,3%*	0,41	1,07	7,9%*
	IO2000	0,33	0,44	6,9%*	0,23	0,35	8,6%
	IO1000	0,30	0,42	7,0%*	0,17	0,30	8,7%
	IT4	0,15	0,41	10,1%	0,14	0,41	11,5%*
	IT25	0,16	0,40	9,8%	0,11	0,37	10,4%
	IT50	0,15	0,42	9,0%	0,11	0,40	10,4%
Région 3 Sud	SM	0,30	0,88	12,4%*	0,15	0,71	11,1%
	IO2000	0,31	0,42	6,7%*	0,28	0,39	7,5%*
	IO1000	0,29	0,40	6,7%*	0,27	0,38	7,3%*
	IT4	0,04	0,29	11,0%	0,01	0,28	10,8%
	IT25	0,02	0,28	11,0%	-0,01	0,27	11,3%
	IT50	0,01	0,29	11,1%	-0,02	0,28	11,9%*
Région 4 Ouest	SM	0,39	0,98	8,9%	0,25	0,79	8,6%
	IO2000	0,32	0,42	6,2%*	0,31	0,41	5,2%*
	IO1000	0,29	0,39	6,2%*	0,28	0,38	5,2%*
	IT4	0,11	0,32	8,6%	0,10	0,31	7,6%*
	IT25	0,10	0,31	9,4%	0,09	0,30	7,5%*
	IT50	0,08	0,31	9,5%	0,08	0,31	8,3%*

Les résultats pour la couverture des intervalles de confiance de chaque procédure d'estimation de la médiane ne sont pas aussi uniformes; ils varient selon le plan de sondage. En présence du plan à un degré sans grappes, les intervalles de confiance établis à partir des ET et des médianes interpolées selon les données transformées offrent la meilleure couverture. Dans chaque population, les casiers qui dépendent des données (toutes largeurs) offrent une couverture presque nominale ou supérieure; en effet, aucun de ces taux d'erreur n'est statistiquement différent de la valeur nominale de 10%. Les intervalles de confiance établis à partir des ET et des médianes interpolées selon les données originales comportent des valeurs extrêmement prudentes. Ici, le biais positif des estimations de la variance donne une largeur inutilement grande à ces intervalles, réduisant ainsi la possibilité de résultats intéressants. La couverture avec la médiane de l'échantillon est irrégulière.

Certains de ces profils de couverture sont répétés dans le plan à deux degrés avec grappes. Encore une fois, la couverture avec la médiane de l'échantillon est irrégulière, et les taux de couverture pour les intervalles de confiance établis à partir des médianes interpolées selon les données originales sont supérieurs à la valeur nominale (bien significativement supérieurs à la valeur nominale dans deux populations seulement). Le profil du taux d'erreur est bien différent pour les médianes interpolées selon les données transformées. Dans toutes les populations sauf celle de la région 4, les taux de couverture des trois procédures sont inférieurs à la valeur nominale. Toutefois, pour les casiers de largeur 4 et 25, un seul taux d'erreur est significativement supérieur à 10%; pour les casiers de largeur 50, deux de ces trois taux d'erreur sont significativement supérieurs à 10%. Toutes les médianes des données interpolées comportent des taux d'erreur significativement plus petits que la valeur nominale dans la population de la région 4; en conformité avec les résultats de l'autre population, les taux d'erreur pour les médianes interpolées selon les données originales sont les plus éloignés de 10%.

Dans les deux études, les médianes interpolées selon les données transformées offrent très nettement de meilleures propriétés d'estimation de la variance pour ce qui est du biais relatif et de la stabilité relative, peu importe la largeur du casier. De même, dans les deux études, les médianes interpolées selon les données transformées à l'aide de casiers de largeur 4 ou de largeur 25 offrent une excellente couverture des intervalles de confiance. Puisque les médianes interpolées selon les données transformées à l'aide de casiers de largeur 50 ou de largeur 25 ont donné les «meilleurs» estimateurs pour ce qui est de la racine de l'EQM et de l'EAM dans les deux études, il semble que l'utilisation de l'interpolation linéaire pour les données transformées comportant des casiers de largeur 25 soit la meilleure procédure d'estimation de la médiane pour ce qui est des propriétés d'estimation et d'estimation de la variance.

6. CONCLUSION

Nous avons examiné l'effet du recours à des variations de deux méthodes d'estimation du prix de vente médian des maisons vendues: l'estimation directe et l'interpolation linéaire. L'interpolation linéaire suppose un classement de données continues en casiers de largeur standard. Cette largeur peut être arbitraire, peut différer considérablement selon le domaine et peut changer à mesure que la distribution de l'échantillon évolue au fil du temps. La transformation linéaire fondée sur le troisième quartile a semblé corriger cette difficulté. En présence des données transformées, la largeur et l'emplacement des casiers dans la distribution changent en fonction des données.

Nos résultats empiriques indiquent que le choix de la méthode exerce une influence marquée sur des estimations de la variance pour la répétition de type MHS. Notre étude de simulation a permis d'examiner les propriétés des différentes procédures d'estimation de la médiane pour les estimations de la variance avec répétition de type MHS. Dans les quatre populations simulées, les médianes interpolées selon les données transformées (largeurs de casier qui dépendent des données) ont fourni les meilleurs résultats, habituellement dans une large mesure. Ce qui plus est, cette méthode réduit grandement la surestimation de la variance. L'utilisation de casiers de largeur 25 pour l'échelle transformée (41 casiers au total) a donné les meilleures estimations du prix de vente médian et de la variance, pour une répétition de type MHS, et c'est elle que nous recommandons pour la Survey of Construction.

La méthode recommandée offre plusieurs avantages. Tout d'abord, elle est adaptable. Elle fonctionne bien pour un choix de distributions, puisque la largeur des casiers eux-mêmes dépend de la distribution en question. Deuxièmement, elle permet d'économiser les ressources informatiques puisque l'on évite le tri de demi-échantillons. Troisièmement, les intervalles qui dépendent des données se laissent aisément intégrer à un logiciel généralisé de traitement des enquêtes. Enfin, elle donne de meilleures estimations et de meilleures estimations de la variance avec répétition de type MHS (du moins pour le prix de vente des maisons vendues). Nous estimons que ces résultats se laissent généraliser pour d'autres distributions continues également, mais il convient bien sûr de vérifier cette conjecture avec d'autres ensembles de données. On pourrait poursuivre la recherche, par exemple en examinant la relation entre la taille de l'échantillon et la précision des estimations de la médiane, en examinant d'autres tailles de casier et en observant la robustesse de la procédure recommandée à l'aide de diverses procédures d'estimation de la variance avec répétition.

REMERCIEMENTS

Les auteurs tiennent à remercier Elizabeth Huang et James Fagan du Bureau of the Census des États-Unis, de même que deux examinateurs anonymes et le rédacteur associé de leurs généreuses remarques au sujet de versions antérieures du présent manuscrit, ainsi que J.N.K. Rao de ses remarques utiles sur l'étude de simulation originale. Le présent exposé décrit les résultats de recherches et d'analyses entreprises par le personnel du Bureau of the Census. Il a subi un examen moins minutieux que les publications officielles du Bureau of the Census. La diffusion du présent rapport doit permettre d'informer les chercheurs intéressés et de stimuler la discussion.

BIBLIOGRAPHIE

DEGROOT, M. (1986). *Probability and Statistics*. Reading, MA: Addison-Wesley Publishing, Inc.

FAY, R.E. (1989). Theory and application of replicate weighting for variance calculations. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

FAY, R.E. (1995). VPLX: Variance Estimation for Complex Surveys. Program Documentation. Non publié Bureau of the Census Report.

HOAGLIN, D.C., et IGLEWICZ, B. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association*, 83, 1147-1149.

JUDKINS, D.R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6, 223-239.

KOVAR, J.G., RAO, J.N.K, et WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *La revue canandienne de statistique*, 16, 25-45.

KOVACEVIC, M., et YUNG, W. (1997). Estimation de la variance des mesures de l'inégalité et de la polarisation du revenu - Étude empirique. *Techniques d'enquête*, 23, 47-59.

LUERY, D.M. (1990). *Survey of Construction Technical Paper.* Ébauche non publié Bureau of the Census documentation interne.

NAYLOR, T.H., BALINTFY, J.L., BURDICK, D. S., et CHU, K. (1968). *Computer Simulation Techniques*. New York: John Wiley and Sons, Inc.

RAO, J.N.K., WU, C.F.J., et YUE, K. (1992). Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes. *Techniques d'enquête*, 18, 225-234.

RAO, J.N.K., et SHAO, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association*, 91, 343-348.

THOMPSON, K.J. (1998). *Evaluation of Modified Half-Sample Replication for Estimating Variances for the Survey of Construction (SOC)*. Washington, DC: U.S. Bureau of the Census. (Rapport technique #ESM-9801, disponible Economic Statistical Methods and Programming Division).

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag, Inc.

WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Techniques d'enquête, décembre 2000
Vol. 26, n° 2, pp. 185-195
Statistique Canada

L'effet de divers plans de renouvellement sur la variance d'échantillonnage des estimations désaisonnalisées et des estimations de la tendance

C.H. McLAREN et D.G. STEEL[1]

RÉSUMÉ

Dans les domaines social et économique, de nombreuses séries chronologiques sont fondées sur des enquêtes par sondage à plan d'échantillonnage complexe. Or, le plan d'échantillonnage influence les propriétés de la série chronologique. Plus précisément, le chevauchement de l'échantillon d'une période à l'autre influence la variabilité de la série chronologique d'estimations basées sur des données d'enquête, ainsi que les estimations désaisonnalisées et les estimations de la tendance produites d'après ces séries chronologiques. La variante X11 de la méthode du US Census Bureau et le programme X11ARIMA, que l'on utilise couramment pour produire des estimations désaisonnalisées, permettent aussi de produire des estimations de la tendance. Le présent article décrit les effets de divers plans de chevauchement des échantillons sur la variance d'échantillonnage des estimations désaisonnalisées et des estimations de la tendance calculées d'après des séries chronologiques fondées sur des enquêtes par sondage.

MOTS CLÉS: X11; X11ARIMA; désaisonnalisation; estimation de la tendance; plan de renouvellement.

1. INTRODUCTION

Nombre de séries chronologiques importantes sont produites d'après les données d'enquêtes par sondage répétées dont le plan de chevauchement des échantillons d'une période à l'autre est complexe. Compte tenu de l'échantillonnage, la variabilité des séries chronologiques estimées tient, en partie, aux erreurs d'échantillonnage. Or, pour nombre de séries, ces erreurs sont une source importante de variabilité. Donc, le plan d'échantillonnage, en particulier le plan de chevauchement des échantillons, influe sur la variabilité de la série chronologique d'estimations d'enquête.

L'étude des séries chronologiques se concentre de plus en plus sur l'évaluation des courbes sous-jacentes de variation ou des tendances par analyse des séries désaisonnalisées. La plupart des bureaux gouvernementaux de la statistique produisent depuis longtemps des séries de données désaisonnalisées. Selon Kenny et Durbin (1982), les analystes de politiques déclarent souvent qu'ils s'intéressent davantage aux tendances sous-jacentes qu'aux fluctuations irrégulières des valeurs mensuelles non désaisonnalisées. Smith (1997) est du même avis. Pendant plus de dix ans, l'Australian Bureau of Statistics (ABS) a publié des séries d'estimations de la tendance obtenues par application de la méthode des moyennes mobiles de Henderson (MMH) aux séries désaisonnalisées pour en lisser les composantes irrégulières (ABS 1987). D'autres bureaux gouvernementaux de la statistique produisent des estimations de la tendance par diverses méthodes (Knowles 1997). Puisqu'on les produit par application de certains traitements à la série originale, les estimations désaisonnalisées et les estimations de la tendance sont également influencées par les erreurs d'échantillonnage. Suivant Bell et Kramer (1999), la composante dominante de la variance des estimations désaisonnalisées est souvent celle due à l'erreur d'échantillonnage. Dans certains cas, les séries sont fondées sur des échantillons indépendants au fil du temps, mais, en général, les échantillons se chevauchent dans une certaine mesure d'une période à l'autre en vue de réduire les coûts et l'écart-type des estimations de la variation entre deux périodes consécutives (Kish 1998).

Un aspect essentiel de l'élaboration du plan d'échantillonnage d'une enquête répétée est le plan de renouvellement, autrement dit, le plan d'intégration d'une unité particulière dans l'enquête au fil du temps, qui déterminera le chevauchement des échantillons. L'objectif de la présente étude est de déterminer les effets du plan de renouvellement choisi sur la variance d'échantillonnage des estimations désaisonnalisées et des estimations de la tendance obtenues par la variante X11 de la méthode du US Census Bureau (méthode Census X11) mise au point par Shiskin, Young et Musgrave (1967) et par la méthode X11ARIMA mise au point par Dagum (1980 et 1988). On se concentre ici sur le niveau et sur la variation d'une période à l'autre des estimations désaisonnalisées et des estimations de la tendance.

2. PLANS DE RENOUVELLEMENT

Considérons une série chronologique univariée dont les valeurs $y_t, t = 1, \ldots T$ proviennent d'une enquête par sondage répétée. La valeur observée au temps t est reliée à la valeur réelle de la série dans la population finie, Y_t, par

$$y_t = Y_t + e_t$$

[1] C.H. McLaren et D.G. Steel, School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522, Australia. Courrier électronique: craigmc@uow.edu.au, dsteel@uow.edu.au.

où e_t est l'erreur d'échantillonnage. On considère que la série Y_t comprend les composantes de la tendance-cycle, de la saisonnalité et des jours irréguliers T_t, S_t et I_t, si bien que

$$y_t = T_t + S_t + I_t + e_t.$$

Dans certains cas, une décomposition multiplicative peut être plus appropriée. Nombre de bureaux de la statistique produisent des séries désaisonnalisées en essayant d'estimer S_t et de l'éliminer de la série, habituellement au moyen d'une combinaison de filtres linéaires. Les méthodes utilisées le plus couramment sont la méthode Census X11 mise au point par Shiskin et coll. (1967) et la variante X11ARIMA élaborée par Dagum (1980 et 1988). Findley, Monsell, Otto, Bell et Pugh (1998) ont décrit d'autres améliorations qui sont intégrées à la variante X12ARIMA. L'ABS publie aussi des estimations de la tendance obtenues par application des MMH à la série désaisonnalisée et encourage les utilisateurs à interpréter la série en se basant sur ces estimations de la tendance (Linacre et Zarb 1991; ABS 1993). Les moyennes mobiles de Henderson (1916), mises au point au départ pour des calculs actuariaux, sont utilisées dans X11, X11ARIMA et X12ARIMA pour extraire la tendance des séries aux fins de la désaisonnalisation. Kenny et Durbin (1982) et Gray et Thomson (1996) expliquent le calcul des MMH. L'utilisateur des séries peut aussi produire des estimations de la tendance en appliquant les filtres aux estimations désaisonnalisées publiées. Kenny et Durbin (1982) font remarquer que la définition de la tendance n'est pas unique et que l'on peut choisir des filtres différents selon le degré de lissage et de sensibilité souhaité. Knowles et Kenny (1997) ont étudié des méthodes d'estimations de la tendance en vue de produire des séries statistiques officielles. Pour les séries mensuelles, ils recommandent l'utilisation des MMH, en choisissant un filtre de 13 ou de 23 termes selon la volatilité de la série en question.

La structure d'autocorrélation de la série observée est déterminée par l'autocorrélation des séries Y_t et e_t, structure qui, à son tour, influe sur l'estimation des composantes saisonnières, de tendance et inégalières. On peut estimer la structure de covariance de la série d'erreurs d'échantillonnage, e_t, d'après les données d'enquête au niveau des unités d'échantillonnage. Le calcul de ces estimations permet d'estimer la variance d'échantillonnage des séries estimatives de la tendance, de la saisonnalité et des irréguliers. À cet égard, diverses méthodes ont été proposées; par exemple, Steel et DeMel (1988) considèrent l'effet des filtres linéaires sur le spectre de la série d'erreurs d'échantillonnage, tandis que Wolter et Monsour (1981) adoptent une méthode fondée sur l'effet de filtres linéaires sur la fonction d'autocovariance. Sutcliffe (1993) adopte une méthode semblable axée sur une approximation linéaire de la méthode X11. Pfeffermann (1994) propose de produire une estimation de l'erreur d'échantillonnage directement d'après la série chronologie estimée, en formulant diverses hypothèses de simplification. Ces méthodes ne modélisent pas explicitement la série chronologique. D'autres auteurs, comme Be0ll et Wilcox (1993), Tiller (1992), Burridge et Wallis (1985) et Hausman et Watson (1985), étudient l'application de modèles ARIMA explicites aussi bien à la série réelle qu'à la série d'erreurs d'échantillonnage, et se concentrent sur l'estimation des paramètres des modèles. Ces travaux ne tiennent pas compte de l'effet des divers plans de renouvellement et visent essentiellement à produire des estimations de la variance des estimations désaisonnalisées pour le plan de renouvellement particulier utilisé.

Le plan de renouvellement de l'échantillon choisi pour l'enquête influence la structure d'autocorrélation de la série d'erreurs d'échantillonnage, donc la variance d'échantillonnage des estimations désaisonnalisées et des estimations originales de la tendance. Le choix du plan de renouvellement dépend de plusieurs facteurs. Un chevauchement important entre les échantillons de deux périodes consécutives réduit la variance d'échantillonnage des estimations de la variation d'une période à l'autre, tandis qu'un chevauchement important entre des périodes séparées par un intervalle de 12 mois réduit la variance d'échantillonnage des estimations de la variation annuelle. Habituellement, la première inclusion d'une unité sélectionnée dans l'enquête est celle qui est la plus coûteuse. Le fait de retenir les unités sélectionnées dans l'échantillon pendant une longue période réduit le coût de l'enquête. On établit donc des plans de renouvellement en vertu desquels une unité sélectionnée est incluse pendant une série aussi longue que possible de périodes consécutives. Cependant, toute unité sélectionnée doit finalement être éliminée de l'échantillon. Outre la nécessité de répartir le fardeau de réponse pour des raisons d'éthique, le fait de retenir une unité dans l'échantillon pendant un grand nombre de périodes d'observation risque de faire baisser le taux de réponse et la qualité des donnée recueillies (pour une discussion de ces problèmes, voir Kalton et Citro 1993).

Les plans de renouvellement varient en ce qui a trait au nombre de fois qu'une unité est incluse dans l'échantillon de l'enquête et à l'intervalle entre les inclusions. Nous nous concentrons ici sur les enquêtes mensuelles sur la population active (EMPA). En pratique, les plans de renouvellement adoptés sont des cas spéciaux du plan de renouvellement *a-b-a(m)* selon lequel les unités sélectionnées sont incluses pendant a mois consécutifs, retirées de l'échantillon pendant b mois, puis réintroduites dans l'échantillon pour une période supplémentaire de a mois. Le plan est répété de sorte que les unités sélectionnées soient incluses, en tout, pour m occasions. Rao et Graham (1964) considèrent l'estimation des moyennes et des totaux de population finies pour sept classes de plans de renouvellement. Aux États-Unis, la *Current Population Survey (CPS)* est exécutée selon un plan 4-8-4(8) (Fuller, Adam et Yansaneh 1992). Si l'on pose $b = 0$, on obtient un plan de renouvellement *dans-pour-m* selon lequel les logements sélectionnés sont inclus pendant m mois après quoi ils sont éliminés de l'échantillon. Le cas où $m = 6$ correspond au

plan de renouvellement utilisé au Canada (Singh, Drew, Gambino et Mayda 1990) et celui où $m = 8$, à celui appliqué en Australie (ABS 1992). Steel (1997) a observé que l'enquête trimestrielle britannique sur la population active correspond à peu près à une enquête mensuelle avec plan de renouvellement 1-2-1(5).

Nous considérons ici la variance d'échantillonnage des estimations désaisonnalisées et des estimations de la tendance associée aux plans de renouvellement utilisés à l'heure actuelle pour les EMPA et à un certain nombre de plans de renouvellement qui, même s'ils ne sont pas utilisés à l'heure actuelle, pourraient présenter certaines propriétés intéressantes. Cette étude donnera une idée des plans de renouvellement qui donnent les meilleurs résultats en ce qui concerne la composante de la variabilité des séries estimées sur laquelle le plan d'échantillonnage exerce une influence.

3. VARIANCE D'ÉCHANTILLONNAGE DES ESTIMATIONS DÉSAISONNALISÉES ET DES ESTIMATIONS DE LA TENDANCE

Représentons par y_T le vecteur qui contient les valeurs de la série chronologique des estimations d'enquête jusqu'au temps T et par Y_T, le vecteur qui contient les valeurs réelles de population. La variance d'échantillonnage de la série originale est représentée par $V(y_T | Y_T)$. Considérons un filtre linéaire utilisé pour obtenir des valeurs à partir de y_T par application d'un vecteur de poids de filtrage w_t. Les poids de filtrage ne sont pas aléatoires et ne présentent aucun lien avec les poids de sondage utilisés pour calculer les estimations d'enquête y_t. Le vecteur de poids de filtrage w_t dépend de la période de référence à laquelle se rapportent les valeurs filtrées. Les poids sont constants dans le corps de la série, mais peuvent être modifiés au début et à la fin de celle-ci. La valeur filtrée au temps t est donnée par

$$\tilde{y}_t = w_t' y_T. \qquad (1)$$

Alors

$$V(\tilde{y}_t | Y_T) = w_t' V(y_T | Y_T) w_t \qquad (2)$$

représente la variance d'échantillonnage de la valeur filtrée au temps t. L'erreur d'échantillonnage de la valeur filtrée est égale à la différence entre $w_t' y_T$ et $w_t' Y_T$, qui est subordonnée aux valeurs de la série réelle, Y_T. Il s'agit de la différence entre la valeur filtrée obtenue d'après la série d'estimations qui se termine au temps T et la valeur que l'on obtiendrait si l'on observait cette série sans erreur d'échantillonnage. Nous nous concentrons sur cette composante, puisque c'est sur la variance d'échantillonnage que peut influer la modification du plan d'échantillonnage. On ne tient pas compte de la variance associée à Y_T. Wolter et Monsour (1981) ont examiné la question de la variance totale par opposition à la variance de l'erreur d'échantillonnage. Tenir compte de la variance totale pour interpréter la série résultante pourrait présenter certains avantages, mais puisque nous nous penchons sur des problèmes d'élaboration du plan d'échantillonnage, comme le choix du plan de renouvellement, nous nous concentrons sur la composante sur laquelle la décision relative au plan d'échantillonnage a un effet direct. Les décisions concernant le plan d'échantillonnage ne sont pas aussi importantes si l'erreur d'échantillonnage ne contribue pas de façon significative à la variabilité de la série que si elle est un facteur important de variabilité, mais il paraît logique d'utiliser un plan d'échantillonnage aussi efficace que possible.

Pour déterminer l'effet de divers plans de renouvellement sur la variance d'échantillonnage d'une série filtrée particulière, nous devons estimer $V(y_T | Y_T)$ pour divers plans de renouvellement. Les auteurs de travaux antérieurs sur l'estimation de la variance des séries désaisonnalisées ont soit omis de tenir compte du plan de renouvellement et supposé que les échantillons étaient indépendants à tout point dans le temps, soit supposé qu'ils étaient constants et utilisé une méthode d'estimation pour en tenir compte. Nous devons définir, pour $V(y_T | Y_T)$, un modèle qui reflète l'effet des divers plans de renouvellement que l'on pourrait utiliser.

L'analyse de l'effet de divers plans de renouvellement se simplifie si la structure d'autocorrélation de la série d'erreurs d'échantillonnage est stable. La forme précise de la fonction d'autocorrélation dépend de la série et doit refléter les complexités du plan d'échantillonnage. Par exemple, Steel et DeMel (1988) proposent un modèle pour les données de l'Enquête mensuelle australienne sur la population active, tandis que Bell et Wilcox (1993) en proposent un pour la série de données sur le commerce de détail aux États-Unis. Bell et Hillmer (1990), ainsi que Miazaki et Dorea (1993) considèrent aussi de modéliser les erreurs d'enquête au moyen de modèles de série chronologique. Dempster et Hwang (1993) et Lee (1990) étudient diverses méthodes en vue d'estimer et de modéliser les corrélations des erreurs d'échantillonnage dans le cas de la CPS américaine.

Nous posons que la variance de la série d'erreurs d'échantillonnage, e_t est constante. Nous devons préciser le modèle de la corrélation entre les erreurs d'échantillonnage de y_t et y_{t+s}. Tous les plans de renouvellement considérés sous-entendent que, à tout point dans le temps, l'échantillon comprendra plusieurs panels. Un panel est un ensemble d'unités qui sont intégrées dans l'enquête et qui en sont éliminées en même temps. Tout panel retiré de l'échantillon est remplacé par un autre. L'ensemble de panels reliés de cette façon forme un groupe de renouvellement. La plupart des EMPA s'appuient sur l'échantillonnage à plusieurs degrés et tout panel qui est éliminé de l'enquête est remplacé par un autre panel de ménages voisins (voir ABS 1992; Singh et coll. 1990). Par conséquent, nous supposons que la corrélation d'échantillonnage entre les estimations obtenues auprès d'un même groupe de renouvellement à s périodes d'intervalle est $r(s)$ si aucun

renouvellement n'a eu lieu et $d(s)$ s'il y a eu renouvellement. Nous supposons aussi que l'estimation au temps t est égale, du moins approximativement, à la moyenne des estimations pour chaque groupe de renouvellement et que les estimations obtenues pour divers groupes de renouvellement, qui correspondent habituellement à des UPE différentes et bien séparées spatialement, sont indépendantes.

Ces hypothèses sous-entendent que la corrélation d'échantillonnage entre y_t et y_{t+s} est

$$R(s) = d(s) + k(s)(r(s) - d(s)) \qquad (3)$$

où $k(s)$ représente la proportion de l'échantillon en commun entre les deux périodes de référence. Le plan de renouvellement détermine le facteur $k(s)$ de chevauchement des échantillons. Par exemple, pour un plan de renouvellement *dans-pour-m*, $k(s) = 1 - s/m$, $s = 0, ..., m - 1$ et est nul autrement, si l'on suppose que le même nombre de logements sont ajoutés et supprimés de l'échantillon chaque mois. Si les divers panels qui font partie d'un groupe particulier de renouvellement sont indépendants, alors $d(s) = 0$, mais, en général, cela n'est pas le cas. Ce modèle est essentiellement le même que celui dérivé par Scott, Smith et Jones (1977). Le tableau 1 donne un exemple de plan de renouvellement *dans-pour*-4 appliqué sur une période de huit mois. Les divers panels sont représentés par des lettres différentes et l'indice indique le nombre de fois que le panel a été inclus dans l'échantillon d'enquête jusqu'à la période de référence indiquée.

Tableau 1
Structure du plan de renouvellement *dans-pour*-4

Groupe de renouvellement	Période de référence							
	t	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$	$t+6$	$t+7$
1	a_1	a_2	a_3	a_4	b_1	b_2	b_3	b_4
2	c_4	d_1	d_2	d_3	d_4	e_1	e_2	e_3
3	f_3	f_4	g_1	g_2	g_3	g_4	h_1	h_2
4	i_2	i_3	i_4	j_1	j_2	j_3	j_4	k_1

Ici, $r(2)$ est la corrélation provenant de, disons, a_2 et a_4, tandis que $d(2)$ est la corrélation associée à a_3 et b_1. Binder et Hidiroglou (1988), ainsi que Fuller et coll. (1992) discutent de la structure de données que suggèrent d'autres plans de renouvellement.

L'hypothèse selon laquelle la variance de la série d'erreurs d'échantillonnage est constante sous-entend qu'aucune modification importante du plan d'échantillonnage ni de la structure de la population n'a lieu, du moins sur la longueur effective des filtres. Envisager l'hypothèse d'autocorrélations stables, $r(s)$ et $d(s)$, pour la corrélation au niveau de la population sous-entend aussi que ni le plan d'échantillonnage ni la population ne varie beaucoup. Les estimations de $r(s)$ et $d(s)$ dans (3) sont tirées d'une étude réalisée par Bell (1998). Les valeurs utilisées, c'est-à-dire les proportions de personnes occupées et de chômeurs tirées de l'enquête australienne sur la population active (EAPA), sont présentées au tableau 2. Pour les obtenir, on a traité les groupes de renouvellement de l'EAPA comme des répétitions et on a déterminé l'autocorrélation au niveau du groupe de renouvellement. On s'est servi d'un modèle proposé par Bell (1998) pour extrapoler les valeurs au-delà des décalages donnés.

Tableau 2
Autocorrélations – EAPA

Proportion de personnes occupées								
décalage	1	2	3	4	5	6	7	8
$r(s)$	0,80	0,71	0,64	0,57	0,50	0,45	0,40	0,36
$d(s)$	0,15	0,15	0,14	0,13	0,12	0,11	0,11	0,10
Proportion de chômeurs								
	1	2	3	4	5	6	7	8
$r(s)$	0,62	0,52	0,44	0,37	0,31	0,26	0,22	0,19
$d(s)$	0,11	0,11	0,10	0,09	0,09	0,08	0,08	0,07

Sutcliffe et Lee (1995) étudient les erreurs types des estimations désaisonnalisées et des estimations de la tendance des niveaux et des flux pour un petit nombre de plans de renouvellement distincts. Ils supposent pour cela que les corrélations entre les estimations d'enquête obéissent à un modèle de décroissance géométrique simple, avec, pour la population, une corrélation $\rho = 0.8$, c'est-à-dire $R(s) = \rho^s$, qui diminue plus rapidement que les valeurs données au tableau 2.

4. APPROXIMATIONS LINÉAIRES DES ESTIMATIONS DÉSAISONNALISÉES ET DES ESTIMATIONS DE LA TENDANCE

La méthode X11 consiste à appliquer itérativement des moyennes mobiles qui produisent un filtre symétrique pour les valeurs centrales et des filtres asymétriques pour les valeurs de début et de fin de série. Le lissage au moyen de filtres linéaires donne une approximation des estimations finales désaisonnalisées et des estimations finales de la tendance produites par la méthode X11. Plusieurs auteurs, dont Young (1968), Cleveland et Tiao (1976), Wallis (1982) et Sutcliffe (1993), ont produit des approximations linéaires de la méthode X11. La méthode X11ARIMA (Dagum 1980, 1988), quant à elle, est une extension de la méthode X11 qui consiste à extrapoler la série originale aux deux extrémités par application d'un modèle autorégressif à moyennes mobiles intégré (ARIMA pour *Autoregressive Integrated Moving Average Model*). On peut intégrer l'effet de l'extrapolation ARIMA à la pondération des filtres et puis appliquer les poids ainsi obtenus aux données uniquement. Dagum, Chhab et Chiu (1996) ont proposé une méthode en cascade, où les filtres en cascade résultent de la convolution de divers filtres linéaires prédéterminés utilisés dans les méthodes X11 et X11ARIMA. Nous utilisons leur

approche pour obtenir une approximation réaliste de ces deux dernières méthodes.

Représentons par $\mathbf{H}_{13}$ la matrice dont les lignes contiennent les poids de filtrage des moyennes mobiles de Henderson sur 13 termes pour les filtres symétriques ainsi qu'asymétriques. Représentons par $\mathbf{S}_{3\times3}$ la matrice des poids qui correspondent à la moyenne mobile (mm) 3×3 et par $\mathbf{S}_{3\times5}$, celle qui correspond à la mm 3×5. Nous utilisons ces matrices pour estimer les facteurs de saisonnalité. Enfin, nous représentons par $\mathbf{D}$ une mm centrée sur 12 termes et par $\mathbf{I}$, une matrice d'identité. La notation c indique le complément d'un filtre, par exemple $\mathbf{D}^c = \mathbf{I} - \mathbf{D}$. Les filtres de désaisonnalisation en cascade sont représentés par

$$\mathbf{S} = \mathbf{I} - \mathbf{D}^c\mathbf{S}_{3\times5}[\mathbf{H}_{13}(\mathbf{D}^c\mathbf{S}_{3\times3}\mathbf{D}^c)^c]^c.$$

Nous obtenons alors le filtre en cascade pour estimer la tendance en multipliant le filtre désaisonnalisé par un filtre de la tendance. Aux extrémités de la série, les filtres en cascade appliqués aux estimations de la tendance et aux estimations désaisonnalisées diffèrent selon qu'on applique la méthode X11 ou X11ARIMA.

Nous considérons les combinaisons des filtres internes de X11 et de X11ARIMA qui suivent:

1. Filtre en cascade X11 type: Ce filtre comprend une MMH sur 13 termes pour l'estimation de la tendance (H13), une mm 3×3 pour la première estimation des facteurs saisonniers ($S1_{3\times3}$), une mm 3×5 pour l'estimation des facteurs saisonniers ($S2_{3\times5}$), mais ne comporte aucune modification pour les valeurs aberrantes.
2. Filtre en cascade X11 type avec prévisions ARIMA: Ce filtre comprend les filtres H13, $S1_{3\times3}$ et $S2_{3\times5}$, ainsi que des prévisions extrapolées d'après un modèle ARIMA de la forme $(1-B)(1-B^{12})y_t = (1-0.4B)(1-0.6B^{12})a_t$, où B est l'opérateur de rétropolation et a_t, un facteur de traitement de l'erreur résiduelle (bruit blanc), mais ne comporte aucune modification pour les valeurs aberrantes.
3. Filtre en cascade X11 court avec prévisions ARIMA: Ce filtre comprend les filtres H9, $S1_{3\times3}$ et $S2_{3\times5}$, ainsi que des prévisions extrapolées d'après un modèle de la forme $(1-B)(1-B^{12})y_t = (1-0.3B)(1-0.3B^{12})a_t$, mais ne comporte aucune modification pour les valeurs aberrantes.
4. Filtre en cascade X11 long avec prévisions ARIMA: Ce filtre comprend les filtres H23, $S1_{3\times3}$ et $S2_{3\times5}$, ainsi que des prévisions extrapolées d'après un modèle de la forme $(1-B)(1-B^{12})y_t = (1-0.8B)(1-0.8B^{12})a_t$, mais ne comporte aucune modification pour les valeurs aberrantes.

Selon Dagum (1983), les combinaisons 2 et 3 sont applicables à plusieurs cas. Les approximations linéaires choisies nous permettent d'examiner l'effet de divers plans de renouvellement pour une gamme de filtres axés sur des MMH de diverses longueurs que l'on utilise en pratique.

Pour chaque combinaison de filtres, le filtre en cascade correspondant fournit un vecteur de poids de filtrage pour l' la désaisonnalisation et un vecteur de poids distinct pour l'estimation finale de la tendance. Ces vecteurs peuvent être introduits dans l'équation (2) pour obtenir la variance d'échantillonnage pour un plan de renouvellement particulier grâce à l'utilisation des valeurs appropriées de $V(y_T|\boldsymbol{Y}_T)$. Si l'on estime les variations, le vecteur de données y_T reste le même, mais les poids appliqués changent. Par exemple, on peut utiliser $w_{t+1} - w_t$ pour un écart d'un mois. Cette méthode de base est la même que celle adoptée par Wolter et Monsour (1981) qui ont proposé d'estimer la variance des estimations désaisonnalisées en se servant de l'équation (2) avec des poids choisis de façon à obtenir une approximation raisonnable de la méthode de désaisonnalisation et d'utiliser une estimation de $V(y_T|\boldsymbol{Y}_T)$ fondée sur des données d'enquête. Nous considérons aussi les filtres de la tendance et différentes réalisations de X11ARIMA et des plans de renouvellement.

Les modèles X11ARIMA étudiés ici sont représentatifs de ceux utilisés couramment en pratique. L'utilisation de prévisions ARIMA, dans le cadre de la méthode X11ARIMA, complique la situation. Par exemple, on suppose que les spécifications du modèle ARIMA ne comportent aucune erreur. Habituellement, on identifie et on estime le modèle ARIMA d'après des données d'enquête antérieures. Or, l'erreur d'échantillonnage commise lors de périodes antérieures risque d'influencer le choix du modèle ARIMA et des filtres X11. On pourrait tenir compte de cette situation en modifiant le terme de variance dans (2).

Nous calculons les premières estimations désaisonnalisées et de la tendance pour le temps t en nous servant de la série chronologique d'estimations qui se termine au temps t, c'est-à-dire y_t, ce qui donne la valeur filtrée $w_t'y_t$. La valeur que nous obtiendrions si aucune erreur d'échantillonnage n'était commise est $w_t'\boldsymbol{Y}_t$. L'erreur d'échantillonnage considérée ici est $w_t'y_t - w_t'\boldsymbol{Y}_t$. Comme nous ajoutons des estimations à la série, la valeur filtrée au temps t pourrait changer, mais à partir d'un certain point, $t+s$, nous ne noterons plus aucune variation appréciable. La valeur filtrée finale au temps t basée sur les estimations d'enquête peut s'écrire $w_t^{*\prime}y_{t+s}$, pour un vecteur final symétrique de poids w_t^*. Pareillement, la valeur finale que nous obtiendrions si aucune erreur d'échantillonnage n'était commise serait $w_t^{*\prime}\boldsymbol{Y}_{t+s}$. Bell et Kramer (1999) considèrent la différence $w_t'y_t - w_t^{*\prime}\boldsymbol{Y}_{t+s}$, qui inclut l'erreur de prévision. Nous pouvons décomposer cette différence comme suit:

$$w_t'y_t - w_t^{*\prime}\boldsymbol{Y}_{t+s} = (w_t'y_t - w_t'\boldsymbol{Y}_t) + (w_t'\boldsymbol{Y}_t - w_t^{*\prime}\boldsymbol{Y}_{t+s}).$$

Nous avons étudié la façon dont divers plans de renouvellement influent sur le premier terme de cette décomposition. Le deuxième terme, qui correspond à la série

observée sans erreur d'échantillonnage, ne dépend pas du plan d'échantillonnage, y compris le plan de renouvellement. Bell et Kramer (1999) ont étudié la série sur les mises en chantier aux États-Unis comportant au moins cinq unités et montrent que la variance totale de la série augmente fortement aux extrémités à cause d'erreurs de prévision. Cette situation est due aux corrections des estimations initiales de la tendance faites à mesure que l'on ajoute des estimations à la série. Steel et McLaren (2000) étudient l'effet de divers plans de renouvellement sur la correction observée des estimations initiales de la tendance, que l'on peut représenter par $w_t' y_t - w_t^{*\prime} y_{t+s}$. Ils constatent que l'importance relative de la composante due à l'erreur d'échantillonnage dépend de la façon dont la série réelle évolue autour de la période observée.

5. RÉSULTATS

Nous utilisons les filtres qui correspondent au niveau, d'une part, et à la variation mensuelle, d'autre part, pour les estimations désaisonnalisées ainsi que les estimations de la tendance à la toute extrémité de la série. Les tableaux 3 à 6 résument les effets de divers plans de renouvellement pour chaque combinaison de filtres en cascade. Ces tableaux donnent, pour certains plans de renouvellement, le rapport de la variance d'échantillonnage des estimations étudiées à la variance d'échantillonnage que l'on obtiendrait si l'on renouvelait complètement l'échantillon chaque mois. Les rapports calculés pour le milieu de la série mènent à des conclusions générales semblables (McLaren 1999).

Tableau 3
Rapport de la variance d'échantillonnage pour certains plans de renouvellement à la variance d'échantillonnage pour un plan d'échantillonnage indépendant
(Combinaison 1)

Plan de renouvellement	$\hat{SA}_t$		$\hat{SA}_{t+1} - \hat{SA}_t$		$\hat{T}_t$		$\hat{T}_{t+1} - \hat{T}_t$	
	occ.	chôm.	occ.	chôm.	occ.	chôm.	occ.	chôm.
complet	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
1-2-1(5)	0,99	0,99	0,99	1,00	0,99	1,00	0,68	0,79
1-2-1(8)	0,98	0,99	0,97	0,99	0,98	0,99	0,64	0,77
1-1-1(6)	1,01	1,01	1,00	1,00	1,17	1,14	0,70	0,82
2-2-2(8)	1,02	1,02	0,61	0,71	1,26	1,23	0,83	0,95
2-10-2(4)	1,04	1,04	0,61	0,71	1,35	1,30	1,32	1,26
3-3-3(6)	1,07	1,06	0,48	0,61	1,52	1,44	1,29	1,25
4-8-4(8)	1,10	1,08	0,42	0,57	1,69	1,57	1,42	1,34
6-6-6(12)	1,10	1,08	0,36	0,52	1,76	1,64	1,22	1,22
dans-pour-6	1,10	1,08	0,36	0,52	1,76	1,64	1,22	1,22
dans-pour-8	1,09	1,08	0,33	0,50	1,78	1,65	1,06	1,13
pas de renouvellement	1,08	1,08	0,24	0,44	1,80	1,69	0,75	0,95

Tableau 4
Rapport de la variance d'échantillonnage pour certains plans de renouvellement à la variance d'échantillonnage pour un plan d'échantillonnage indépendant
(Combinaison 2)

Plan de renouvellement	$\hat{SA}_t$		$\hat{SA}_{t+1} - \hat{SA}_t$		$\hat{T}_t$		$\hat{T}_{t+1} - \hat{T}_t$	
	occ.	chôm.	occ.	chôm.	occ.	chôm.	occ.	chôm.
complet	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
1-2-1(5)	1,01	1,01	0,99	1,00	1,06	1,05	0,69	0,80
1-2-1(8)	1,00	1,00	0,96	0,99	1,07	1,05	0,66	0,78
1-1-1(6)	1,04	1,03	1,00	1,00	1,22	1,17	0,65	0,77
2-2-2(8)	1,05	1,04	0,60	0,71	1,32	1,26	0,81	0,92
2-10-2(4)	1,02	1,03	0,60	0,71	1,26	1,23	1,19	1,17
3-3-3(6)	1,08	1,06	0,49	0,61	1,49	1,40	1,19	1,16
4-8-4(8)	1,06	1,06	0,41	0,56	1,56	1,47	1,13	1,13
6-6-6(12)	1,08	1,07	0,35	0,52	1,67	1,56	0,93	1,01
dans-pour-6	1,10	1,08	0,36	0,52	1,69	1,56	0,94	1,01
dans-pour-8	1,11	1,08	0,32	0,49	1,75	1,61	0,82	0,93
pas de renouvellement	1,14	1,11	0,24	0,43	1,89	1,73	0,59	0,78

Tableau 5
Rapport de la variance d'échantillonnage pour certains plans de renouvellement à la variance d'échantillonnage pour un plan d'échantillonnage indépendant
(Combinaison 3)

Plan de renouvel lement	$\hat{SA}_t$		$\hat{SA}_{t+1} - \hat{SA}_t$		$\hat{T}_t$		$\hat{T}_{t+1} - \hat{T}_t$	
	occ.	chôm.	occ.	chôm.	occ.	chôm.	occ.	chôm.
complet	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
1-2-1(5)	0,99	0,99	0,96	0,98	0,99	0,99	0,68	0,79
1-2-1(8)	0,97	0,99	0,93	0,97	0,98	0,99	0,64	0,77
1-1-1(6)	1,04	1,02	0,99	0,99	1,11	1,08	0,60	0,72
2-2-2(8)	1,07	1,06	0,60	0,71	1,23	1,19	0,89	0,95
2-10-2(4)	1,05	1,06	0,61	0,72	1,21	1,20	1,07	1,08
3-3-3(6)	1,15	1,12	0,51	0,63	1,41	1,32	1,02	1,02
4-8-4(8)	1,12	1,11	0,44	0,58	1,41	1,35	0,85	0,93
6-6-6(12)	1,14	1,13	0,37	0,53	1,47	1,39	0,69	0,82
dans-pour-6	1,16	1,13	0,38	0,53	1,49	1,40	0,70	0,81
dans-pour-8	1,17	1,14	0,34	0,51	1,52	1,42	0,61	0,76
pas de renouvellement	1,22	1,17	0,25	0,44	1,62	1,50	0,44	0,64

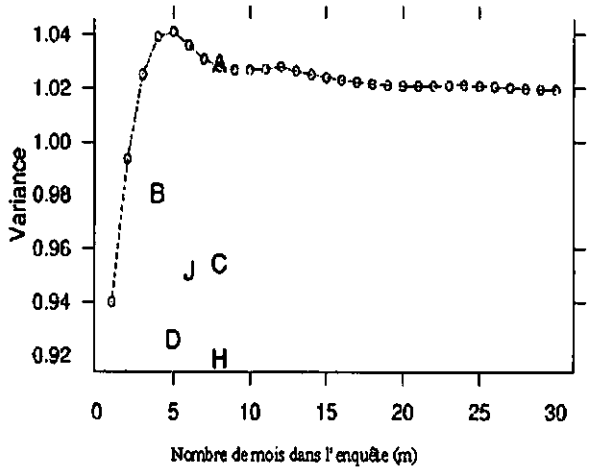
Tableau 6
Rapport de la variance d'échantillonnage pour certains plans de renouvellement à la variance d'échantillonnage pour un plan d'échantillonnage indépendant
(Combinaison 4)

Plan de renouvellement	$\hat{SA}_t$		$\hat{SA}_{t+1} - \hat{SA}_t$		$\hat{T}_t$		$\hat{T}_{t+1} - \hat{T}_t$	
	occ.	chôm	occ.	chôm	occ.	chôm	occ.	chôm.
complet	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
1-2-1(5)	1,02	1,02	0,99	1,00	1,25	1,19	0,75	0,87
1-2-1(8)	1,02	1,02	0,97	0,99	1,28	1,21	0,70	0,84
1-1-1(6)	1,06	1,04	1,00	1,00	1,49	1,39	0,92	1,01
2-2-2(8)	1,06	1,04	0,60	0,71	1,57	1,47	0,98	1,09
2-10-2(4)	1,00	1,01	0,60	0,70	1,30	1,27	1,49	1,37
3-3-3(6)	1,07	1,05	0,48	0,61	1,64	1,54	1,34	1,36
4-8-4(8)	1,05	1,04	0,41	0,56	1,73	1,63	1,92	1,69
6-6-6(12)	1,08	1,06	0,35	0,51	2,00	1,84	1,87	1,68
dans-pour-6	1,09	1,07	0,35	0,52	2,00	1,84	1,90	1,70
dans-pour-8	1,11	1,08	0,32	0,49	2,15	1,96	1,73	1,62
pas de renouvellement	1,17	1,12	0,24	0,43	2,56	2,27	1,11	1,33

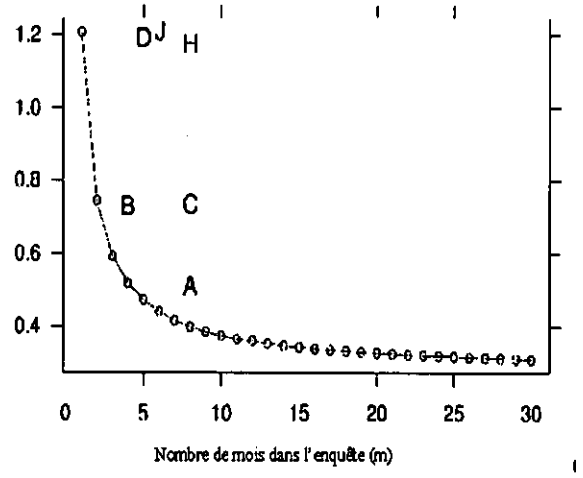
5.1 X11 – Filtres en cascade types concourants

Les résultats obtenus en utilisant les filtres types X11 (combinaison 1) sont présentés au tableau 3. Les figures 1(a) à 1(d) montrent la variance d'échantillonnage du niveau et de la variation mensuelle des estimations désaisonnalisées et des estimations de la tendance à la fin de la série divisée par la variance de l'estimation originale du niveau en fonction du nombre total de fois qu'une unité sélectionnée est incluse. Les résultats pour la variable employée sont portés en graphique pour certains plans *a-b-a(m)* et pour le plan de renouvellement *dans-pour-m* où *m* varie de 1 à 30. Un plan de renouvellement *dans-pour*-30 signifie qu'il n'y a pas de renouvellement.

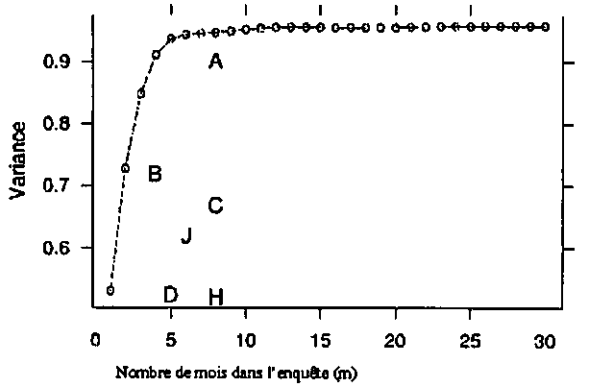
Les colonnes 1 et 2 du tableau 3 montrent que les plans de renouvellement sans chevauchement mensuel donnent de bons résultats en ce qui concerne la variance des estimations désaisonnalisées du niveau. L'utilisation de plans de renouvellement avec chevauchement annuel ne produit aucune amélioration appréciable. Cependant, dans le cas des estimations désaisonnalisées de la variation mensuelle, l'avantage d'un chevauchement mensuel important est manifeste (voir la figure 1(b), ainsi que les colonnes 3 et 4 du tableau 3). La variance associée aux plans de renouvellement *dans-pour-m* est, en fait, une fonction de $1/m$, c'est-à-dire la proportion de l'échantillon pour laquelle il n'y a pas de chevauchement. Au Canada et en Australie, l'utilisation de ces plans de renouvellement donne de bons résultats. La meilleure solution est celle où il n'y a pas de renouvellement, mais, comme on l'a mentionné à la section 2, on ne peut l'adopter en pratique. Les figures 1(a) et 1(b) montrent que, pour la série désaisonnalisée, la variance des estimations du niveau et de la variation mensuelle est la même pour des plans de renouvellement qui présentent le même degré de chevauchement mensuel des échantillons.



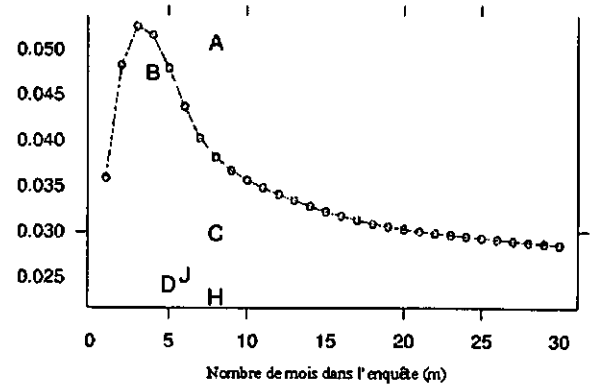
1a) Désaisonnalisation: estimations du niveau



1b) Désaisonnalisation: variations mensuelles



1c) Tendance: estimations du niveau



1d) Tendance: variations mensuelles

Figure 1. Rapport de la variance d'échantillonnage à la variance de la série originale pour le plan de renouvellement choisi pour la combinaison 1 (X11) pour la variable personne occupée où A = 4-8-4(8), B = 2-10-2(4), C = 2-2-2(8), D = 1-2-1(5), H = 1-2-1(8), J = 1-1-1(6).

En ce qui concerne les estimations du niveau de la tendance, la variance augmente parallèlement au chevauchement des échantillons mensuels (voir la figure 1(c), ainsi que les colonnes 5 et 6 du tableau 3). Dans le cas des plans de renouvellement *dans-pour-m*, la variance augmente rapidement quand m passe de 1 à 5. Les plans de renouvellement 1-2-1(5) et 1-2-1(8) donnent d'aussi bons résultats que si l'on sélectionnait un échantillon indépendant chaque mois et de nettement meilleurs résultats que les plans de renouvellement avec chevauchement mensuel. Cette situation tient surtout au fait que, pour une moyenne mobile, il est préférable de calculer la moyenne d'observations indépendantes que celle d'observations positivement corrélées. La variance plus grande du plan 1-1-1(6) que des plans 1-2-1(5) et 1-2-1(8) donne à penser que, dans le cas de plans de renouvellement sans chevauchement mensuel, l'intervalle entre deux inclusions successives des unités dans l'échantillon exerce un certain effet.

La figure 1(d), ainsi que les colonnes 7 et 8 du tableau 3 montrent que, pour les estimations de la variation mensuelle de la tendance, la variance augmente très rapidement quand m passe de 1 à 3 et diminue très rapidement quand m augmente à partir de 4. De tous les plans de renouvellement considérés, le plan *dans-pour*-3 semble être le moins bon et le plan utilisé à l'heure actuelle pourrait donc être amélioré considérablement. Par exemple, le choix d'un plan 1-2-1(8) au lieu de 4-8-4(8) réduirait la variance des estimations de la variation mensuelle de la tendance de 55% pour l'emploi et de 43% pour le chômage. Bien que le degré de chevauchement mensuel demeure un facteur essentiel, le plan d'inclusion joue un rôle également. Par exemple, le plan de 2-2-2(8) produit une variance plus faible que le plan *dans-pour*-2 ou 2-10-2(4). Qui plus est, pour les estimations de la variation mensuelle de la tendance, les plans de renouvellement les meilleurs sont les plans 1-2-1(5) et 1-2-1(8), qui donnent des résultats nettement supérieurs à ceux obtenus en procédant à un renouvellement complet chaque mois. Ce résultat tient au fait que les estimations de la variation mensuelle de la tendance représentent, en réalité, l'écart entre des séries désaisonnalisées décalées de quelques mois et que les plans de renouvellement 1-2-1(m) produisent des corrélations positives entre les estimations espacées de trois mois. McLaren et Steel (1997) ont obtenu des résultats comparables en se servant de l'approximation de X11 proposée par Sutcliffe (1993).

Les résultats montrent que, pour les estimations du niveau et du dernier mouvement de la tendance, les plans de renouvellement 1-2-1(m) produisent une variance d'échantillonnage nettement plus faible que les plans de renouvellement utilisé à l'heure actuelle.

5.2 X11ARIMA – Filtres en cascade concourants avec extrapolation

Les résultats de l'application des combinaisons de filtres 2, 3 et 4 sont présentés aux tableaux 4, 5 et 6, respectivement. Les figures 2(a) à 2(d) présentent les résultats pour la combinaison 4 dans le cas de la variable "personne occupée".

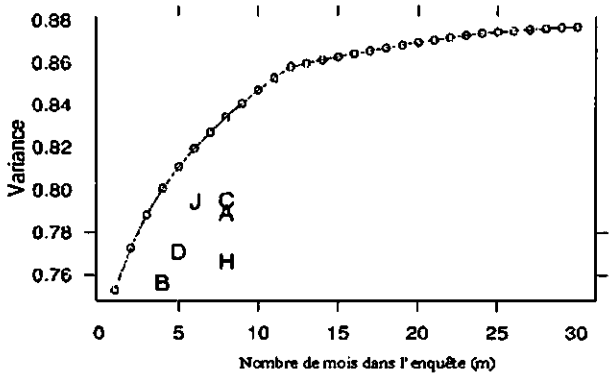
Les colonnes 1 et 2 des tableaux 4, 5 et 6 montrent que les plans de renouvellement avec faible chevauchement mensuel donnent des résultats presque aussi bons que les plans à renouvellement complet pour les estimations désaisonnalisées du niveau. La variance est plus forte pour les plans de renouvellement avec chevauchement mensuel important, particulièrement pour la combinaison 3 qui correspond à l'utilisation de la MMH sur 9 termes.

L'écart entre les rapports calculés pour les quatre combinaisons est minime dans le cas des estimations désaisonnalisées de la variation mensuelle (colonnes 3 et 4 dans tous les tableaux). Les plans de renouvellement avec chevauchement mensuel important demeurent supérieurs aux plans avec chevauchement mensuel faible ou nul, quelle que soit la combinaison X11/X11ARIMA utilisée.

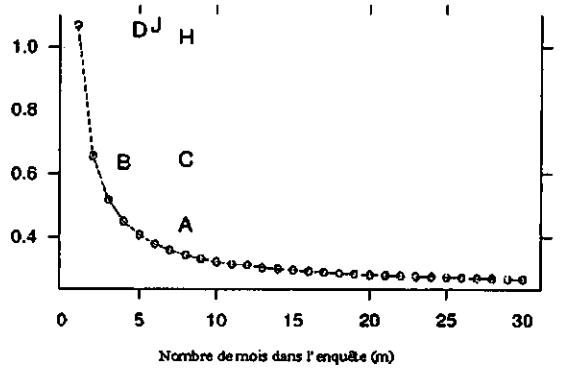
En ce qui concerne les estimations du niveau de la tendance, les plans de renouvellement avec chevauchement important de l'échantillon produisent de nouveau un rapport plus grand des variances. Les plans de renouvellement 1-2-1(5) et 1-2-1(8) continuent de donner de meilleurs résultats que les autres pour chaque combinaison de filtres, mais donnent de moins bons résultats qu'un échantillon indépendant pour les combinaisons 2 et 4.

Dans le cas des estimations de la variation mensuelle de la tendance, les meilleurs plans de renouvellement sont de nouveau les plans 1-2-1(5) et 1-2-1(8), qui donnent des résultats supérieurs à ceux de l'échantillon indépendant pour les quatre combinaisons de filtres. Pour la combinaison 3, les plans de renouvellement avec chevauchement mensuel important sont aussi bons que les plans de renouvellement 1-2-1(m). Pour les combinaisons 2 et 3, le plan 1-1-1(6) est légèrement supérieur au plan 1-2-1(m). Le remplacement des plans de renouvellement utilisés à l'heure actuelle par les plans de renouvellement 1-2-1(m) produirait des améliorations considérables. Par exemple, pour la variable d'emploi, passer du plan 4-8-4(8) au plan a 1-2-1(8) produirait des gains de 42%, 25% et 64% en choisissant les combinaisons 2, 3 et 4, respectivement.

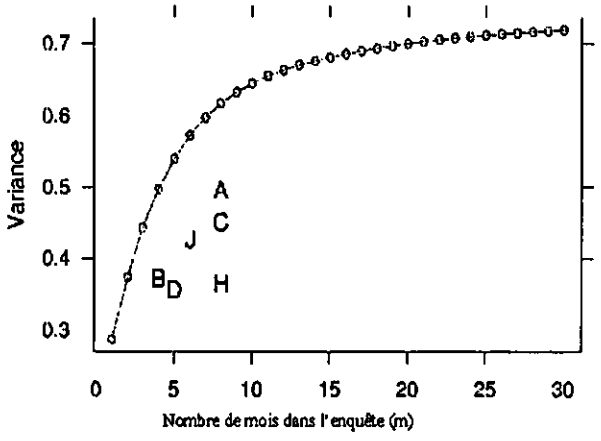
Ces résultats se fondent sur les estimations de corrélation de l'EAPA qui, puisqu'elles sont fondées sur des estimations d'enquête, sont entachées d'une erreur d'échantillonnage. Les filtres de la tendance étudiés ne sont pas calculés en se servant de ces estimations. Nous arrivons aux mêmes conclusions générales quant à l'effet de divers plans de renouvellement pour les deux modèles de corrélation basés sur des corrélations raisonnablement différentes. Selon nous, les conclusions sont applicables à la gamme de modèles de corrélation compris entre ces deux modèles extrêmes. McLaren et Steel (1997) arrivent aux mêmes conclusions en se servant du modèle de corrélation calculé par Steel (1996) pour l'emploi et le chômage au Royaume-Uni.



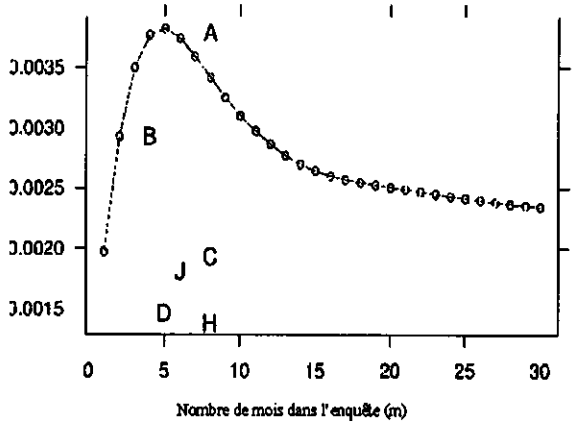
2a) Désaisonnalisation: estimations du niveau



2b) Désaisonnalisation: variations mensuelles



2c) Tendance: estimations du niveau



2d) Tendance: variations mensuelles

Figure 2. Rapport de la variance d'échantillonnage à la variance de la série originale pour le plan de renouvellement choisi pour la combinaison 4 (X11 ARIMA) pour la variable "personne occupée" où A= 4-8-4(8), B=2-10-2(4), C=2-2-2(8), D=1-2-1(5), H=1-2-1(8), J=1-1-1(6).

6. DISCUSSION

Les plans de renouvellement utilisés à l'heure actuelle, tels que *dans-pour*-8, *dans-pour*-6 et 4-8-4(8), sont raisonnables si la statistique principale sur laquelle porte l'analyse est la variation mensuelle des estimations désaisonnalisées. Selon nous, très souvent, l'examen de cette variation n'est pas un moyen fiable d'évaluer les tendances courantes. Il est nécessaire d'étudier la courbe de variation au cours des mois récents. On peut, pour cela, utiliser des filtres qui produisent une estimation de la tendance. Les résultats obtenus ici donnent à penser que, si l'utilisation des données d'enquête visent principalement à déterminer la tendance, il est préférable d'utiliser des plans de renouvellement assez différents. Plus précisément, les plans de renouvellement 1-2-1(m) donnent de bons résultats si l'on cherche à réduire la variance des estimations du niveau de la tendance et de l'écart entre deux estimations consécutives de la tendance pour une gamme de combinaisons de filtres. Les plans de renouvellement 1-2-1(m) donnent aussi de bons résultats pour la variance d'échantillonnage des estimations désaisonnalisées de niveau. Par conséquent, lors de l'élaboration du plan de renouvellement de l'échantillon d'une enquête répétée, il faut tenir compte de l'importance relative des estimations désaisonnalisées et des estimations de la tendance. L'examen des figures 1 et 2 montre que le plan de renouvellement 2-2-2(8) est un compromis raisonnable si l'on juge important d'étudier à la fois le niveau et la variation mensuelle des estimations désaisonnalisées et des estimations de la tendance. Bell (1999) considère aussi l'effet de quatre plans de renouvellement distincts sur la variance d'échantillonnage du niveau et de la variation mensuelle des estimations originales, non corrigées, et des estimations de la tendance obtenues en se servant de la méthode X11 et d'une MMH sur 13 tèrmes. Il juge aussi que le plan de renouvellement 2-2-2(8) est un bon compromis.

Même si les analystes n'utilisent pas officiellement les estimations de la tendance, l'évaluation de cette dernière demande l'examen de la variation d'estimations désaisonnalisées à intervalle de quelques mois. McLaren (1999)

montre que les plans de renouvellement 1-2-1(m) conviennent bien si l'évaluation de la tendance consiste à examiner les variations d'estimations désaisonnalisées sur trois ou sur six mois. Ces résultats donnent aussi à penser que ce genre de plan de renouvellement donne de bons résultats pour les estimations de la variation des estimations de la tendance sur les périodes de trois ou de six mois les plus récentes.

Le critère d'évaluation utilisé ici est la variance d'échantillonnage des estimations de la tendance et des estimations désaisonnalisées, facteur qui est influencé par le plan d'échantillonnage. Steel et McLaren (2000) évaluent divers plans de renouvellement en ce qui concerne l'importance des corrections de ces estimations aux extrémités de la série et tirent des conclusions comparables quant aux plans de renouvellement.

REMERCIEMENTS

Les présents travaux ont été financés par l'*Australian Research Council* et l'*Australian Bureau of Statistics (ABS)*. Les opinions exprimées dans le présent article ne représentent pas nécessairement celles de ces organismes. Nous remercions le rédacteur adjoint et les examinateurs de leurs commentaires, ainsi que Geoff Lee, Andrew Sutcliffe et Phillip Bell de l'ABS, et Norma Chhab, de Statistique Canada.

BIBLIOGRAPHIE

AUSTRALIAN BUREAU OF STATISTICS (1987). *A Guide to Smoothing Time Series – Estimates of "Trend"*. Australian Bureau of Statistics, numéro 1316.0 au catalogue, Canberra.

AUSTRALIAN BUREAU OF STATISTICS (1992). *Information Paper: Labour Force Survey Sample Design*. Australian Bureau of Statistics, numéro 6269.0 au catalogue, Canberra.

AUSTRALIAN BUREAU OF STATISTICS (1993). *A Guide to Interpreting Time Series – Monitoring "Trends", An Overview*. Australian Bureau of Statistics, numéro 1348.0 au calatogue, Canberra.

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 dans Econometrics and Applied Statistics, 98/2, Australian Bureau of Statistics, numéro 1351.0 au catalogue, Canberra.

BELL, P.A. (1999). *The Impact of Sample Rotation Patterns and Composite Estimation on Survey Outcomes*. Document de travail dans Econometrics and Applied Statistics, No. 99/1, Australian Bureau of Statistics, numéro 1352.0 au catalogue, Canberra.

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

BELL, W.R., et KRAMER, M. (1999). Vers des variances pour la désaisonnalisation X-11. *Techniques d'enquête*, 25, 13-32.

BELL, W.R., et WILCOX, D.W. (1993). The effect of sampling error on the time series behavior of consumption data. *Journal of Econometrics*, 555, 235-265.

BINDER, D.A., et HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, (P.R. Krishaiah et C.R. Rao, Éds.). Amsterdam: Elsevier Science Publishers, B.V., 6, 187-211.

BURRIDGE, P., et WALLIS, K.F. (1985). Calculation of seasonally adjusted series. *Journal of the American Statistical Association*, 80, 541-552.

CLEVELAND, W.P., et TIAO, G.C. (1976). Decomposition of seasonal time series: a model for the X-11 program. *Journal of the American Statistical Association*, 71, 581-587.

DAGUM, E.B. (1980). *The X-11ARIMA Seasonal Adjustment Method*. Numéro 12-564E au catalogue, Statistique Canada, Ottawa.

DAGUM, E.B. (1983). Spectral properties of the concurrent and forecasting seasonal linear filters of the X-11ARIMA method. *La revue canadienne de statistique*, 11, 73-90.

DAGUM, E.B. (1988). *The X-11ARIMA/88 Seasonal Adjustment Methods – Foundations and User's Manual*. Statistique Canada, Ottawa.

DAGUM, E.B., CHHAB, N., et CHIU, K. (1996). Derivation and properties of the X-11ARIMA and Census X11 linear filters. *Journal of Official Statistics*, 12, 329-347.

DEMPSTER, A.P. et HWANG, J.-S. (1993). Component models and Bayesian technology for estimation of state employment and unemployment rates. *Proceedings of the Bureau of the Census Annual Research Conference*, 571-581.

FINDLEY, D.F., MONSELL, B.C., OTTO, M.C., BELL, W.R., et PUGH, M.G. (1998). New capabilities and methods of the X-12 ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics*, 16, 127-177.

FULLER, W.A., ADAM, A., et YANSANEH, I.S. (1992). Estimateurs pour des enquêtes longitudinales avec application à l'enquête "Current Population Survey" des É.-U. *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales*, Statistique Canada, 301-324.

GRAY, A., et THOMSON, P. (1996). Design of moving-average trend filters using fidelity, smoothness and minimum revisions criteria. *Time Series Analysis in Memory of E.J. Hannan*. (P. Robinson et M. Rosenblatt, Éds.), 205-219. Springer lecture notes in statistics, 115.

HENDERSON, R. (1916). Note on graduation by adjusted averages. *Transactions of the Actuarial Society of America*, 17, 43-48.

HAUSMAN, J.A., et WATSON, M.W. (1985). Error in variables and seasonal adjustment procedures. *Journal of the American Statistical Association*, 80, 531-540.

KALTON, G., et CITRO, C.F. (1993). Enquêtes par panel: ajout d'une quatrième dimension. *Techniques d'enquête*, 19, 217-227.

KENNY, P.B., et DURBIN, J. (1982). Local trend estimation and seasonal adjustment of economic and social time series. *Journal of the Royal Statistical Society A*, 145, 1-41.

KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, 14, 31-46.

KNOWLES, J. (1997). Trend Estimation Practices of National Statistical Institutes. Office for National Statistics, Methods and Quality Division, UK, MQ 044.

KNOWLES, J., et KENNY (1997). An Investigation of Trend Estimation Methods. Office for National Statistics, Methods and Quality Division, UK, MQ 044.

LEE, H. (1990). Estimation des coefficients de corrélation de panel pour l'enquête sur la population active du Canada. *Techniques d'enquête*, 16, 297-306.

LINACRE, S., et ZARB, J. (1991). Picking turning points in the economy. *Australian Economic Indicators*. Australian Bureau of Statistics, numéro 1350.0 au catalogue.

MCLAREN, C.H. (1999). Designing Rotation Patterns and Filters for Trend Estimation in Repeated Surveys. Non publiée Thèse de doctorat, School of Mathematics and Applied Statistics, University of Wollongong.

MCLAREN, C.H., et STEEL, D.G. (1997). The effect of different rotation patterns on the sampling variance of seasonal and trend filters. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1997, 790-795.

MIAZAKI, E.S., et DOREA, C.C.Y. (1993). Estimation of the parameters of a time series subject to the error of rotation sampling. *Communications in Statistics*, A, 22, 805-825.

PFEFFERMANN, D. (1994). A general method for estimating the variances of X-11 seasonally adjusted estimators. *Journal of Time Series Analysis*, 15, 85-116.

RAO, J.N.K., et GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 69, 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, 3-73.

SHISKIN, J., YOUNG, A.H., et MUSGRAVE, J.C. (1967). *X-11 Variant of the Census Method II Seasonal Adjustment Program*. Article technique 15, Bureau of the Census, U.S. Department of Commerce, Washington, D.C.

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

SMITH, T.M.F. (1997). Discussion of paper by Steel. *Journal of the Royal Statistical Society A*, 160, 33-34.

STEEL, D.G. (1996). Options for Producing Monthly Estimates of Unemployment According to the ILO Definition. Central Statistical Office, U.K.

STEEL, D.G. (1997). Producing monthly estimates of unemployment and employment according to the international labour office definition. *Journal of the Royal Statistical Society A*, 160, 5-46.

STEEL, D.G., et MCLAREN, C.H. (2000). The effect of different rotation patterns on the revisions of trend estimates. *Journal of Official Statistics*, 16, 61-76.

STEEL, D.G., et DEMEL, R. (1988). The Contribution of Sampling Error to the Variability of Statistical Series. Article présenté à la National Mathematical Sciences Congress, Canberra.

SUTCLIFFE, A. (1993). *X-11 Time Series Decomposition and Sampling Errors*. Document de travail dans Econometrics and Applied Statistics, numéro 93/2. Australian Bureau of Statistics, numéro 1351 au catalogue.

SUTCLIFFE, A., et LEE, G. (1995). Seasonal Analysis and Sample Design. Article présenté à The Conference of Survey Measurement and Process Quality. Bristol 1995.

TILLER, R.B. (1992). Time series modeling of sample survey data from the U.S. Current Population Survey. *Journal of Official Statistics*, 8, 149-166.

WALLIS, K.F. (1982). Seasonal adjustment and revision of current data: linear filters for the x11-method. *Journal of the Royal Statistical Society A*, 145, 74-85.

WOLTER, K.M., et MONSOUR, N.J. (1981). On the problem of variance estimation for a deseasonalized series. *Current Topics in Survey Sampling*, (D. Krewski, R. Platek et J.N.K. Rao, Éds.), Academic Press, New York, 367-407.

YOUNG, A.H. (1968). Linear Approximations to the Census and BLS Seasonal Adjustment Methods. *Journal of the American Statistical Association*, 63,

Techniques d'enquête, décembre 2000
Vol. 26, n° 2, pp. 197-206
Statistique Canada

Estimation bayésienne hiérarchique des moyennes pour petites régions à l'aide de modèles à plusieurs niveaux

YONG YOU et J.N.K. RAO[1]

RÉSUMÉ

On examine les modèles standard à plusieurs niveaux comportant des paramètres de régression aléatoires pour les estimations régionales. Les auteurs élargissent également les modèles en admettant une variance inégale de l'erreur ou en supposant des modèles à effets aléatoires tant pour les paramètres de régression que pour la variance de l'erreur. Les auteurs présentent ces modèles en fonction d'un cadre hiérarchique de Bayes et ils estiment une moyenne régionale en fonction de sa moyenne a posteriori. La variance a posteriori de la moyenne régionale sert de mesure de la précision de l'estimation. Elle tient compte automatiquement de l'incertitude supplémentaire associée aux hyperparamètres du modèle à plusieurs niveaux. Les auteurs utilisent l'échantillonnage de Gibbs pour calculer les moyennes a posteriori et la variance a posteriori de moyennes régionales. Ils obtiennent des estimateurs Rao-Blackwellisés réduisant les erreurs de Monte Carlo. On étudie également la sélection d'un modèle bayésien et l'analyse de la sensibilité. La procédure est illustrée à l'aide de données sur le revenu des ménages de quelques comtés (régions) du Brésil.

MOTS CLÉS: Échantillonnage de Gibbs; hiérarchique de Bayes; modèle à plusieurs niveaux; variance de l'erreur d'échantillonnage; région.

1. INTRODUCTION

Depuis quelques années, on accorde beaucoup d'attention à l'estimation régionale à cause de la demande accrue d'estimateurs régionaux fiables. Les estimateurs directs traditionnels propres à une région ne fournissent pas une précision adéquate puisque la taille des échantillons régionaux est rarement assez grande. Il est donc nécessaire d'utiliser des estimateurs indirects qui s'appuient sur des régions connexes, notamment des estimateurs indirects modélisés. Battese, Harter et Fuller (1988) ont proposé et appliqué un modèle de régression à erreur emboîtée afin d'obtenir des estimations régionales modélisées. Le modèle se présente comme suit:

$$y_{ij} = x_{ij}^T \beta + v_{0i} + e_{ij}, j = 1, ..., n_i; i = 1, ..., m, \qquad (1)$$

où y_{ij} représente les observations associées aux unités échantillonnées de la i-ième région, $i = 1, ..., m$, x_{ij} représente le vecteur $p \times 1$ des variables explicatives au niveau de l'unité, β est un ensemble de p paramètres de régression fixes, v_{0i} sont des effets régionaux indépendants avec $E(v_{0i}) = 0$ et $V(v_{0i}) = \sigma_v^2$. Nous supposons que les e_{ij} sont des variables de l'erreur aléatoire indépendantes avec $E(e_{ij}) = 0$ et $V(e_{ij}) = \sigma_e^2$. v_{0i} et e_{ij} sont également supposés indépendants. Pour l'ensemble de la population, le modèle (1) s'applique avec n_i remplacé par N_i, la taille de la population régionale. Le modèle (1) se laisse exprimer en fonction d'une notation matricielle comme suit:

$$Y_i = X_i \beta + v_{0i} 1_{n_i} + e_i, i = 1, ..., m,$$

où $Y_i = (y_{i1}, ..., y_{i,n_i})^T, X_i = (x_{i1}, ..., x_{i,n_i})^T$ est une matrice $n_i \times p$, $1_{n_i} = (1, ..., 1)^T$ est le vecteur unité de longueur n_i et $e_i = (e_{i1}, ..., e_{i,n_i})^T$.

Holt et Moura (1993) ont élargi le cadre ci-dessus pour en faire un modèle à plusieurs niveaux en introduisant des coefficients de régression aléatoires et en les rattachant à des variables explicatives au niveau de la région afin d'expliquer une partie des variations entre les régions. Le modèle se laisse énoncer comme suit:

$$Y_i = X_i \beta_i + e_i, \beta_i = Z_i \gamma + v_i \qquad (2)$$

où Z_i est la matrice du plan $p \times q$ des variables au niveau de la région, γ est un vecteur $q \times 1$ de coefficients fixes et $v_i = (v_{i1}, ..., v_{ip})^T$ est un vecteur $p \times 1$ d'effets aléatoires pour la i-ième région. Les v_i sont indépendants d'une région à l'autre et comportent une distribution composée au sein de chaque région avec $E(v_i) = 0$ et $V(v_i) = \Phi$, où la matrice de variances-covariances Φ est inconnue. À noter que le modèle (2) intègre effectivement le recours à des covariables au niveau de l'unité et au niveau de la région en un même modèle. Holt et Moura (1993) et Moura et Holt (1999) ont élargi le cadre adopté par Prasad et Rao (1990) pour en faire le modèle à plusieurs niveaux ci-dessus de façon à obtenir le meilleur prédicteur linéaire sans biais (MPLSB) de la moyenne régionale $\mu_i = \bar{X}_i^T \beta_i$ en supposant N_i grand, où $\bar{X}_i$ est le vecteur $p \times 1$ des moyennes connues de la population des variables auxiliaires pour la i-ième région. Ils ont également obtenu le MPLSB empirique et une approximation d'ordre deux de l'erreur quadratique moyenne (EQM) du MPLSB empirique pour le modèle à plusieurs niveaux. À l'aide de données sur le revenu des ménages de

[1] Yong You, Division des métodes des enquêtes auprès des ménages, Statistique Canada, Ottawa (Ontario) Canada K1A 0T6; J.N.K. Rao, School of Mathematics and Statistics, Université Carleton, Ottawa (Ontario) Canada K1S 5B6.

comtés (régions) du Brésil, ils ont obtenu un accroissement de l'efficacité des estimateurs de type MPLSB empirique relativement aux estimateurs MPLSB empiriques obtenus à l'aide de modèles de régression à erreur emboîtée. Ghosh et Rao (1994) et Rao (1999) ont donné un aperçu détaillé des méthodes d'estimation régionale modélisées.

Dans le présent exposé, nous examinons le modèle (2) à plusieurs niveaux selon un cadre hiérarchique de Bayes et nous l'étendons à des modèles à plusieurs niveaux plus généraux qui admettent des variances de l'erreur inégale ou des variances de l'erreur aléatoire fixes. La moyenne régionale μ_i est estimée à l'aide de sa moyenne a posteriori, et sa précision est mesurée en fonction de sa variance a posteriori. La variance a posteriori tient compte automatiquement de l'incertitude supplémentaire associée aux hyperparamètres du modèle à plusieurs niveaux. Nous utilisons la méthode d'échantillonnage de Gibbs afin d'obtenir les estimations bayésiennes hiérarchiques et les variances a posteriori connexes. À la section 2, nous présentons les modèles bayésiens hiérarchiques à plusieurs niveaux avec diverses hypothèses pour ce qui est de la variance de l'erreur et de l'inférence de l'échantillonnage de Gibbs connexe. À la section 3, nous illustrons notre méthode et nous abordons la sélection du modèle et l'analyse de la sensibilité à l'aide de données sur le revenu des ménages de comtés (régions) du Brésil. Enfin, à la section 4, nous présentons des remarques et nos conclusions.

2. MODÈLE À PLUSIEURS NIVEAUX ET INFÉRENCE DE L'ÉCHANTILLONNAGE DE GIBBS

2.1 Variance de l'erreur égale

Nous considérons une représentation bayésienne hiérarchique du modèle (2) à plusieurs niveaux comme suit:

Modèle 1:

(i) Sous réserve que β_i et σ_e^2, les y_{ij} sont indépendants avec

$$y_{ij}|\beta_i, \sigma_e^2 \sim N(x_{ij}^T\beta_i, \sigma_e^2), \quad (i = 1, ..., m; j = 1, ..., n_i); \tag{3}$$

(ii) Sous réserve que γ et Φ, les β_i sont indépendants avec

$$\beta_i|\gamma, \Phi \sim N_p(Z_i\gamma, \Phi), (i = 1, ..., m). \tag{4}$$

Afin de compléter notre description de modèle bayésien, nous adaptons les distributions a priori pour des paramètres comme suit:

(iii) Distributions a priori marginales: $\gamma \sim N_q(0, D)$, $\tau_e \sim G(a, b)$ et $\Omega \sim W_p(\alpha, R)$, où $\tau_e = \sigma_e^{-2}$, $\Omega = \Phi^{-1}$ et D, a, b, α et R connus.

À l'étape (iii) du modèle 1, $G(a, b)$ désigne une distribution gamma avec une densité donnée par $f(x) = b^a/\Gamma(a)x^{a-1}e^{-xb}$, $a > 0, b > 0, x \geq 0$ et $W_p(\alpha, R)$ est une distribution de Wishart avec comme fonction de densité

$$f(X) = \frac{|R|^{\frac{\alpha}{2}}}{2^{\alpha p/2}\Gamma_p\left(\frac{\alpha}{2}\right)}|X|^{\frac{\alpha-p-1}{2}}\exp\{-\frac{1}{2}\operatorname{tr}(RX)\},$$

où $X > 0, R > 0$ et $\Gamma_p(\alpha)$ est une fonction gamma à plusieurs variables définie comme

$$\Gamma_p(\alpha) = \pi^{\frac{p(p-1)}{4}}\prod_{j=1}^{p}\Gamma\left(\alpha + \frac{1}{2}(1-j)\right).$$

Remarque 1.1: Les distributions a priori à l'étape (iii) et les distributions d'échantillonnage et de population données par (3) et par (4) sont conjuguées en ce sens qu'elles mènent à des distributions conditionnelles complètes pour γ, τ_e et Ω qui sont encore une fois des distributions normale, gamma et de Wishart respectivement. La distribution de Wishart est la version à plusieurs variables d'une distribution gamma pour la matrice de variances-covariances inverse d'effets aléatoires. L'importance de la conjugaison peut être exploitée comme suit: (1) à l'étape de l'échantillonnage de Gibbs, en l'absence de conjugaison, la distribution conditionnelle complète pour un paramètre quelconque sera connue jusqu'aux constantes normalisatrices; dans ce cas, une génération aléatoire plus perfectionnée sera requise; (2) des distributions conditionnelles complètes de forme fermée peuvent servir à trouver les estimateurs Rao-Blackwellisés des moyennes a posteriori et des variances a posteriori, et donc à améliorer l'estimation a posteriori. En général, pour une inférence bayésienne, le choix de distributions a priori n'est pas une tâche simple puisque toute distribution a priori appropriée sur les paramètres de modèle est un candidat plausible. C'est là une limitation des méthodes bayésiennes.

Remarque 1.2: Il importe de noter que nous avons utilisé des distributions a priori appropriées sur tous les paramètres inconnus afin de nous assurer que toutes les distributions a posteriori sont appropriées (Hobert et Casella 1996). Ainsi, nous n'avons pas à craindre la présence de distributions a posteriori inappropriées. La valeur des paramètres des distributions a priori (les hyperparamètres) est choisie de façon à refléter une connaissance assez vague des distributions a priori. Des détails seront fournis à la section 3 sur l'analyse des données.

Remarque 1.3: Dans le modèle 1, nous supposons une variance de l'erreur égale σ_e^2 pour toutes les régions. Dans la pratique, toutefois, les variances de l'erreur d'échantillonnage pourraient être différentes pour diverses régions. Un modèle plus général devrait admettre des variances d'erreur possiblement différentes. Dans les sections 2.2 et 2.3, nous introduirons des modèles de variance d'erreur inégale et de variance d'erreur aléatoire.

Nous voulons trouver les distributions a posteriori des β_i pour les données $Y = (\{y_{ij}\}, i = 1, ..., m; j = 1, ..., n_i)$ et, en particulier, trouver les estimations a posteriori des moyennes régionales $\mu_i = \overline{X}_i^T \beta_i$, qui dépendent des estimations de β_i. L'évaluation directe de la distribution a posteriori composée comporte une intégration numérique très dimensionnelle, et le calcul n'est pas faisable. Par conséquent, nous utilisons la méthode d'échantillonnage de Gibbs (Gelfand et Smith 1990) afin de produire des échantillons à partir des distributions a posteriori composées. Pour appliquer l'échantillonnage de Gibbs dans le cadre du modèle 1, il nous faut les distributions conditionnelles complètes données par:

(i) Pour $i = 1, ..., m$,

$$[\beta_i | Y, \gamma, \Omega, \tau_e] \overset{\text{ind}}{\sim} N_p\left((\tau_e X_i^T X_i + \Omega)^{-1} (\tau_e X_i^T Y_i + \Omega Z_i \gamma), (\tau_e X_i^T X_i + \Omega)^{-1}\right)$$

(ii) $$[\gamma | Y, \beta, \Omega, \tau_e] \sim N_q\left(\left(\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1}\right)^{-1} \left(\sum_{i=1}^m Z_i^T \Omega \beta_i\right), \left(\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1}\right)^{-1}\right)$$

(iii) $$[\Omega | Y, \beta, \gamma, \tau_e] \sim W_p\left(\alpha + m, R + \frac{1}{2}\sum_{i=1}^m (\beta_i - Z_i\gamma)(\beta_i - Z_i\gamma)^T\right)$$

(iv) $$[\tau_e | Y, \beta, \gamma, \Omega] \sim G\left(a + \frac{1}{2}\sum_{i=1}^m n_i, b + \frac{1}{2}\left(\sum_{i=1}^m (Y_i - X_i\beta_i)^T (Y_i - X_i\beta_i)\right)\right).$$

Toutes les distributions conditionnelles complètes sont de forme fermée et permettent facilement de produire des échantillons. La méthode d'échantillonnage de Gibbs se présente comme suit: a) À l'aide de valeurs de départ $\gamma^{(0)}, \Omega^{(0)}$ et $\tau_e^{(0)}$, il s'agit de tirer $\beta_i^{(1)}$, $i = 1, ..., m$, à partir de $[\beta_i | Y, \gamma, \Omega, \tau_e]$; b) de tirer $\gamma^{(1)}$ à partir de $[\gamma | Y, \beta, \Omega, \tau_e]$ à l'aide de $\beta_i^{(1)}$, $i = 1, ..., m$, $\Omega^{(0)}$ et $\tau_e^{(0)}$; c) de tirer $\Omega^{(1)}$ à partir de $[\Omega | Y, \beta, \gamma, \tau_e]$ à l'aide de $\beta_i^{(1)}$, $i = 1, ..., m$, $\gamma^{(1)}$ et $\tau_e^{(0)}$; d) de tirer $\tau_e^{(1)}$ à partir de $[\tau_e | Y, \beta, \gamma, \Omega]$ à l'aide de $\beta_i^{(1)}$, $i = 1, ..., m$, $\gamma^{(1)}$ et $\Omega^{(1)}$. Les étapes (a)-(d) permettent de compléter un cycle d'échantillonnage. On exécute un nombre élevé de cycles, disons t, période dite de rodage, jusqu'à la convergence, puis on traite $\{\beta_i^{(t+k)}, i = 1, ..., m; \gamma^{(t+k)}; \Omega^{(t+k)}; \tau_e^{(t+k)}; k = 1, ..., G\}$ comme des échantillons G à partir de la distribution a posteriori composée de β_i, $i = 1, ..., m$, γ, Ω et τ_e.

Supposons qu'un échantillon de taille G, de la forme $\{\beta_i^{(k)}, i = 1, ..., m; \gamma^{(k)}; \Omega^{(k)}; \tau_e^{(k)}; k = 1, ..., G\}$ soit obtenu. Pour obtenir un estimateur de la moyenne a posteriori de β_i, on peut utiliser la moyenne d'échantillon du $\{\beta_i^{(k)}\}$. Puisque β_i comporte une distribution conditionnelle complète de forme fermée, nous pouvons utiliser la moyenne d'échantillon des espérances conditionnelles $\{E[\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)}]\}$ afin d'améliorer notre estimation, puisque $E(\beta_i | Y) = E(E(\beta_i | Y, \gamma, \Omega, \tau_e))$, et $\text{Var}(\beta_i | Y) \geq \text{Var}(E(\beta_i | Y, \gamma, \Omega, \tau_e))$. Cette modification se fonde sur le théorème bien connu de Rao-Blackwell, et l'estimateur correspondant est l'estimateur soi-disant Rao-Blackwellisé (Gelfand et Smith 1990, 1991). Nous avons ainsi les deux autres estimateurs ci-dessous pour β_i:

$$\hat{\beta}_i^{(E)} = \frac{1}{G}\sum_{k=1}^G \beta_i^{(k)} \tag{5}$$

et

$$\hat{\beta}_i^{(\text{RB})} = \frac{1}{G}\sum_{k=1}^G E(\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)}) = \frac{1}{G}\sum_{k=1}^G (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)}), \tag{6}$$

où $\hat{\beta}_i^{(E)}$ est l'estimateur empirique et $\hat{\beta}_i^{(\text{RB})}$ est l'estimateur Rao-Blackwellisé. $\hat{\beta}_i^{(E)}$ et $\hat{\beta}_i^{(\text{RB})}$ sont tous deux sans biais pour la moyenne a posteriori. Toutefois, $\hat{\beta}_i^{(\text{RB})}$ est meilleur que $\hat{\beta}_i^{(E)}$ pour ce qui est de l'erreur-type de simulation (Gelfand et Smith 1991).

Les estimateurs correspondants pour la moyenne régionale μ_i sont donnés comme suit

$$\hat{\mu}_i^{(E)} = \overline{X}_i^T \hat{\beta}_i^{(E)} = \frac{1}{G}\sum_{k=1}^G \overline{X}_i^T \beta_i^{(k)} \tag{7}$$

et

$$\hat{\mu}_i^{(\text{RB})} = \overline{X}_i^T \hat{\beta}_i^{(\text{RB})} = \frac{1}{G}\sum_{k=1}^G \overline{X}_i^T (\tau_e^{(k)} X_i^T X_i + \Omega^{(k)})^{-1} (\tau_e^{(k)} X_i^T Y_i + \Omega^{(k)} Z_i \gamma^{(k)}). \tag{8}$$

Nous nous attendons à ce que $\hat{\mu}_i^{(E)}$ et $\hat{\mu}_i^{(\text{RB})}$ donnent tous deux presque les mêmes estimations ponctuelles. Toutefois, il est intéressant de calculer et de comparer les erreurs-types de simulation de ces deux estimateurs en vue de l'évaluation des effets de la Rao-Blackwellisation (voir la section 3).

Afin d'obtenir la variance a posteriori de μ_i, nous trouvons d'abord la variance a posteriori de β_i, puisque $V(\mu_i | Y) = \overline{X}_i^T V(\beta_i | Y) \overline{X}_i$. À noter que

$$\begin{aligned} V(\beta_i | Y) &= E(V(\beta_i | Y, \gamma, \Omega, \tau_e)) + V(E(\beta_i | Y, \gamma, \Omega, \tau_e)) \\ &= E(V(\beta_i | Y, \gamma, \Omega, \tau_e)) + E(E(\beta_i | Y, \gamma, \Omega, \tau_e)^2) \\ &\quad - [E(E(\beta_i | Y, \gamma, \Omega, \tau_e))]^2. \end{aligned} \tag{9}$$

À l'aide de (9), l'estimateur Rao-Blackwellisé de la variance a posteriori de β_i, noté $\hat{V}(\beta_i)$, s'obtient à l'aide des échantillons de Gibbs $\{\beta_i^{(k)}, i = 1, ..., m; \gamma^{(k)}; \Omega^{(k)}; \tau_e^{(k)}; k = 1, ..., G\}$; voir l'annexe A1. La variance a posteriori de la moyenne régionale μ_i est alors estimée à l'aide de

$$\hat{V}(\mu_i) = \overline{X}_i^T \hat{V}(\beta_i) \overline{X}_i. \quad (10)$$

On peut appliquer la même procédure d'estimation à la variance de l'erreur d'échantillonnage σ_e^2. Puisque, conditionnellement, σ_e^2 comporte une distribution gamma inverse, l'estimateur Rao-Blackwellisé de la moyenne a posteriori de σ_e^2 s'obtient sous la forme

$$\hat{\sigma}_e^{2(\mathrm{RB})} = \frac{1}{G}\sum_{k=1}^{G}\left[b + \frac{1}{2}\sum_{i=1}^{m}(Y_i - X_i\beta_i^{(k)})^T(Y_i - X_i\beta_i^{(k)})\right] \left(a + \frac{1}{2}\sum_{i=1}^{m} n_i - 1\right)^{-1}. \quad (11)$$

Puisque nous nous intéressons surtout à l'estimation des moyennes régionales, le calcul de la variance d'échantillonnage sert uniquement à la sélection du modèle. On trouvera des détails sur la sélection du modèle à la section 3.2.

2.2 Variance de l'erreur inégale

Dans la pratique, il est plus réaliste d'admettre une variance d'erreur inégale pour les erreurs d'échantillonnage. Soit σ_i^2, la vraie variance de l'erreur d'échantillonnage pour la i-ième région. Une extension directe du modèle 1 mène au modèle bayésien hiérarchique à plusieurs niveaux de la variance d'erreur inégale que voici:

Modèle 2:

(i) Sous réserve que β_i et σ_i^2, les y_{ij} sont indépendants avec

$$y_{ij} \mid \beta_i, \sigma_i^2 \sim N(x_{ij}^T\beta_i, \sigma_i^2), \quad (i = 1, ..., m; j = 1, ..., n_i); \quad (12)$$

(ii) Sous réserve que γ et Φ, les β_i sont indépendants avec

$$\beta_i \mid \gamma, \Phi \sim N_p(Z_i\gamma, \Phi), (i = 1, ..., m); \quad (13)$$

(iii) Distributions a priori marginales: $\gamma \sim N_q(0, D)$, $\tau_i \overset{\text{ind}}{\sim} G(a_i, b_i)$ et $\Omega \sim W_p(\alpha, R)$, où $\tau_i = \sigma_i^{-2}$, $\Omega = \Phi^{-1}$ et D, a_i, b_i, α et R connus.

Remarque 2.1: Le modèle 2 se réduit au modèle 1 lorsque $\sigma_i^2 = \sigma_e^2$ pour tous les i. D'un point de vue hiérarchique de Bayes, l'extension du modèle de la variance d'erreur égale au modèle de la variance d'erreur inégale va de soi. Il n'y a pas non plus de difficulté pour la mise en œuvre de l'échantillonnage de Gibbs.

Remarque 2.2: Les τ_i sont supposés indépendants et comportent des distributions a priori $G(a_i, b_i)$, où a_i et b_i sont des hyperparamètres connus normalement choisis très petits afin de refléter une vague connaissance des τ_i.

Les distributions conditionnelles complètes pour un échantillonnage de Gibbs dans le cadre du modèle 2 sont données par:

(i) Pour $i = 1, ..., m$,

$$[\beta_i \mid Y, \gamma, \Omega, \tau] \overset{\text{ind}}{\sim} N_p((\tau_i X_i^T X_i + \Omega)^{-1}(\tau_i X_i^T Y_i + \Omega Z_i\gamma), (\tau_i X_i^T X_i + \Omega)^{-1})$$

(ii)
$$[\gamma \mid Y, \beta, \Omega, \tau] \sim N_q\left(\left(\sum_{i=1}^{m} Z_i^T\Omega Z_i + D^{-1}\right)^{-1}\left(\sum_{i=1}^{m} Z_i^T\Omega\beta_i\right), \left(\sum_{i=1}^{m} Z_i^T\Omega Z_i + D^{-1}\right)^{-1}\right)$$

(iii)
$$[\Omega \mid Y, \beta, \gamma, \tau] \sim W_p\left(\alpha + m, R + \frac{1}{2}\sum_{i=1}^{m}(\beta_i - Z_i\gamma)(\beta_i - Z_i\gamma)^T\right)$$

(iv) Pour $i = 1, ..., m$,

$$[\tau_i \mid Y, \beta, \gamma, \Omega] \sim G\left(a_i + \frac{1}{2}n_i, b_i + \frac{1}{2}\times(Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i)\right).$$

Pour le modèle 2, les estimateurs empiriques des moyennes a posteriori de β_i et μ_i ont la même forme que (5) et (7). Les estimateurs Rao-Blackwellisés $\hat{\beta}_i^{(\mathrm{RB})}$ et $\hat{\mu}_i^{(\mathrm{RB})}$ sont les estimateurs donnés par (6) et (8) avec $\tau_e^{(k)}$ remplacé par $\tau_i^{(k)}$. L'estimateur de la variance a posteriori est $\hat{V}(\mu_i)$ donné par (10) avec $\tau_e^{(k)}$ remplacé par $\tau_i^{(k)}$.

En vue de la sélection du modèle et de la comparaison des modèles, nous trouvons également l'estimateur Rao-Blackwellisé de la moyenne a posteriori de σ_i^2 dans le cadre du modèle 2 de la forme

$$\hat{\sigma}_i^{2(\mathrm{RB})} = \frac{1}{G}\sum_{k=1}^{G}\left[b_i + \frac{1}{2}(Y_i - X_i\beta_i^{(k)})^T(Y_i - X_i\beta_i^{(k)})\right] \times (a_i + \frac{1}{2}n_i - 1)^{-1}. \quad (14)$$

2.3 Variance de l'erreur aléatoire

Dans le modèle 2, nous avons supposé une variance d'erreur inégale pour les erreurs d'échantillonnage. Kleffe et Rao (1992) ont utilisé un modèle simple de la variance d'erreur aléatoire afin d'obtenir les meilleurs prédicteurs linéaires sans biais pour des moyennes régionales. Dans la présente section, nous étendons leur modèle au cas comportant plusieurs niveaux. Nous supposons des modèles d'effets aléatoires sur les coefficient de régression β_i aussi bien que sur les variances de l'erreur d'échantillonnage σ_i^2, ce qui mène au modèle 3 donné ci-dessous.

Modèle 3:

(i) Comme pour le modèle 2;

(ii) Comme pour le modèle 2;

(iii) Sous réserve de η et de λ, les τ_i sont indépendants avec

$$\tau_i \mid \eta, \lambda \overset{\text{iid}}{\sim} G(\eta, \lambda), \quad (15)$$

où $\tau_i = \sigma_i^{-2}$;

(iv) Distributions a priori marginales : $\gamma \sim N_q(0, D)$, $\Omega \sim W_p(\alpha, R)$, $\eta \sim U^+$ et $\lambda \sim U^+$, où U^+ désigne une distribution uniforme sur un sous-ensemble de R^+ avec une longueur grande mais finie, D, α et R connus.

Remarque 3.1: Dans le modèle 3, nous supposons que les τ_i sont des variables aléatoires gamma indépendantes et distribuées de façon identique avec des hyperparamètres η et λ inconnus. Nous avons ainsi des modèles de population pour les coefficients de régression β_i aussi bien que pour la variance d'échantillonnage σ_i^2. Dans les modèles 1 et 2, nous avons considéré la modélisation de β_i seulement et nous avons supposé de vagues distributions a priori appropriées sur σ_e^2 ou σ_i^2.

Remarque 3.2: L'hypothèse (iii) n'est peut-être pas un bon modèle de population pour tous les τ_i. Comme autre possibilité, nous pouvons modéliser τ_i de façon plus réaliste, comme dans le cas de β_i, en précisant un modèle de régression pour le logarithme de τ_i. Cela peut exiger des renseignements auxiliaires liés à τ_i. Dans l'analyse des données de la section 3, toutefois, nous utilisons simplement $G(\eta, \lambda)$ comme modèle de population pour τ_i. Il n'est généralement pas facile de modéliser les variances d'échantillonnage lorsqu'elles sont inconnues.

Les distributions conditionnelles complètes pour un échantillonnage de Gibbs dans le cadre du modèle 3 sont données par:

(i) Pour $i = 1, ..., m$,

$$[\beta_i | Y, \tau, \gamma, \Omega, \eta, \lambda] \overset{\text{ind}}{\sim} N_p\left((\tau_i X_i^T X_i + \Omega)^{-1}\left(\tau_i X_i^T Y_i + \Omega Z_i \gamma\right), (\tau_i X_i^T X_i + \Omega)^{-1}\right)$$

(ii) Pour $i = 1, ..., m$,

$$[\tau_i | Y, \beta, \gamma, \Omega, \eta, \lambda] \overset{\text{ind}}{\sim} G\left(\eta + \frac{n_i}{2}, \frac{1}{2}(Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i) + \lambda\right)$$

(iii) $$[\gamma | Y, \beta, \tau, \eta, \lambda] \sim N_q\left(\left(\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1}\right)^{-1}\left(\sum_{i=1}^m Z_i^T \Omega \beta_i\right), \left(\sum_{i=1}^m Z_i^T \Omega Z_i + D^{-1}\right)^{-1}\right)$$

(iv) $$[\Omega | Y, \beta, \sigma^2, \gamma, \eta, \lambda] \sim W_p\left(\alpha + m, R + \frac{1}{2}\sum_{i=1}^m (\beta_i - Z_i\gamma)(\beta_i - Z_i\gamma)^T\right)$$

(v) $$[\eta | Y, \beta, \tau, \gamma, \Omega, \lambda] \propto [\Gamma(\eta)]^{-m} \lambda^{m\eta} (\Pi_{i=1}^m \tau_i)^{\eta}$$

(vi) $$[\lambda | Y, \beta, \tau, \gamma, \Omega, \eta] \sim G\left(m\eta + 1, \sum_{i=1}^m \tau_i\right)$$

Pour le modèle 3, les estimateurs a posteriori de β_i et μ_i ont les mêmes formes que celles qui ont été données pour le modèle 2. Dans le cadre du modèle 3, l'estimateur Rao-Blackwellisé de la moyenne a posteriori de σ_i^2 est donné par

$$\hat{\sigma}_i^{2(\text{RB})} = \frac{1}{G}\sum_{k=1}^G [\lambda^{(k)} + \frac{1}{2}(Y_i - X_i\beta_i^{(k)})^T \times (Y_i - X_i\beta_i^{(k)})](\eta^{(k)} + \frac{1}{2}n_i - 1)^{-1}. \quad (16)$$

Dans le cadre du modèle 3, $[\eta | Y, \beta, \tau, \gamma, \Omega, \lambda]$ est connu jusqu'à une constante multiplicatrice seulement. Toutefois, puisque $[\eta | Y, \beta, \tau, \gamma, \Omega, \lambda]$ est une fonction concave logarithmique de η (voir l'annexe A2), il est possible d'utiliser la méthode d'échantillonnage à rejet adapté de Gilks, Best et Tan (1995) dans l'échantillonneur de Gibbs afin de produire des échantillons à partir de la distribution conditionnelle $[\eta | Y, \beta, \tau, \gamma, \Omega, \lambda]$.

3. ANALYSE DES DONNÉES

3.1 Description des données et des modèles

Suivant Holt et Moura (1993) et Moura et Holt (1999), nous avons considéré l'estimation du revenu des ménages de comtés (régions) du Brésil. Les données originales de Holt et Moura comportent 140 régions, les unités d'échantillonnage étant tirées de chaque région par échantillonnage aléatoire simple. La méthode hiérarchique de Bayes n'exige pas que le nombre de régions soit grand, contrairement à la méthode du MPLSB empirique, pour l'obtention des erreurs-types. Par conséquent, nous avons utilisé uniquement une faible partie de l'ensemble original de données dans notre analyse, à simple titre d'illustration. Notre ensemble de données contient un sous-ensemble de 10 régions comportant 28 unités d'échantillonnage obtenues par échantillonnage aléatoire simple dans chaque région.

Soit y_{ij}, le revenu du j-ième ménage dans la i-ième région. Il y a deux variables auxiliaires au niveau de l'unité, c'est-à-dire x_1 et x_2, où x_1 désigne le nombre de pièces dans un ménage et x_2 désigne le niveau de scolarité atteint par le chef du ménage. Le modèle d'échantillonnage est donné par

$$y_{ij} = x_{ij}^T\beta_i + e_{ij} = \beta_{0i} + x_{1ij}\beta_{1i} + x_{2ij}\beta_{2i} + e_{ij}, \quad (17)$$

où x_{1ij} désigne le nombre de pièces dans le j-ième ménage de la région i et x_{2ij} désigne le niveau de scolarité correspondant atteint par le chef du ménage. Les valeurs de x_{1ij} et de x_{2ij} sont centrées autour de leurs moyennes d'échantillon global respectives et e_{ij} est la variable de l'erreur d'échantillonnage, sa distribution étant précisée par les trois modèles de la variance d'erreur discutées à la section 2.

Dans le modèle d'échantillonnage (17), β_i est le coefficient de régression aléatoire correspondant à la i-ième région et se laisse modéliser comme

$$\beta_{0i} = \gamma_0 + \nu_{0i}, \beta_{1i} = \gamma_{10} + \gamma_{11}z_i + \nu_{1i}, \beta_{2i} = \gamma_{20} + \gamma_{21}z_i + \nu_{2i},$$

où $\gamma = (\gamma_0, \gamma_{10}, \gamma_{11}, \gamma_{20}, \gamma_{21})^T$ est le vecteur inconnu de paramètres de régression fixes, $\nu_i = (\nu_{0i}, \nu_{1i}, \nu_{2i})^T$ est le vecteur d'effets aléatoires de la i-ième région distribué comme

$v_i \sim N_3(0, \Phi)$, et z_i est une variable au niveau de la région définie comme le nombre moyen de voitures par ménage dans chaque région. La valeur de z_i est également centrée autour de sa moyenne d'échantillon globale.

Nous avons utilisé les trois modèles discutés à la section 2 pour notre analyse des données. De vagues distributions a priori appropriées sur des paramètres inconnus sont spécifiées comme suit: $\gamma \sim N_5(0, D)$, où $D = \text{diag}(10^4, 10^4, 10^4, 10^4, 10^4)$, donc $\gamma_0, \gamma_{10}, \gamma_{11}, \gamma_{20}, \gamma_{21}$ sont supposées être des variables normales indépendantes de moyenne 0 et d'écart-type 100, de sorte qu'un intervalle a priori de 95% correspond à ±200 environ, et la distribution a priori est uniforme localement pour la région appuyée par la vraisemblance. Comme autre possibilité, une distribution a priori uniforme sur un intervalle convenablement large pourrait être donnée, par exemple $U(-200, 200)$. Une distribution a priori de Wishart $W_3(\alpha, R)$ est spécifiée pour la matrice de covariances inverse $\Omega = \Phi^{-1}$. Pour représenter une vague connaissance a priori, nous avons choisi les degrés de liberté α pour que cette distribution soit aussi petite que possible, $\alpha = 3$, le rang de Ω (Spiegelhalter, Thomas, Best et Gilks 1996). La matrice scalaire R est décrite avec des éléments diagonaux égaux à 1 et des éléments non diagonaux égaux à 0,001, ce qui représente notre estimation préliminaire pour l'ordre de grandeur de la matrice de covariances. Pour les modèles 1 et 2, une distribution a priori gamma G(0,001, 0,001) est supposée pour τ_e et les τ_i. Pour le modèle 3, $\tau_i \sim G(\eta, \lambda)$, et on suppose que η et λ sont distribués indépendamment comme U(0,10000), c'est-à-dire la distribution uniforme sur un grand intervalle. Nous nous attendons à ce que les vagues distributions a priori appropriées sur les hyperparamètres se rapprochent raisonnablement bien des distributions a priori uniformes et qu'elles exercent ainsi un effet minimal sur l'estimation a posteriori.

Nous avons mis en œuvre l'échantillonneur de Gibbs pour les trois modèles à l'aide du programme BUGS (Spiegelhalter et coll. 1996) et de la fonction CODA Splus (Best, Cowles et Vines 1996) pour l'évaluation de la convergence. Le programme BUGS permet de construire les distributions conditionnelles complètes nécessaires et d'exécuter l'échantillonnage de Gibbs pourvu que nos modèles soient décrits à l'aide du langage BUGS. Les distributions a priori et les valeurs initiales des paramètres doivent être décrites dans le programme. Pour chaque modèle, l'échantillonneur de Gibbs a d'abord été exécuté au cours d'une période dite de rodage de 2 000 itérations, puis 5 000 autres itérations ont été exécutées et conservées en vue de l'analyse et de l'estimation du modèle.

Nous cherchons à estimer la moyenne régionale $\mu_i = \bar{X}_i^T \beta_i = \beta_{0i} + \bar{X}_{1i}\beta_{1i} + \bar{X}_{2i}\beta_{2i}$, où $\bar{X}_{1i}$ et $\bar{X}_{2i}$ sont les moyennes de population de la i-ième région des variables auxiliaires x_1 et x_2, respectivement. Pour ce faire, nous allons d'abord choisir un modèle optimal pour l'ensemble de données, et nous allons présenter les estimations modélisées pour les moyennes régionales fondées sur le modèle choisi.

3.2 Sélection du modèle

À la section 2, nous avons proposé trois modèles en fonction d'hypothèses différentes quant aux variances d'échantillonnage. Afin de déterminer quel modèle ajuste les données, nous avons d'abord obtenu les estimations a posteriori des variances d'échantillonnage dans le cadre des trois modèles. Nous avons également calculé les estimations des moindres carrés ordinaires (MCO) des variances d'échantillonnage au sein de chaque région à l'aide des données propres à la région seulement. Le tableau 1 indique les estimations Rao-Blackwellisées des variances d'échantillonnage dans le cadre des trois modèles de même que les estimations MCO.

Tableau 1
Variances estimatives de l'erreur d'échantillonnage

Région	MCO	Modèle 1	Modèle 2	Modèle 3
1	38,17	76,86	40,18	63,6
2	31,75	76,86	34,24	62,13
3	81,26	76,86	94,77	79,58
4	48,73	76,86	52,01	67,27
5	115,98	76,86	121,65	87,70
6	90,74	76,86	94,35	79,78
7	101,67	76,86	101,67	82,14
8	135,65	76,86	159,94	97,96
9	59,1	76,86	63,37	70,57
10	62,86	76,86	65,72	71,22

Dans le tableau 1, les estimations MCO indiquent de fortes variations parmi les 10 régions. Le modèle 1 suppose une variance de l'erreur égale σ_e^2 pour toutes les régions, et on estime σ_e^2 par $\hat{\sigma}_e^{2(RB)} = 76,86$, valeur beaucoup plus faible que les estimations MCO pour certaines régions. Le modèle 2 suppose une variance de l'erreur inégale σ_i^2 d'une région à l'autre. Dans le cadre du modèle 2, les variances de l'erreur estimée $\hat{\sigma}_i^{2(RB)}$ montrent dans une certaine mesure l'aspect des régions; les $\hat{\sigma}_i^{2(RB)}$ concordent avec la tendance des estimations MCO. Le résultat le plus remarquable est $\hat{\sigma}_5^{2(RB)} = 121,65$ et $\hat{\sigma}_8^{2(RB)} = 159,94$, indiquant de plus fortes variations au sein des régions 5 et 8. Le modèle 3 suppose que les σ_i^2 sont des variables aléatoires distribuées comme $G(\eta, \lambda)$. Dans le cadre du modèle 3, toutes les $\hat{\sigma}_i^{2(RB)}$ tendent à être égales à $\hat{\sigma}_e^{2(RB)} = 76,86$ et y sont déplacées. Les résultats du tableau 1 indiquent que le modèle 2, c'est-à-dire le modèle de la variance de l'erreur inégale, pourrait être le modèle optimal pour notre ensemble de données. En guise d'approfondissement, nous présentons maintenant une étude de validation croisée en vue de la sélection d'un modèle d'ajustement optimal.

Afin d'examiner comment les données appuient chaque modèle, nous avons calculé la densité prédictive de validation croisée pour chaque donnée simple y_{ij}. La densité de validation croisée pour y_{ij} est la densité conditionnelle $f(y_{ij} | Y_{(ij)})$, où $Y_{(ij)}$ désigne toutes les données sauf y_{ij}. Nous

avons examiné la valeur de $f(y_{ij} \mid Y_{(ij)})$ pour la donnée simple observée, la soi-disant ordonnée prédictive conditionnelle (CPO) pour chacun des trois modèles. Nous avons

$$\text{CPO}_{ij} = f(y_{ij,\,\text{obs}} | Y_{(ij),\,\text{obs}}),$$

où $y_{ij,\,\text{obs}}$ désigne la donnée simple observée. Puisque les CPO ne sont que les vraisemblances observées, les modèles comportant des CPO plus élevées fournissent un meilleur ajustement pour les données observées. À l'aide de la sortie de l'échantillonneur de Gibbs, nous pouvons calculer les CPO pour toutes les données simples. Ainsi, dans le cadre du modèle 1, nous avons

$$f(y_{ij}|Y_{(ij)}) = \frac{f(Y)}{f(Y_{(ij)})}$$

$$= \frac{1}{\int \frac{f(Y_{(ij)}, \beta_i, \sigma_e^2)}{f(Y, \beta_i, \sigma_e^2)} \cdot f(\beta_i, \sigma_e^2|Y) d\beta_i d\sigma_e^2}$$

$$= \frac{1}{\int \frac{1}{f(y_{ij}|Y_{(ij)}, \beta_i, \sigma_e^2)} \cdot f(\beta_i, \sigma_e^2|Y) d\beta_i d\sigma_e^2}.$$

Nous notons maintenant que les y_{ij} sont conditionnellement indépendantes, donc $f(y_{ij}|Y_{(ij)}, \beta_i, \sigma_e^2) = f(y_{ij}|\beta_i, \sigma_e^2)$, et les valeurs CPO sont calculées comme suit:

$$\widehat{\text{CPO}}_{ij} = \frac{1}{\frac{1}{G}\sum_{k=1}^{G} \frac{1}{f(y_{ij,\,\text{obs}}|\beta_i^{(k)}, \sigma_e^{2(k)})}}, \qquad (18)$$

où $f(y_{ij}|\beta_i, \sigma_e^2)$ est la fonction de densité normale donnée par (3). Pour les modèles 2 et 3, on calcule les CPO avec $\sigma_e^{2(k)}$ remplacée par $\sigma_i^{2(k)}$ en (18). On trouvera une discussion plus détaillée dans Gelfand (1995).

Nous présentons un tracé CPO pour les trois modèles à la figure 1. Le modèle 2 est clairement le meilleur, puisqu'une majorité de valeurs CPO pour le modèle 2 est appréciablement plus grande que celles des modèles 1 et 3. Le modèle 3 est légèrement meilleur que le modèle 1 pour ce qui est des valeurs CPO. Il y a aussi de petites valeurs CPO pour les trois modèles, ce qui indique que les hypothèses de nos modèles ne sont peut-être pas bien respectées par notre ensemble de données.

Compte tenu des estimations de la variance d'échantillonnage que l'on trouve dans le tableau 1 et le tracé CPO, nous concluons que le modèle 2 est le meilleur modèle d'ajustement pour nos données. Par conséquent, nous avons utilisé le modèle 2 pour obtenir des estimations modélisées de moyennes régionales et des erreurs-types a posteriori connexes.

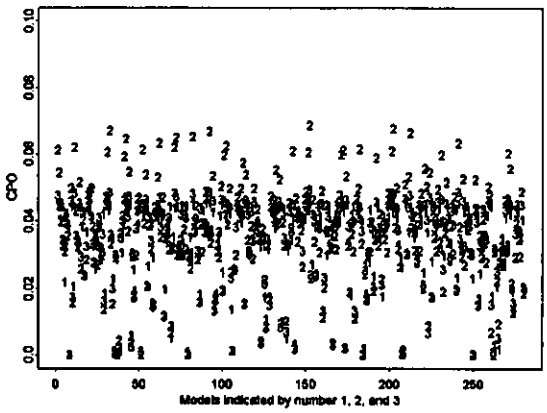


Figure 1. Sélection du modèle: tracé de comparaison CPO

3.3 Résultat de l'estimation

Nous présentons maintenant les estimations des moyennes régionales en fonction du modèle 2 seulement. Le tableau 2 montre les moyennes régionales a posteriori estimées et les erreurs-types a posteriori correspondantes. Notre étude a montré que l'estimateur empirique $\hat{\mu}_i^{(E)}$ et l'estimateur Rao-Blackwellisé $\hat{\mu}_i^{(RB)}$ fournissent presque les mêmes estimations ponctuelles, et nous avons donc signalé uniquement les estimations obtenues à l'aide de $\hat{\mu}_i^{(RB)}$. Pour la comparaison, nous avons calculé les estimations directes (moyennes d'échantillon) et les erreurs-types directes correspondantes pour les 10 régions. Le tableau 2 montre clairement que les estimations modélisées sont appréciablement plus efficaces que les estimations directes. Les erreurs-types a posteriori sont beaucoup plus petites que les erreurs-types directes.

Tableau 2
Estimations de moyennes régionales

Région	$\bar{y}_i$	é.-t.	$\hat{\mu}_i^{(RB)}$	é.-t.
1	11,08	9,53	10,23	0,81
2	7,91	6,82	9,84	0,85
3	13,48	14,15	13,01	1,08
4	6,53	8,01	10,95	1,11
5	19,52	14,96	17,87	1,57
6	11,21	11,38	10,21	0,93
7	8,72	11,24	9,58	0,97
8	12,81	13,99	10,30	1,19
9	10,18	8,76	11,34	1,01
10	10,01	11,30	9,79	0,87

Afin d'étudier les effets de la Rao-Blackwellisation, nous avons calculé les erreurs-types de simulation de $\hat{\mu}_i^{(E)}$ et $\hat{\mu}_i^{(RB)}$, qui sont respectivement les erreurs-types d'échantillon de $\{\bar{X}_i^T \hat{\beta}_i^{(k)}\}$ et $\{\bar{X}_i^T E[\beta_i | Y, \gamma^{(k)}, \Omega^{(k)}, \tau_e^{(k)}]\}$. Le tableau 3 montre les erreurs-types de simulation. Le tableau 3 indique clairement que l'estimateur Rao-Blackwellisé $\hat{\mu}_i^{(RB)}$ comporte une erreur-type de simulation beaucoup plus

faible que l'estimateur empirique $\hat{\mu}_i^{(E)}$ pour toutes les régions. Dans tous les cas, l'erreur-type de $\hat{\mu}_i^{(RB)}$ représente de 50% à 75% environ de l'erreur-type de $\hat{\mu}_i^{(E)}$, ce qui confirme l'avantage de la Rao-Blackwellisation. Ainsi, $\hat{\mu}_i^{(RB)}$ est plus stable que $\hat{\mu}_i^{(E)}$ lorsqu'on l'utilise pour obtenir des estimations ponctuelles pour les moyennes a posteriori dans une analyse bayésienne computationnelle. Il convient de mentionner que l'erreur-type de simulation de $\hat{\mu}_i^{(E)}$ est également un estimateur de l'erreur-type a posteriori. Ainsi, l'erreur-type de simulation de $\hat{\mu}_i^{(E)}$ au tableau 3 est presque identique à l'erreur-type estimée $\hat{\mu}_i^{(RB)}$ au tableau 2.

Tableau 3
Erreurs-types de simulation

Région	$\hat{\mu}_i^{(E)}$	$\hat{\mu}_i^{(RB)}$
1	0,817	0,506
2	0,862	0,498
3	1,090	0,548
4	1,101	0,604
5	1,583	0,878
6	0,930	0,481
7	0,978	0,480
8	1,208	0,842
9	0,997	0,524
10	0,869	0,513

3.4 Analyse de sensibilité

Dans le modèle 2, les variances d'erreur $\tau_i = \sigma_i^{-2}$ sont supposées indépendantes avec des distributions a priori $G(a_i, b_i)$ ou σ_i^2 avec la distribution gamma inverse $IG(a_i, b_i)$, où a_i et b_i sont des valeurs connues choisies de façon à refléter nos connaissances a priori de σ_i^2. Dans la pratique, il est toujours difficile d'obtenir des informations exactes au sujet des variances d'échantillonnage. De plus, à mesure que le nombre de régions m augmente, le nombre de composantes des variances σ_i^2 augmente. Nous nous intéressons aux effets possibles du choix des distributions a priori sur les σ_i^2; en particulier, nous aimerions évaluer la sensibilité des moyennes a posteriori au choix des distributions a priori sur les variances d'échantillonnage σ_i^2. Dans notre analyse des données, a_i et b_i ont été choisies égales à 0,001. Ainsi, nous avons utilisé des distributions a priori appropriées avec de très faibles valeurs paramétriques pour les composantes de variance de façon à refléter nos connaissances vagues de σ_i^2. Afin de vérifier la sensibilité des estimations a posteriori au choix de a_i et b_i dans le cadre du modèle 2, nous avons fixé six valeurs différentes de $a_i = b_i$ c'est-à-dire 0,0001, 0,001, 0,01, 0,1, 1 et 10. Puisque

$$\{\tau_i | Y, \beta, \gamma, \Omega\} \sim G\left(a_i + \frac{1}{2}n_i, b_i + \frac{1}{2}(Y_i - X_i\beta_i)^T (Y_i - X_i\beta_i)\right), \quad (19)$$

les effets d'échantillon $n_{i/2}$ et $(Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i)/2$ dominent l'information a priori a_i et b_i lorsque a_i et b_i sont petites. Ainsi IG(0,0001, 0,0001), IG(0,001, 0,001) et IG(0,01,0,01) peuvent être considérées comme des distributions a priori non informatives tandis que IG(1, 1) et IG(10, 10) peuvent être considérées comme des distributions a priori informatives. Le tableau 4 présente des moyennes a posteriori dans le cadre du modèle 2 à l'aide des différentes distributions a priori sur σ_i^2, et le tableau 5 montre les variances a posteriori correspondantes.

Tableau 4
Comparaison de moyennes régionales estimées

Région	$IG(a_i, b_i), a_i = b_i$					
	0,001	0,001	0,01	0,1	1	10
1	10,23	10,23	10,23	10,24	10,25	10,37
2	9,84	9,84	9,84	9,83	9,82	9,62
3	13,00	13,00	13,01	13,01	13,07	13,09
4	10,95	10,95	10,95	10,95	10,94	10,61
5	17,86	17,87	17,85	17,76	17,78	18,27
6	10,21	10,21	10,21	10,21	10,25	10,28
7	9,58	9,58	9,59	9,58	9,63	9,57
8	10,29	10,30	10,30	10,26	10,37	10,86
9	11,34	11,34	11,35	11,32	11,32	11,23
10	9,79	9,79	9,80	9,79	9,82	9,92

Le tableau 4 indique clairement que les estimations de moyennes régionales sont très stables; elles ne sont pas sensibles au choix de a_i et b_i. Toutefois, comme le montre le tableau 5, les variances a posteriori diminuent à mesure que les distributions a priori sur σ_i^2 deviennent plus informatives, et mènent à de plus petits coefficients de variation (cv), Cela indique que nous pouvons améliorer les résultats des estimations pour les régions en ce qui concerne le cv si nous avons plus de renseignements a priori sur les variances de l'erreur d'échantillonnage, Dans notre étude, nous avons considéré uniquement le cas $a_i = b_i$. Une étude plus approfondie comporterait différentes combinaisons de a_i et b_i.

Tableau 5
Comparaison de variances a posteriori de l'erreur d'échantillonnage estimée

Région	$IG(a_i, b_i), a_i = b_i$					
	0,0001	0,001	0,01	0,1	1	10
1	0,658	0,658	0,658	0,656	0,653	0,499
2	0,724	0,724	0,724	0,711	0,684	0,462
3	1,167	1,167	1,167	1,161	1,152	0,917
4	1,220	1,220	1,218	1,217	1,202	0,919
5	2,455	2,455	2,454	2,462	2,139	1,335
6	0,871	0,870	0,870	0,830	0,826	0,699
7	0,933	0,933	0,931	0,930	0,914	0,779
8	1,418	1,417	1,418	1,375	1,351	1,337
9	1,015	1,014	1,014	1,011	0,975	0,790
10	0,760	0,760	0,760	0,750	0,745	0,613

Le tableau 6 présente les estimations a posteriori de σ_i^2 à l'aide des différentes distributions a priori sur σ_i^2. Comme l'indique le tableau 6, lorsque a_i et b_i sont petites (<= 0.01), il n'existe pratiquement pas de différence entre les estimations. À mesure que a_i et b_i augmentent, les estimations $\hat{\sigma}_i^{2(RB)}$ deviennent plus petites. Toutefois, en présence de solides informations a priori sur a_i et b_i, par exemple $a_i = b_i = 10$, les estimations a posteriori de σ_i^2 sont appréciablement différentes de celles qui sont obtenues pour des distributions a priori non informatives.

Tableau 6
Comparaison de variances de l'erreur d'échantillonnage estimée

Région	$IG(a_i, b_i), a_i = b_i$					
	0,0001	0,001	0,01	0,1	1	10
1	40,09	40,10	40,05	39,64	37,14	22,29
2	34,19	34,18	34,17	33,97	31,74	19,05
3	94,48	94,49	94,42	93,76	86,73	50,60
4	52,08	52,08	52,04	51,63	48,21	28,82
5	121,60	121,70	121,60	121,40	113,70	66,75
6	94,03	94,03	93,83	92,96	87,21	52,90
7	102,30	102,30	102,20	101,40	94,85	57,58
8	160,10	160,00	159,90	159,10	147,60	86,61
9	63,46	63,46	63,38	62,99	58,46	34,85
10	65,88	65,87	65,89	65,40	60,76	36,60

4. CONCLUSION

Dans le présent exposé, nous avons présenté des méthodes bayésiennes hiérarchiques pour l'estimation régionale, à l'aide de modèles à plusieurs niveaux. Il n'est clairement pas facile de fournir un modèle approprié pour toutes les régions de façon à obtenir des résultats satisfaisants, même si des méthodes bayésiennes de type MCCM (Monte Carlo à chaîne markovienne) comme l'échantillonnage de Gibbs nous permettent d'ajuster les données à l'aide de modèles bayésiens d'une complexité virtuellement illimitée. La taille et l'homogénéité des régions et la présence de solides informations auxiliaires auront un effet sur le résultat final. Des modèles appropriés dans certaines situations pourront ne pas se prêter à d'autres situations. La méthode bayésienne hiérarchique comporte également des limitations, par exemple le choix de distributions a priori sur les paramètres de modèle et certaines questions d'échantillonnage liées à la méthode d'échantillonnage de Gibbs. Néanmoins, la méthode bayésienne hiérarchique générale s'applique à tout un choix de situations pour l'estimation de paramètres régionaux. Le choix du modèle est un aspect important de l'analyse bayésienne hiérarchique. Il est également important de comparer la méthode bayésienne hiérarchique à d'autres méthodes largement utilisées pour l'estimation régionale, par exemple la méthode bayésienne empirique et la méthode du meilleur prédicteur linéaire sans biais empirique. Les travaux se poursuivent afin d'y intégrer les poids de plan d'enquête, en conformité avec You et Rao (1999).

REMERCIEMENTS

Nous tenons à remercier deux examinateurs et le rédacteur de leurs remarques et suggestions utiles. Nous tenons également à remercier le professeur F. Moura de l'Université fédérale de Rio de Janeiro (Brésil) d'avoir fourni l'ensemble de données utilisé à la section 3. Nos travaux ont été appuyés en partie par une subvention du Conseil de recherches en sciences naturelles et en génie du Canada.

ANNEXE

A1:

L'estimateur Rao-Blackwellisé de la variance a posteriori de β_i est donné par:

$$\begin{aligned}\hat{V}(\beta_i) &= \frac{1}{G}\sum_{k=1}^{G} V(\beta_i|Y,\gamma^{(k)},\Omega^{(k)},\tau_e^{(k)}) \\ &\quad + \frac{1}{G}\sum_{k=1}^{G}[E(\beta_i|Y,\gamma^{(k)},\Omega^{(k)},\tau_e^{(k)})]^2 \\ &\quad - \left[\frac{1}{G}\sum_{k=1}^{G} E(\beta_i|Y,\gamma^{(k)},\Omega^{(k)},\tau_e^{(k)})\right]^2 \\ &= \frac{1}{G}\sum_{k=1}^{G}(\tau_e^{(k)}X_i^TX_i+\Omega^{(k)})^{-1} \\ &\quad + \frac{1}{G}\sum_{k=1}^{G}(\tau_e^{(k)}X_i^TX_i+\Omega^{(k)})^{-1}(\tau_e^{(k)}X_i^TY_i+\Omega^{(k)}Z_i\gamma^{(k)}) \\ &\quad \times(\tau_e^{(k)}X_i^TY_i+\Omega^{(k)}Z_i\gamma^{(k)})^T(\tau_e^{(k)}X_i^TX_i+\Omega^{(k)})^{-1} \\ &\quad - \frac{1}{G^2}\left[\sum_{k=1}^{G}(\tau_e^{(k)}X_i^TX_i+\Omega^{(k)})^{-1}(\tau_e^{(k)}X_i^TY_i+\Omega^{(k)}Z_i\gamma^{(k)})\right] \\ &\quad \times\left[\sum_{k=1}^{G}(\tau_e^{(k)}X_i^TX_i+\Omega^{(k)})^{-1}(\tau_e^{(k)}X_i^TY_i+\Omega^{(k)}Z_i\gamma^{(k)})\right]^T.\end{aligned}$$

A2:

Lemme: $[\eta|Y,\beta,\tau,\gamma,\Omega,\lambda]$ est une fonction concave logarithmique de η.

Preuve: Soit $h(\eta) = \log[\eta|Y,\beta,\tau,\gamma,\Omega,\lambda]$. Il est suffisant de montrer que

$$\frac{\partial^2 h(\eta)}{\partial^2\eta} \le 0.$$

Manifestement,

$$\frac{\partial h(\eta)}{\partial\eta} = -m\frac{\Gamma'(\eta)}{\Gamma(\eta)} + m\log(\lambda) + \log(\Pi_{i=1}^{m}\tau_i).$$

Soit $\psi(\eta) = \Gamma'(\eta)/\Gamma(\eta)$, nous avons alors

$$\frac{\partial^2 h(\eta)}{\partial^2 \eta} = -m\psi'(\eta) \leq 0$$

puisque $m > 0$ et $\psi'(\eta)$ est positif sur $(0, \infty)$ (Temme, 1994, 54-55).

BIBLIOGRAPHIE

BATTESE, G.E., HARTER, R.M., et FULLER, W.A. (1988). An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

BEST, N., COWLES, M.K., et VINES, K. (1996). CODA, *Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output*, Version 0.30. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR.

GELFAND, A.E. (1995). Model determination using sampling-based methods. Dans *Markov Chain Monte Carlo in Practice* (W.R. Gilks, S. Richardson, et D.J. Spiegelhalter, Éds.), 145-161. London: Chapman and Hall.

GELFAND, A.E., et SMITH, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

GELFAND, A.E., et SMITH, A.F.M. (1991). Gibbs sampling for marginal posterior expections. *Communications In Statistics - Theory and Methods*, 20, 1747-1766.

GHOSH, M., et RAO, J.N.K. (1994). Small area estimation: An appraisal (avec discussion). *Statistical Science*, 9, 55-93.

GILKS, W.R., BEST, N.G., et TAN, K.K.C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of Applied Statistics*, 44, 455-472.

HOBERT, J.P., et CASSELLA, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91, 1461-1473.

HOLT, D., et MOURA, F. (1993). Small area estimation using multi-level models. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 21-30.

KLEFFE, J., et RAO, J.N.K. (1992). Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model. *Journal of Multivariate Analysis*, 43, 1-15.

MOURA, F., et HOLT, D. (1999). Production d'estimations régionales à partir de modèles multiniveau. *Techniques d'enquête*, 25, 81-89.

PRASAD, N.G.N., et RAO, J.N.K. (1990). The estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

RAO, J.N.K. (1999).Quelques progrès récents concernant l'estimation régionale fondée sur un modèle. *Techniques d'enquête*, 25, 199-212.

SPIEGELHALTER, D., THOMAS, A., BEST, N., et GILKS, W. (1996). *BUGS 0.5, Bayesian Inference Using Gibbs Sampling Manual*. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR.

TEMME, N.M. (1994). *Special Functions: An Introduction to the Classical Functions of Mathematical Physics*. New York: John Wiley.

YOU, Y., et RAO, J.N.K. (1999). Pseudo hierarchical Bayes small area estimation using sampling weights. *Recueil de la section des méthodes d'enquête, Société Statistique du Canada*, 117-122.

Techniques d'enquête, décembre 2000
Vol. 26, n° 2, pp. 207-212
Statistique Canada

Échantillonnage à deux phases pour l'estimation par quotient ou par régression avec sous-échantillonnage des non-répondants

FABIAN C. OKAFOR et HYUNSHIK LEE[1]

RÉSUMÉ

Cochran (1977, p. 374) a proposé certains estimateurs par quotient ou par régression de la moyenne de population fondés sur la méthode de Hansen et Hurwitz (1946) consistant à sous-échantillonner les non-répondants en supposant que l'on connaît la moyenne de population de la variable auxiliaire. Le présent article décrit certains estimateurs par quotient ou par régression axés sur un échantillonnage double (à deux phases) applicables aux cas où l'on ne connaît pas la moyenne de population de la variable auxiliaire. On y compare aussi la performance de ces estimateurs à celle de l'estimateur proposé par Hansen et Hurwitz (1946).

MOTS CLÉS: Estimateur de Hansen et Hurwitz; coût d'enquête; fraction optimale d'échantillonnage.

1. INTRODUCTION

Très souvent, lors d'enquêtes visant des personnes, on ne réussit pas à recueillir les renseignements auprès de toutes les unités d'échantillonnage, même après plusieurs rappels. Or, une estimation calculée d'après des données incomplètes peut être trompeuse, surtout si les caractéristiques des répondants diffèrent de celles des non-répondants, auquel cas l'estimation risque d'être biaisée. Hansen et Hurwitz (1946) ont proposé une méthode de correction pour la non-réponse afin de résoudre le problème du biais. Cette méthode consiste à sélectionner un sous-échantillon de non-répondants afin d'obtenir une estimation pour la sous-population que ces derniers représentent.

En s'appuyant sur la méthode de Hansen et Hurwitz (1946), Cochran (1977) a proposé les estimateurs par quotient ou par régression de la moyenne de population de la variable étudiée pour lesquels les renseignements sur la variable auxiliaire proviennent de toutes les unités d'échantillonnage, alors que certaines unités n'ont pas toutes fourni les renseignements sur la variable étudiée. En outre, on connaît la moyenne de population de la variable auxiliaire. Ici, nous supposons que l'on ne connaît pas la moyenne de population de la variable auxiliaire. Par conséquent, nous utilisons la méthode d'échantillonnage à deux phases pour estimer d'abord la moyenne de la variable auxiliaire, puis la moyenne de la variable étudiée selon une méthode similaire à celle de Cochran (1977).

En pratique, on compense souvent la non-réponse par rajustement de la pondération (Oh et Scheuren 1983) ou par imputation (Kalton et Karsprzyk 1986). Les méthodes appliquées pour rajuster la pondération ou pour procéder à l'imputation visent à éliminer le biais dû à la non-réponse. Cependant, cette méthode se fonde sur des hypothèses insoutenables quant au mécanisme de réponse. Si le mécanisme hypothétique est erroné, les estimations résultantes risquent d'être fortement biaisées. De surcroît, il est difficile d'éliminer entièrement le biais s'il y a confusion de la non-réponse, en ce sens que la probabilité de réponse dépend de la variable étudiée. Rancourt, Lee et Särndal (1994) ont réussi à corriger partiellement cette situation. La méthode de sous-échantillonnage de Hansen et Hurwitz ne présente pas ce défaut, mais elle est plus coûteuse à cause du travail supplémentaire qu'exige le sous-échantillonnage des non-répondants. Néanmoins, si le biais est important, la méthode offre un moyen viable de résoudre le problème sans recourir à la réponse totale qui pourrait coûter fort cher.

À la section suivante, nous examinons les estimateurs par quotient ou par régression axés sur l'échantillonnage à deux phases. En général, on recoure à l'échantillonnage à deux phases lorsqu'il faut se servir de données auxiliaires pour améliorer la précision d'une estimation, mais que l'on ne connaît pas la loi de distribution de la population pour les variables auxiliaires. Nous nous servons de l'échantillon de la première phase pour estimer la distribution de population de la variable auxiliaire et de l'échantillon de la deuxième phase pour obtenir les renseignements nécessaires sur la variable étudiée. Nous calculons la fraction optimale d'échantillonnage pour les divers estimateurs, pour un coût prédéterminé. Enfin, nous comparons théoriquement et empiriquement les estimateurs proposés à l'estimateur de Hansen et Hurwitz.

2. ESTIMATEURS PAR QUOTIENT OU PAR RÉGRESSION AVEC ÉCHANTILLONNAGE À DEUX PHASES

2.1 Renseignements généraux

Pour estimer la moyenne de population $\bar{X}$ de la variable auxiliaire, nous sélectionnons un échantillon de première

[1] Fabian C. Okafor, Dept. of Statistics, University of Nigeria, Nsukka, Nigeria; Hyunshik Lee, anciennement Statistique Canada, maintenant Westat, 1650 Research Boulevard, Rockville, Maryland, 20850, U.S.A.

phase de grande taille n' à partir de N unités de la population par échantillonnage aléatoire simple sans remise (EASSR). Puis, nous sélectionnons par EASSR un échantillon de deuxième phase de plus petite taille n à partir de n' et nous mesurons la caractéristique y sur cet échantillon. L'estimateur par quotient de la moyenne de y est $\bar{y}'_r = (\bar{y}/\bar{x})\bar{x}'$, où $\bar{x}'$ est la moyenne d'échantillon calculée pour n' unités, et où $\bar{y}$ et $\bar{x}$ sont obtenues d'après l'échantillon de deuxième phase s'il n'y a pas de non-réponse dans cet échantillon. Cependant, en cas de non-réponse dans l'échantillon de deuxième phase, nous pouvons utiliser un estimateur fondé uniquement sur les répondants ou sélectionner un sous-échantillon de non-répondants et reprendre contact avec eux. La première option est beaucoup moins coûteuse que la seconde, parce que recueillir les renseignements manquants auprès des non-répondants en les contactant de nouveau nécessite habituellement beaucoup plus d'efforts et de dépenses. Cependant, il se pourrait fort bien qu'en ce qui concerne la caractéristique étudiée, les non-répondants diffèrent des répondants au point que les résultats soient sérieusement biaisés. Le cas échéant, le sous-échantillonnage des non-répondants pourrait être souhaitable. Par conséquent, nous poursuivons l'idée du sous-échantillonnage de Hansen et Hurwitz dans le cas d'un échantillonnage à deux phases. Fondamentalement, les estimateurs que nous proposons ici constituent une version à échantillonnage à deux phases des estimateurs proposés par Cochran (1977, p. 374), c'est-à-dire des estimateurs de $\bar{Y}$ par quotient ou par régression avec échantillonnage à deux phases, corrigés pour la non-réponse par la méthode de Hansen et Hurwitz (1946).

Supposons que les n' unités fournissent toutes des renseignements sur la variable auxiliaire x à la première étape d'échantillonnage. Mais posons que n_1 unités fournissent des renseignements sur y et que n_2 unités refusent de répondre lors de la deuxième étape. À partir des n_2 non-respondants, nous sélectionnons un échantillon aléatoire simple sans remise de m unités au taux inverse d'échantillonnage k, où $m = n_2/k$, $k > 1$. Cette fois-ci, les m unités répondent toutes. Ces conditions pourraient s'appliquer à une enquête-ménages où l'on se sert de la taille du ménage comme variable auxiliaire pour l'estimation, disons, des dépenses familiales. On pourrait obtenir des renseignements complets sur la taille de la famille durant l'établissement de la liste des ménages, mais faire face à une non-réponse en ce qui concerne les dépenses du ménage.

Dans l'exposé qui suit, nous supposons que l'ensemble de la population (représenté par A) est divisé en deux strates: une strate(représentée par A_1) de N_1 unités qui répondent lors de la première visite à la deuxième étape et une strate (représentée par A_2) de N_2 unités qui ne répondent pas lors de la première visite à la deuxième étape d'échantillonnage, mais qui répondent lors de la deuxième visite. Représentons les échantillons de premier et de deuxième phase par a' et a , respectivement, et posons que $a_1 = a \cap A_1$ et $a_2 = a \cap A_2$. Le sous-échantillon de a_2 sera représenté par a_{2m}. La somme sur l'ensemble des unités est un ensemble s que nous représentons par Σ_s.

En règle générale, les paramètres de population sont représentés par des lettres majuscules, sauf les lettres grecques, et les statistiques d'échantillon, par les lettres minuscules correspondantes.

2.2 Estimateur par quotient avec échantillonnage à deux phases

Nous définissons l'estimateur par quotient avec échantillonnage à deux phases comme suit:

$$d^* = \frac{\bar{y}^*}{\bar{x}^*}\bar{x}' = r^*\bar{x}' \tag{2.1}$$

où $\bar{x}^*$ et $\bar{y}^*$ sont, respectivement, les estimateurs de $\bar{X}$ et $\bar{Y}$ de Hansen-Hurwitz qui sont donnés par

$$\bar{u}^* = w_1\bar{u}_1 + w_2\bar{u}_{2m}, \quad u = x, y. \tag{2.2}$$

Conformément à la règle générale, nous définissons $W_j = N_j/N$ et $w_j = n_j/n$, $j = 1$ ou 2. Les statistiques d'échantillon calculées d'après a_{2m} sont marquées de l'indice «$2m$», (par exemple, $\bar{u}_{2m} = (1/m)\Sigma_{a_{2m}} u_i$); celles calculées d'après a_1 sont marquées de l'indice «1», (par exemple, $\bar{u}_1 = (1/n_1)\Sigma_{a_1} u_i$), et celles calculées d'après l'échantillon de première phase a' sont marqués d'un signe prime (par exemple, $\bar{x}' = (1/n')\Sigma_{a'} x_i$).

Pour un grand échantillon, l'approximation de premier ordre de la variance de d^*, calculée selon la méthode de linéarisation par série de Taylor, est donnée par

$$V(d^*) \cong \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)S_r^2 + \frac{W_2(k-1)}{n}S_{2r}^2 \tag{2.3}$$

où

$$S_r^2 = S_y^2 + R^2S_x^2 - 2RS_{xy},$$

$$S_{2r}^2 = S_{2y}^2 + R^2S_{2x}^2 - 2RS_{2xy}, \tag{2.4}$$

R représente le rapport de population de $\bar{Y}$ à $\bar{X}$. S_u^2 et S_{2u}^2 représentent, respectivement, la variance de la variable u pour l'ensemble de la population et pour la strate de non-répondants. S_{xy} et S_{2xy} représentent, respectivement, la covariance pour l'ensemble de la population et pour la population de non-répondants.

Nous pouvons estimer la variance approximative de d^* par

$$v(d^*) = \left(\frac{1}{n'} - \frac{1}{N}\right)\hat{S}_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)\hat{S}_r^2 + \frac{w_2(k-1)}{n}\hat{S}_{2r}^2 \tag{2.5}$$

où

$$\hat{S}_y^2 = \frac{1}{n-1}\left\{\sum_{a_1} y_i^2 + k\sum_{a_{2m}} y_i^2 - n\bar{y}^{*2} + w_2(k-1)s_{2my}^2\right\},$$

$$\hat{S}_r^2 = \frac{1}{n-1}\left\{\sum_{a_1} (y_i - r^* x_i)^2 + k\sum_{a_{2m}} (y_i - r^* x_i)^2\right\} \text{ et}$$

$$\hat{S}_{2r}^2 = \frac{1}{m-1}\sum_{a_{2m}} (y_i - r^* x_i)^2. \qquad (2.6)$$

Il convient de souligner que $\hat{S}_y^2$ est un estimateur non biaisé de S_y^2. Il semble naturel d'utiliser $\hat{S}_r^2$ pour estimer S_r^2 puisque l'expression obtenue d'après $\hat{S}_r^2$ en remplaçant r^* par R est un estimateur convergent de S_r^2. Nous pouvons nous servir du même argument pour justifier l'emploi de $\hat{S}_{2r}^2$.

Nous obtenons un autre estimateur de $V(d^*)$ en remplaçant $\hat{S}_r^2$ par $\hat{S}_{2r}^2$, c'est-à-dire:

$$\tilde{S}_r^2 = \hat{S}_y^2 + r^{*2} s_x'^2 - 2r^* s_{xy}^* \text{ et}$$

$$\tilde{S}_{2r}^2 = s_{2my}^2 + r^{*2} s_{2x}^2 - 2r^* s_{2mxy}, \qquad (2.7)$$

respectivement, dans (2.5), où

$$s_x'^2 = \frac{1}{n'-1}\sum_{a'} (x_i - \bar{x}')^2,$$

$$s_{2my}^2 = \frac{1}{m-1}\sum_{a_{2m}} (y_i - \bar{y}_{2m})^2,$$

$$s_{2x}^2 = \frac{1}{n_2-1}\sum_{a_2} (x_i - \bar{x}_2)^2,$$

$$s_{2mxy} = \frac{1}{m-1}\left(\sum_{a_{2m}} x_i y_i - m\bar{x}_{2m}\bar{y}_{2m}\right)$$

et où s_{xy}^* a la même forme que dans (2.9). La variance de cet estimateur de rechange sera vraisemblablement plus faible que celle de l'estimateur (2.5), puisque les estimateurs $s_x'^2$ et s_{2x}^2 sont fondés sur des échantillons plus grands et sont par conséquent plus précis.

2.3 Estimateur par régression avec échantillonnage à deux phases

Nous définissons l'estimateur par régression comme suit:

$$t^* = \bar{y}^* + \hat{\beta}^* (\bar{x}' - \bar{x}^*) \qquad (2.8)$$

où $\hat{\beta}^*$ est un estimateur de $\beta = S_{xy}/S_x^2$. Il pourrait exister plusieurs solutions pour $\hat{\beta}^*$, mais un choix naturel semble être $\hat{\beta}^* = s_{xy}^*/s_x^{*2}$, où

$$s_{xy}^* = \frac{1}{n-1}\left(\sum_{a_1} x_i y_i + k\sum_{a_{2m}} x_i y_i - n\bar{x}\bar{y}^*\right) \text{ et}$$

$$s_x^{*2} = \frac{1}{n-1}\left(\sum_{a_1} x_i^2 + k\sum_{a_{2m}} x_i^2 - n\bar{x}\bar{x}^*\right). \qquad (2.9)$$

Il est facile de montrer que s_{xy}^* et s_x^{*2} sont des estimateurs non biaisés de S_{xy} et S_x^2, respectivement. La variance approximative de t^* est donnée par

$$V(t^*) = \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)S_l^2 + \frac{W_2(k-1)}{n}S_{2l}^2 \qquad (2.10)$$

où S_l^2 et S_{2l}^2 sont calculés d'après (2.4) en remplaçant R par β.

Pour estimer $V(t^*)$, nous pouvons appliquer la formule suivante:

$$v(t^*) = \left(\frac{1}{n'} - \frac{1}{N}\right)\hat{S}_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)\hat{S}_l^2 + \frac{w_2(k-1)}{n}\hat{S}_{2l}^2 \qquad (2.11)$$

où

$$\hat{S}_l^2 = \frac{1}{n-1}\left\{\sum_{a_1} (y_i - y_i^*)^2 + k\sum_{a_{2m}} (y_i - y_i^*)^2\right\},$$

$$\hat{S}_{2l}^2 = \frac{1}{m-1}\sum_{a_{2m}} (y_i - y_i^*)^2, \text{ et}$$

$$y_i^* = \bar{y}_i^* - \hat{\beta}^* (x_i - \bar{x}^*). \qquad (2.12)$$

Comme pour (2.7), nous obtenons un estimateur un peu meilleur de $V(t^*)$ en utilisant:

$$\tilde{S}_l^2 = \hat{S}_y^2 + \hat{\beta}^{*2} s_x'^2 - 2\hat{\beta}^* s_{xy}^* \text{ et}$$

$$\tilde{S}_{2l} = s_{2my}^2 + \hat{\beta}^{*2} s_{2x}^2 - 2\hat{\beta}^* s_{2mxy}. \qquad (2.13)$$

3. CHOIX DES FRACTIONS D'ÉCHANTILLONNAGE

Nous allons maintenant déterminer les valeurs optimales de k, n et n' qui réduisent au minimum la variance des estimateurs proposés pour un coût particulier ou qui réduisent au minimum le coût pour une variance particulière.

Considérons pour d^* la fonction de coût donnée par

$$C = c'n' + cn + c_1 n_1 + c_2 m \tag{3.1}$$

où les c représentent les coûts unitaires définis comme suit:

c' : coût unitaire associé à l'échantillon de première phase, a';

c : coût unitaire associé au premier effort de collecte de renseignements sur y auprès de l'échantillon de deuxième phase, a;

c_1 : coût unitaire du traitement des renseignements sur y fournis par les répondants lors du premier effort de collecte auprès de a_1;

c_2 : coût unitaire associé au sous-échantillon a_{2m} de a_2.

Puisqu'on ne connaît pas la valeur de n_1 tant qu'on n'a pas fait la première collecte de données, on se sert du coût prévu pour la minimisation. Le coût prévu est donné par

$$E(C) = C^* = c'n' + \left(c + c_1 W_1 + \frac{c_2 W_2}{k} \right) n. \tag{3.2}$$

Nous nous servons du multiplicateur de Lagrange pour recalculer les valeurs optimales de k, n et n' qui réduisent au minimum la variance de d^* pour un coût prévu fixe C^*. Les valeurs optimales ainsi obtenues sont:

$$k_o = \sqrt{\frac{c_2\left(S_r^2 - W_2 S_{2r}^2\right)}{S_{2r}^2 (c + c_1 W_1)}},$$

$$n_o = \frac{C^* \sqrt{A}}{D\sqrt{G}} \text{ et } n_o' = \frac{C^* \sqrt{S_y^2 - S_r^2}}{D\sqrt{c'}} \tag{3.3}$$

où

$$A = S_r^2 + W_2 (k_o - 1) S_{2r}^2,$$

$$G = c + c_1 W_1 + \frac{c_2 W_2}{k_o} \text{ et}$$

$$D = \sqrt{(S_y^2 - S_r^2) c'} + \sqrt{AG}.$$

Si nous supposons que $\gamma = c_2/(c + c_1 W_1)$, $\delta = S_r^2/S_{2r}^2$ et $\xi = S_y^2/S_r^2$, alors nous obtenons

$$k_o = \sqrt{\gamma(\delta - W_2)},$$

$$n_o = \frac{C^* \sqrt{1 + W_2(k_o - 1)/\delta}}{\sqrt{Gc'(\xi - 1)} + G\sqrt{1 + W_2(k_o - 1)/\delta}} \text{ et}$$

$$n_o' = \frac{C^* \sqrt{\xi - 1}}{c'\sqrt{\xi - 1} + \sqrt{Gc'\{1 + W_2(k_o - 1)/\delta\}}}. \tag{3.4}$$

Les valeurs optimales de n_o et n_o' sont proportionnelles au coût prévu, C^*. Pour obtenir les valeurs optimales de k, n et n' qui réduisent au minimum la valeur de $V(t^*)$, nous remplaçons simplement dans l'expression (3.3) susmentionnée S_r^2 et S_{2r}^2 par S_l^2 et S_{2l}^2, respectivement. Le tableau 1 montre les valeurs optimales de k_o, n_o et n_o' pour des paramètres donnés.

Tableau 1
Valeur optimale de k_o, n_o, and n_o'

C^*	c'	c	c_1	c_2	δ	ξ	W_2	γ	k_o	G	n_o	n_o'
200	0,1	0,5	1	2	1	2	0,3	1,67	1,08	1,76	92	382
200	0,1	0,5	1	2	1	4	0,3	1,67	1,08	1,76	81	580
200	0,1	0,5	1	2	2	2	0,3	1,67	1,68	1,56	104	389
200	0,1	0,5	1	2	2	4	0,3	1,67	1,68	1,56	91	590
200	0,1	0,5	1	4	1	2	0,3	3,33	1,52	1,99	83	345
200	0,1	0,5	1	4	1	4	0,3	3,33	1,52	1,99	74	531
200	0,1	0,5	1	4	2	2	0,3	3,33	2,38	1,70	96	361
200	0,1	0,5	1	4	2	4	0,3	3,33	2,38	1,70	85	553
200	0,5	0,5	1	2	1	2	0,3	1,67	1,08	1,76	85	250
200	0,5	0,5	1	2	1	4	0,3	1,67	1,08	1,76	72	366
200	0,5	0,5	1	2	2	2	0,3	167	1,68	1,56	96	255
200	0,5	0,5	1	2	2	4	0,3	167	1,68	1,56	80	372
200	0,5	0,5	1	4	1	2	0,3	3,33	1,52	1,99	78	228
200	0,5	0,5	1	4	1	4	0,3	3,33	1,52	1,99	67	338
200	0,5	0,5	1	4	2	2	0,3	3,33	2,38	1,70	89	238
200	0,5	0,5	1	4	2	4	0,3	3,33	2,38	1,70	76	351

4. COMPARAISON DES ESTIMATEURS

À la présente section, nous comparons théoriquement la performance des estimateurs proposés à celle des estimateurs de Hansen et Hurwitz (1946), d'abord sans tenir compte du coût, puis en tenant compte de celui-ci.

4.1 Sans tenir compte du coût

La variance de l'estimateur de Hansen-Hurwitz est donnée par

$$V(\bar{y}^*) = \left(\frac{1}{n} - \frac{1}{N} \right) S_y^2 + \frac{W_2(k - 1)}{n} S_{2y}^2 \tag{4.1}$$

où $\bar{y}^*$ est défini comme dans (2.2).

$$V(\bar{y}^*) - V(d^*) = \left(\frac{1}{n} - \frac{1}{n'}\right)\left(2RS_{xy} - R^2S_x^2\right) + \frac{W_2(k-1)}{n}\left(2RS_{2xy} - R^2S_{2x}^2\right). \quad (4.2)$$

La différence est positive (autrement dit, d^* est plus efficace que $\bar{y}^*$) si $R < 2\beta$ et $R < 2\beta_2$, où $\beta_2 = S_{2xy}/S_{2x}^2$. Par ailleurs, nous avons

$$V(\bar{y}^*) - V(t^*) = \left(\frac{1}{n} - \frac{1}{n'}\right)\frac{S_{xy}^2}{S_x^2} + \frac{W_2(k-1)}{n}\beta S_{2x}^2(2\beta_2 - \beta). \quad (4.3)$$

Par conséquent, t^* est plus efficace que l'estimateur de Hansen-Hurwitz si la différence (4.3) est positive. Cette condition s'observe notamment si $\beta_2 \geq \beta/2$, avec $\beta \geq 0$. Comme les conditions dont nous discutons ici sont suffisantes, d^* ou t^* peut être plus efficace que $\bar{y}^*$ dans des conditions moins rigoureuses.

4.2 En tenant compte du coût

Nous allons maintenant comparer les estimateurs proposés à l'estimateur de Hansen-Hurwitz ($\bar{y}^*$) en nous servant de la fonction de coût donnée à la section 3.

Pour l'estimateur $\bar{y}^*$, si nous sélectionnons un échantillon aléatoire simple (sans recourir à l'échantillonnage à deux phases) pour y, nous pouvons calculer la taille optimale d'échantillon pour un coût prévu,

$$C^* = \left(c + c_1W_1 + \frac{c_2W_2}{k}\right)n,$$

semblable à celui donné par (3.2) selon la même technique (c'est-à-dire, le multiplicateur de Lagrange) que celle utilisée à la section 3, comme suit:

$$n_{oHH} = \frac{C^*}{c + c_1W_1 + c_2W_2/k_{oHH}} \text{ et}$$

$$k_{oHH} = \sqrt{\frac{c_2(S_y^2 - W_2S_{2y}^2)}{S_{2y}^2(c + c_1W_1)}}. \quad (4.4)$$

Alors, la variance optimale de l'estimateur de Hansen-Hurwitz devient

$$V(\bar{y}^*) = \left(\frac{1}{n_{oHH}} - \frac{1}{N}\right)S_y^2 + \frac{W_2(k_{oHH} - 1)}{n_{oHH}}S_{2y}^2. \quad (4.5)$$

Si nous comparons cette expression à celle de $V(d^*)$ avec les valeurs optimales de k, n et n' données par (3.3), alors la condition voulant que d^* soit plus précis que $\bar{y}^*$ s'écrit

$$2\rho - Rh > \frac{1}{1-\theta_1} \times \left\{\frac{1}{\beta h}(1 - \theta_2 + Q_y - \theta_2Q_{HHy}) - hQ_x(2\beta_2 - R)\right\} \quad (4.6)$$

où

$$h = \frac{S_x}{S_y}, \theta_1 = \frac{n_o}{n_o'}, \theta_2 = \frac{n_o}{n_{oHH}}, Q_{HHy} = \frac{W_2(k_{oHH} - 1)S_{2y}^2}{S_y^2},$$

$$Q_u = \frac{W_2(k_o - 1)S_{2u}^2}{S_u^2}, u = x, y,$$

et où ρ représente le coefficient de corrélation de x et y.

Nous pouvons procéder à une comparaison similaire entre $\bar{y}^*$ et t^*. Autrement dit, t^* est plus efficace que $\bar{y}^*$ si

$$2\rho - \beta h > \frac{1}{1-\theta_1} \times \left\{\frac{1}{\beta h}(1 - \theta_2 + Q_y - \theta_2Q_{HHy}) - hQ_x(2\beta_2 - \beta)\right\}. \quad (4.7)$$

4.3 Comparaison empirique des estimateurs proposés

Nous nous servons d'une population générée artificiellement pour comparer l'efficacité relative des estimateurs d^* et t^* par rapport à $\bar{y}^*$. Les paramètres de la population sont les suivants:

$$R = 1{,}92, \beta = 1{,}52, \rho = 0{,}85, R_2 = 1{,}88, \beta_2 = 1{,}47,$$

$$\rho_2 = 0{,}83, N = 1000, N_2 = 302, S_x^2 = 766{,}54,$$

$$S_y^2 = 2426{,}82, S_{xy} = 1164{,}08, S_{2x}^2 = 433{,}63,$$

$$S_{2y}^2 = 1350{,}05 \text{ et } S_{2xy} = 638{,}32.$$

Les efficacités relatives de d^* et t^* sont présentées au tableau 2. Notons que R diffère nettement de β, autrement

dit que la droite de régression ne passe pas l'origine. Dans le cas de cette population, l'estimateur par régression t^* est plus efficace que l'estimateur par quotient d^*. Nous constatons aussi que la taille initiale optimale d'échantillon, n_o', est plus grande pour t^* que pour l'estimateur d^*. Pour la taille optimale d'échantillon de deuxième phase, n_o, nous observons l'inverse, parce qu'on peut obtenir un estimateur par régression plus précis avec un plus petit échantillon de deuxième phase, si bien qu'il est possible d'allouer davantage à l'échantillon de première phase. Enfin, le taux d'échantillonnage inverse optimal k_o est pratiquement le même pour les deux estimateurs.

Si la droite de régression passe par l'origine, l'avantage de t^* sur d^* disparaît, comme le prédit et le confirme une autre comparaison empirique que nous ne présentons pas ici.

Tableau 2
Efficacité relative de d^* et t^* par rapport à $\bar{y}^*$ ($C^* = 200$, $c = 0{,}5$, $c_1 = 1$)

c'	c_2	k_{oHH}	n_{oHH}	k_o	n_o	n_o'	Efficacité
				d^*			
0.1	2	1,58	127	1,46	92	514	1,85
0.1	4	2,23	115	2,06	85	477	1,91
0.3	2	1,58	127	1,46	78	250	1,23
0.3	4	2,23	115	2,06	73	234	1,32
				t^*			
0.1	2	1,58	127	1,47	89	563	2,11
0.1	4	2,23	115	2,08	83	523	2,19
0.3	2	1,58	127	1,47	74	269	1,34
0.3	4	2,23	115	2,08	70	253	1,45

5. CONCLUSIONS

Nous proposons des estimateurs par quotient ou par régression fondés sur une méthode d'échantillonnage à deux phases pour tenir compte de la non-réponse au sujet de la variable principale quand on ne connaît pas la moyenne de population de la variable auxiliaire. Nous éliminons le biais éventuellement important dû à la non-réponse par sous-échantillonnage des non-répondants, conformément à la méthode de Hansen et Hurwitz (1946). Nous calculons les tailles optimales d'échantillon pour un ensemble donné de coûts unitaires, puis nous comparons théoriquement et empiriquement la performance de nos estimateurs à celle de l'estimateur de Hansen et Hurwitz.

S'il existe une relation linéaire prononcée entre la variable principale et la variable auxiliaire et que l'on peut recueillir à faible coût les données sur la variable auxiliaire auprès d'un échantillon de grande taille, nos estimateurs donnent de nettement meilleurs résultats que l'estimateur de Hansen et Hurwitz. Notre méthode pourrait être utile s'il existe un biais important dû à la non-réponse que l'on ne peut corriger par rajustement de la pondération ni par imputation.

REMERCIEMENTS

Nous remercions les examinateurs et les rédacteurs en chef adjoints de leurs commentaires qui nous ont aidés à améliorer notre article.

BIBLIOGRAPHIE

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

HANSEN, M. H., et HURWITZ, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

KALTON, G., et KASPRZYK, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.

OH, H.L., et SCHEUREN, F.J. (1983). Chapter 13. Weighting adjustment for unit non-response. Dans *Incomplete Data in Sample Surveys*, (I. Olkin, W.G. Madow, et D.B. Rubin, Éds.). Theory and Bibliographies, 2, 143-184. New York: Academic Press.

RANCOURT, E., LEE, H., et SÄRNDAL, C.-E. (1994). Corrections du biais pour des estimations d'enquête tirées de données comprenant des valeurs imputées par quotient par suite d'une non-réponse selon un mécanisme avec confusion. *Techniques d'enquête*, 20, 143-153.

Techniques d'enquête, décembre 2000
Vol. 26, n° 2, pp. 213-224
Statistique Canada

Modélisation des effets d'intervieweur dans le cas des enquêtes par panel: Une application

JAN PICKERY et GEERT LOOSVELDT[1]

RÉSUMÉ

Le présent article décrit la combinaison de deux applications de modèles multiniveaux. Le modèle multiniveaux convient bien à l'analyse des effets d'intervieweur sur les données d'enquête. On peut aussi l'utiliser pour analyser des données longitudinales, c'est-à-dire des «mesures répétées». Nous analysons un indicateur de la qualité des données que nous appliquons à des données de panel recueillies dans le cadre des enquêtes post-électorales auprès de la population belge. Ces données de panel comprennent les données de deux cycles d'enquête seulement. La plupart des répondants qui ont participé aux deux cycles n'ont pas été interviewés par la même personne les deux fois. Par conséquent, il en résulte une structure complexe où les mesures sont emboîtées dans les répondants et les répondants sont emboîtés dans les intervieweurs, sans toutefois présenter une structure hiérarchique globale: la classification croisée. Nous procédons à deux analyses distinctes de cette structure complexe des données, l'une portant sur tous les répondants et l'autre uniquement sur ceux qui ont été interviewés deux fois par la même personne. Puis, nous comparons les résultats de ces deux analyses. Nous concluons que le modèle multiniveaux à classification croisée est un outil très souple et fort utile pour analyser les effets d'intervieweur en cas d'enquête par panel.

MOTS CLÉS: Classification croisée; effets d'intervieweur; enquêtes par panel; modèles multiniveaux; réponse «Ne sait pas».

1. INTRODUCTION

Dans le présent article, nous analysons l'effet des caractéristiques du répondant et de l'intervieweur sur le nombre de réponses «Ne sait pas» recueillies lors de deux cycles de l'enquête par panel menée dans le cadre des enquêtes post-électorales auprès de la population belge. Nous utilisons divers modèles multiniveaux que nous appliquons à un sous-ensemble de l'ensemble de données, d'une part, et à l'ensemble de données complet, d'autre part. Le but principal de l'article est d'illustrer l'utilisation de modèles multiniveaux pour analyser les effets d'intervieweur dans le cas d'une enquête par panel.

Le modèle multiniveaux ou hiérarchique est un outil approprié à l'analyse des données emboîtées, c'est-à-dire à plusieurs degrés, comme les élèves groupés dans des classes ou les patients groupés dans des hôpitaux. Un modèle multiniveaux peut inclure les variables qui correspondent aux divers niveaux de groupement, mais aussi tenir compte de la variabilité associée à chaque niveau. La qualité type de ces modèles tient non pas à la forme fonctionnelle établissant le lien entre les variables des divers niveaux, mais plutôt au traitement plus complexe de la structure de l'erreur (DiPrete et Forristal 1994, p. 334). Par exemple, dans le cas de la recherche en éducation, un modèle multiniveaux permet de rendre compte de la variation entre écoles et de la variation entre élèves. De surcroît, on s'efforce de remplacer dans le modèle cette variance attribuable aux deux niveaux par des variables liées à chaque niveau. Ces modèles sont décrits dans divers manuels, comme Bryk et Raudenbush (1992), Goldstein (1995), Kreft et de Leeuw (1998) et Snijders et Bosker (1999).

Les modèles multiniveaux ou hiérarchiques sont aussi ceux qui conviennent le mieux à l'analyse des effets d'intervieweur sur les données d'enquête (Hox 1994). Le modèle hiérarchique est l'outil le plus approprié en cas de plan de sondage avec «répondants nichés dans les intervieweurs». L'application d'autres méthodes statistiques exige que les caractéristiques des intervieweurs et des répondants soit mutuellement indépendantes, condition qui, la plupart du temps, n'est pas satisfaite à cause de la structure hiérarchique des données. Dans le cas d'un modèle multiniveaux, aussi bien les coefficients de régression que les composantes de la variance sont subordonnés aux variables explicatives du modèle, propriété utile si l'orthogonalisation des variables caractérisant les intervieweurs et les répondants est incomplète (Hox 1994, p. 307). Si les répondants ne sont pas affectés au hasard aux intervieweurs, les caractéristiques des répondants et des intervieweurs peuvent éventuellement se confondre puisque les répondants d'une région particulière seront vraisemblablement interviewés par des intervieweurs de cette région. Le cas échéant, si les variables pertinentes caractérisant les répondants sont introduites dans un modèle multiniveaux, l'égalisation des variables relatives aux intervieweurs se fait par des moyens statistiques. Par conséquent, les hypothèses qui sous-tendent l'analyse des effets d'intervieweur au moyen d'un modèle multiniveaux sont plus réalistes que celles formulées dans le cas d'un analyse de la variance ou d'un analyse de la covariance. De surcroît, le modèle

[1] Jan Pickery et Geert Loosveldt, Department of Sociology, University of Leuven, E. Van Evenstraat 2B, 3000 Leuven, Belgium. courriel : jan.pickery@soc. kuleuven.ac.be; geert.loosveldt@soc.kuleuven.ac.be.

hiérarchique permet d'estimer la variance liée à l'intervieweur, ainsi que les effets des variables explicatives mesurées au niveau de l'intervieweur et du répondant. La possibilité de remplacer la variance attribuée au répondant et à l'intervieweur par les effets des caractéristiques du répondant et de l'intervieweur permet de procéder à des généralisations de plus grande portée.

L'utilisation du modèle multiniveaux est également fructueuse lors de l'analyse de données longitudinales, c'est-à-dire de «mesures répétées» (voir, par exemple, Goldstein 1995, p. 87-95; Snijders 1996; Yang et Goldstein 1996). Il existe d'autres moyens d'analyser le plan de sondage avec «mesures emboîtées dans les unités», mais l'analyse multiniveaux présente certains avantages manifestes. Le modèle hiérarchique peut s'appliquer aux plans de sondage non équilibrés, c'est-à-dire ceux où le nombre de mesures n'est pas le même pour toutes les unités observées, et permet d'intégrer assez facilement les covariables qui évoluent. Qui plus est, le modèle permet de tenir compte d'un plus grand nombre de niveaux hiérarchiques. Les unités observées peuvent être regroupées dans une autre unité de niveau supérieur.

Nous analyserons les effets de répondant et d'intervieweur sur le nombre de réponses «Ne sait pas» à une série de questions sur les partis politiques posées lors d'une enquête par panel. Les mesures (cycle 1 et cycle 2) sont emboîtées dans les répondants (plan de sondage longitudinal) et les répondants sont emboîtés dans les intervieweurs. Nos données de panel englobent celles de deux cycles d'enquête. Durant le deuxième cycle, la plupart des répondants n'ont pas été interviewés par le même intervieweur que la première fois, situation qui interrompt le groupement purement hiérarchique pour donner lieu à une structure plus complexe des données. Pour tenir compte de cette structure, il faut concevoir les mesures comme étant groupées dans deux structures de classification distinctes, à savoir les répondants et les intervieweurs. Nous parlerons ici de plan de sondage à classification croisée, car le groupement par niveau n'est pas purement hiérarchique.

Dans le présent article, nous partons d'une structure simple de données et du modèle approprié. Puis, le modèle devient plus complexe. Nous réalisons deux analyses. Dans la première, nous nous attachons uniquement aux répondants qui sont interviewés deux fois par le même intervieweur. Ensuite, nous nous intéressons à l'ensemble des répondants, y compris ceux qui n'ont été interviewés qu'une seule fois. Lors de la première analyse, la structure purement hiérarchique reste intacte. Le modèle est un modèle «simple» à trois niveaux: mesures emboîtées dans les répondants qui sont eux-mêmes emboîtés dans les intervieweurs. Dans le cas de la deuxième analyse, nous établissons un modèle à classification croisée. Dans ce modèle, les mesures sont classées selon le répondant et selon l'intervieweur.

À la section suivante, nous décrivons la nature de notre variable dépendante, c'est-à-dire la réponse «Ne sait pas», et la façon de l'analyser. À la troisième section, nous décrivons les données en détail, afin de préciser leur structure complexe. La quatrième section traite brièvement des divers modèles que nous combinons. À la cinquième section, nous présentons les variables de l'analyse. Aux sixième et septième sections, nous discutons de l'élaboration des deux modèles que nous utilisons et nous présentons les résultats de l'analyse. Enfin, à la huitième section, nous formulons les conclusions de l'article.

2. RÉPONSE «NE SAIT PAS»

Il est généralement reconnu que l'utilisation d'un filtre «Ne sait pas» ou «Sans opinion» augmente la proportion de répondants qui donnent cette réponse et que l'augmentation proprement dite est fonction de la nature du filtre utilisé (Schuman et Presser 1981, p. 143). Krosnick soutient que répondre «Ne sait pas» est une forme de solution de facilité. Cette situation se présente si le répondant n'est pas motivé à déployer l'effort mental nécessaire pour donner la réponse optimale. La réponse «Sans opinion» est une réponse acceptable, mais elle est le résultat d'un processus cognitif «faible». Choisir la solution de facilité est fonction de la difficulté de la tâche, ainsi que du niveau de connaissances, de compétences et de motivation du répondant. Ce raisonnement théorique corrobore l'observation selon laquelle donner la possibilité de répondre «Ne sait pas» fait augmenter la proportion de répondants qui choisissent cette réponse, particulièrement chez les personnes peu instruites et celles qui accordent peu d'intérêt au problème en question (Krosnick 1991). Selon cette thèse, les caractéristiques des répondants que l'on peut associer à l'aspect cognitif de l'acte consistant à répondre aux questions explique en grande partie le choix de la réponse «Ne sait pas». Selon des études antérieures, les caractéristiques pertinentes sont le niveau de scolarité (par exemple, Sudman et Bradburn 1974), l'âge (voir, par exemple, Groves 1989, p. 441-443), le sexe (par exemple, Hox, de Leeuw et Kreft 1991) et l'intérêt pour le sujet (par exemple, Groves 1989, p. 419).

Cependant, répondre à une question comprend non seulement le processus cognitif du répondant, mais aussi un processus de communication (Schwarz et Sudman 1995). Or, l'intervieweur joue un rôle important dans ce second processus. Nombreux sont les articles qui traitent de l'intervieweur en tant que source d'erreur de mesure dans les enquêtes (Groves 1989). Le thème principal de ces études est qu'au lieu d'être des collecteurs «neutres» de données, les intervieweurs influencent les réponses que donnent les personnes qu'ils interrogent. La non-réponse à une question est, elle aussi, sujette à l'effet d'intervieweur, comme l'ont montré il y a déjà longtemps, par exemple, Hanson et Marks (1958), ainsi que Bailar, Bailey et Stevens (1977). Par conséquent, un spécialiste des sciences sociales qui cherche à expliquer les réponses «Ne sait pas» doit inclure dans l'analyse les répondants ainsi que les intervieweurs. Le

nombre de réponses «Ne sait pas» est la variable dépendante de nos analyses.

3. DESCRIPTION DE LA STRUCTURE DES DONNÉES

Après les élections générales de 1991 en Belgique, on a réalisé une enquête nationale comptant 4 544 interviews directes réalisées dans les trois régions du pays au cours des premiers mois de 1992. Le plan de sondage est un plan à deux degrés autopondéré (voir, par exemple, Särndal, Swensson et Wretman 1992, p. 141-144). L'échantillon est représentatif de la population de 18 à 74 ans (ISPO/PIOP 1995). Dans le présent article, nous utilisons les données recueillies dans la région flamande, auprès de 2 691 répondants flamands, interviewés par 163 intervieweurs (Carton, Swyngedouw, Billiet et Beerten 1993). Après les élections de 1995, on a réalisé une enquête post-électorale comparable. Étant donné les contraintes budgétaires, il a fallu sélectionner un échantillon plus petit pour le deuxième cycle. Donc, on a commencé par sélectionner un échantillon à partir du groupe de 2 691 répondants, puis on a sélectionné de nouveaux répondants pour compenser le vieillissement de la cohorte la plus jeune sélectionnée en 1991. En dernière analyse, 2 099 répondants ont été interviewés par 167 intervieweurs. Cet échantillon contient 1 762 répondants appartenant au panel et 337 nouveaux répondants (voir Beerten, Billiet, Carton et Swyngedouw, 1997, pour un rapport technique détaillé sur le plan de sondage). En tout, seuls 55 intervieweurs qui avaient participé au premier cycle ont de nouveau participé à l'enquête. Par conséquent, 112 nouveaux intervieweurs ont participé au deuxième cycle.

Nous obtenons ainsi un ensemble de données correspondant à 3 028 répondants (2 691 + 337) et à 275 intervieweurs (163+112). Pour 1762 répondants, nous disposons de mesures faites lors des deux cycles, et pour le reste (1 266), d'une seule mesure. On peut représenter la structure de l'ensemble de données dans un tableau comparable au tableau 1 (voir aussi Goldstein 1995, p. 114). Chaque x du tableau représente une observation. L'ensemble complet de données contient 4 790 observations ((1 762 × 2) + 1 266). Chaque catégorie de répondants qui figure dans le tableau représente une occurrence éventuelle dans l'ensemble de données.

Le tableau montre qu'il existe trois groupes de répondants: les répondants membres du panel qui ont été interviewés deux fois par le même intervieweur (catégorie 1), les répondants qui ont participé deux fois à l'enquête, mais qui ont été interviewés par deux intervieweurs différents (catégorie 2) et les répondants qui n'ont été interviewés qu'une seule fois (catégories 3 ou 4). Nos deux analyses se fondent sur ces différentes catégories de répondants. Dans le cadre de la première, nous examinons les répondants dont la situation correspond à celle de la catégorie 1. Cette dernière ne compte que 394 répondants qui ont été interviewés deux fois par le même intervieweur. La deuxième analyse porte sur l'ensemble des 3 028 répondants (répondants des catégories 1 à 4 du tableau).

Tableau 1
Une représentation de l'ensemble de données

	Catégorie 1 de répondants		Catégorie 2 de répondants		Catégorie 3 de répondants	Catégorie 4 de répondants	
Cycle	1	2	1	2	1	2	N (Inter-vieweurs)
Catégorie A d'intervieweurs	x	x	x			x	47
Catégorie B d'intervieweurs				x	x		8
Catégorie C d'intervieweurs			x		x		108
Catégorie D d'intervieweurs				x		x	112
N (Répondants)	374		1388		929	337	

Le tableau montre aussi que nous pouvons faire la distinction entre les intervieweurs, à savoir ceux qui ont collaboré deux fois à l'enquête (catégories A et B) et ceux qui n'ont collaboré qu'au premier cycle (catégorie C) ou qu'au deuxième cycle (catégorie D). Les intervieweurs de la catégorie B ont collaboré aux deux cycles, mais n'ont jamais interviewé un même répondant deux fois (contrairement à ceux de la catégorie A).

Pour analyser ces données de structure complexe, nous combinons trois modèles distincts que nous présentons à la section qui suit.

4. BRÈVE DESCRIPTION DES DIFFÉRENTS MODÈLES MULTINIVEAUX UTILISÉS POUR LES ANALYSES

4.1 Modèle multiniveaux général

Le premier modèle que nous utilisons est le modèle multiniveaux général, dont la forme est:

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + e_{ij}, \qquad (1)$$

$$\beta_{0j} = \beta_0 + u_{0j} \text{ et } \beta_{1j} = \beta_1 + u_{1j} \qquad (2)$$

ou

$$\beta_{0j} = \beta_0 + \gamma_{01} z_{1j} + u_{0j} \text{ et } \beta_{1j} = \beta_1 + \gamma_{11} z_{1j} + u_{1j}. \qquad (3)$$

L'indice i désigne l'unité de niveau 1 et l'indice j, l'unité de niveau 2. Ici, le niveau 1 correspond au répondant et le niveau 2, à l'intervieweur. Donc, la variable de réponse Y du répondant i, interviewé par l'intervieweur j, dépend de la variable x de ce répondant. Cette relation ressemble à un modèle de régression ordinaire, mais les paramètres sont propres à l'intervieweur. Les coefficients β varient d'un

intervieweur à l'autre. Pour chaque valeur de β, il existe un résidu lié à l'intervieweur (u_{0j} ou u_{1j}). On peut aussi faire dépendre les coefficients β de variables de plus haut niveau (caractéristiques de l'intervieweur) pour permettre les généralisations sur l'ensemble des intervieweurs. Nous avons une variable de deuxième niveau z. En introduisant (3) par substitution dans (1), nous obtenons le modèle global suivant:

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \gamma_{01} z_{1j} + \gamma_{11} z_{1j} x_{1ij} + u_{1j} x_{1ij} + u_{0j} + e_{ij}. \quad (4)$$

Naturellement, on pourrait inclure un plus grand nombre de variables x et z dans ces équations. Nous supposons que les résidus u_{0j}, u_{1j} et e_{ij} ont une moyenne nulle, étant donné les valeurs des variables explicatives z et x. De surcroît, nous supposons que les résidus de niveau 1 (e_{ij}) sont indépendants. Nous supposons en outre que les résidus de niveau 2 (u_{0j} et u_{1j}) sont indépendants de e_{ij} et qu'ils suivent une loi normale multivariée commune avec matrice de covariance Ω. Ces résidus ne doivent pas être indépendants l'un de l'autre. Habituellement, ils sont corrélés.

4.2 Modèle multiniveaux pour l'analyse longitudinale

Le deuxième modèle que nous utilisons est le modèle multiniveaux longitudinal. Lors de l'analyse de «mesures répétées» au moyen d'un modèle hiérarchique, on considère que les mesures représentent le premier niveau et les unités mesurées, le deuxième. La plupart du temps les unités individuelles sont des personnes, mais il pourrait naturellement s'agir d'autres unités, comme des écoles ou des pays. Dans le cas qui nous occupe, les unités sont des répondants. L'analyse a pour but d'estimer une courbe de croissance d'après les diverses mesures et d'étudier la variation de la courbe selon les caractéristiques individuelles. Chaque valeur observée est subordonnée au moment de la mesure, lequel peut coïncider avec une mesure de temps ou avec l'âge, ainsi qu'aux transformations éventuelles de cette mesure. Habituellement, on suppose que la courbe a une forme polynomiale dont l'équation est la suivante:

$$Y_{it} = \pi_{0i} + \pi_{1i} t + \pi_{2i} t^2 + \ldots + \pi_{ki} t^k + e_{it}. \quad (5)$$

Y_{it} représente la valeur observée pour le répondant i au moment t, t pouvant être le moment où est faite la mesure ou l'âge. Les π_{hi} ($h = 0 \ldots k$) sont les paramètres de trajectoire ou paramètres de croissance pour le sujet i, et k représente le degré du polynôme. Dans un cas simple, k est égal à 1 et la courbe est linéaire. S'il existe m moments de mesure, un polynôme de degré $m - 1$ représentera exactement la courbe. Évidemment, il est plus intéressant d'utiliser un polynôme de faible degré si celui-ci donne une représentation satisfaisante de la courbe. On peut faire l'essai pour voir si le modèle de degré $k + 1$ donne de nettement meilleurs résultats que le modèle de degré k.

On donne aussi aux paramètres de croissance un indice qui représente l'unité (répondant). Le modèle précise que ces paramètres diffèrent d'une unité à l'autre. La deuxième partie du modèle définit ces paramètres:

$$\pi_{0i} = \pi_0 + r_{0i} \quad (6)$$

ou

$$\pi_{0i} = \pi_0 + \beta_{01} x_{1i} + r_{0i}. \quad (7)$$

Les paramètres individuels sont équivalents à un paramètre général (π_0) + un résidu individuel (r_{0i}). Grâce à l'inclusion de caractéristiques individuelles (x), il est possible de réduire la partie individuelle spécifique, donc de généraliser sur l'ensemble des répondants. Conformément à (1) et à (2), nous choisissons x pour représenter les caractéristiques individuelles (répondant). Il convient toutefois de noter que, dans ce modèle, les variables x sont d'un niveau plus élevé (niveau 2). Les caractéristiques individuelles peuvent être constantes (les mêmes pour tous les moments de mesure) ou variables.

4.3 Modèles à classification croisée

Le troisième modèle que nous utilisons est le modèle à classification croisée. Les données n'ont pas toute une structure purement hiérarchique. Certaines unités peuvent être classées en fonction de plus d'une variable (voir Goldstein 1995, p. 113-116). Par exemple, des élèves peuvent être classés selon l'école qu'ils fréquentent ou selon le quartier où ils vivent. Dans notre exemple, les mesures sont classées selon le répondant et selon l'intervieweur. Un modèle à classification croisée prend la forme qui suit (les indices j_1 et j_2 désignent les deux structures de classification distinctes):

$$Y_{ij_1j_2} = \beta_{0j_1j_2} + \beta_{1j_1j_2} x_{1ij_1j_2} + e_{ij_1j_2} \quad (8)$$

$$\beta_{0j_1j_2} = \beta_0 + u_{0j_1} + u_{0j_2} \text{ et } \beta_{1j_1j_2} = \beta_{1j_1j_2} + u_{1j_1} + u_{1j_2}. \quad (9)$$

L'équation (9) peut être reformulée de la même façon que l'équation (3).

$Y_{ij_1j_2}$ représente la valeur observée pour l'unité i, classifiée selon j_1 et j_2, c'est-à-dire, ici, la valeur de la mesure i faite sur le répondant j_1, interviewé par l'intervieweur j_2. Les paramètres associés à la variable indépendante x ont un résidu pour chaque structure de classification. Pour ce modèle, on pose en outre que les résidus obtenus pour des structures de classification différentes (ici les résidus liés au répondant et à l'intervieweur) sont mutuellement indépendants (u_{0j_1} et u_{1j_1} contre u_{0j_2} et u_{1j_2}).

Raudenbush (1993) examine ce genre de modèles et l'utilisation de l'algorithme EM pour les estimer. Rasbash et Goldstein (1994) et Goldstein (1995, p. 123-124) montrent comment spécifier et estimer ces modèles au

moyen d'une formulation purement hiérarchique et (conséquemment) d'un logiciel multiniveaux type. La méthode consiste à établir les spécifications de l'une des classifications conformément à une classification hiérarchique type, puis à définir une variable fictive pour chaque unité de l'autre classification, à préciser que chaque variable fictive possède un coefficient aléatoire au niveau supérieur et à poser la condition que les ensembles de variances résultants doivent être égaux.

Aux sections 6 et 7, nous nous servons de ces trois modèles, qui peuvent tous être appliqués au moyen du logiciel de modélisation multiniveaux *MLn/MLwiN*. Au préalable, nous examinons plus en détail les variables que nous utilisons dans l'analyse.

5. VARIABLES ANALYSÉES

Durant l'interview de l'enquête post-électorale, l'une des tâches les plus difficiles consistaient à évaluer six partis politiques au moyen de diverses échelles à 11 points. Trois échelles ont été présentées au répondant, pour évaluer, respectivement, le catholicisme, le libéralisme économique et le fédéralisme. La question comportait un filtre explicite (Ne sait pas), mais celui-ci n'était pas mentionné sur la carte décrivant les divers choix donnés au répondant. L'énoncé complet de la question figure en annexe. Nous nous attendions à recueillir un nombre considérable de réponses «Ne sait pas», à cause de la complexité de la tâche. Nous nous attendions aussi à ce que le filtre explicite fasse augmenter ce nombre (voir, par exemple Schuman et Presser 1981).

Lors du premier cycle, le nombre moyen de réponses «Ne sait pas» était supérieur à quatre par répondant. Presque 20 % de répondants ont choisi cette option au moins neuf fois sur 18. Si nous considérons uniquement les répondants membres du panel, le nombre moyen est légèrement inférieur (3,8). Ces résultats ne sont pas surprenants, puisque l'on pouvait s'attendre à ce que les «utilisateurs multiples» de la réponse «Ne sait pas» soient sous-représentés lors du deuxième cycle, à cause du manque d'intérêt pour le sujet de l'enquête et(ou) de la difficulté des questions. Au deuxième cycle, le nombre global moyen de réponses «Ne sait pas» était de 3,6 et la moyenne pour les répondants membres du panel se chiffrait à 3,4. Pour les répondants qui ont été interviewés deux fois par le même intervieweur, le nombre moyen est de 3,9 et de 4,2, respectivement. Rien ne permet d'expliquer pourquoi le nombre de réponses «Ne sait pas» enregistré lors du deuxième cycle pour ces répondants est plus élevé que le nombre moyen observé pour l'ensemble des répondants.

Au niveau des mesures, nous utilisons l'année de l'interview comme indicateur du moment de la mesure. Nous avons recodé cette variable de sorte que le temps soit égal à zéro pour le premier cycle et égal à trois pour le deuxième.

Au niveau des répondants, nous considérons trois variables indépendantes: le sexe (0 = homme, 1 = femme), le niveau de scolarité (0 = faible 1 = élevé) et la mesure dans laquelle les répondants suivent les nouvelles politiques dans la presse (presse: 1 = (presque) toujours - 5 = jamais). Les deux premières variables sont constantes pour les deux moments de mesure. La troisième est une covariable qui varie en fonction du temps; qui plus est l'énoncé de la question a été légèrement modifié pour le deuxième cycle. Les deux variantes de la question sur la presse figurent également en annexe. Cette dissimilarité de l'énoncé des questions complique encore davantage l'établissement du modèle. Le moyen de traiter ce genre de variable consiste à la normaliser (moyenne 0, variance 1) pour chaque moment de mesure, puis de lui attribuer la valeur 0 au moment de la mesure si la question n'a pas été posée. La valeur de référence pour ces deux variables est leur moyenne (voir Snijders 1996, p. 422). Nous obtenons ainsi deux variables, à savoir presse1 pour le premier cycle et presse2, pour le deuxième. La première est nulle pour tous les répondants dans le cas de la deuxième mesure et la seconde, dans le cas de la première mesure. Nous n'introduisons pas l'âge du répondant dans le modèle, car cette variable serait corrélée trop fortement au moment de la mesure au niveau du cycle d'enquête.

Pour éviter de trop compliquer l'analyse, nous n'introduisons pas de variables caractéristiques de l'intervieweur. Nous supposons simplement qu'il existe un effet d'intervieweur, sans essayer d'expliquer cet effet en fonction des caractéristiques de l'intervieweur.

6. PREMIÈRE ANALYSE: CATÉGORIE 1 DE RÉPONDANTS DU TABLEAU 1

La première analyse ne porte que sur les répondants qui ont été interviewés deux fois par le même intervieweur (c'est-à-dire la catégorie 1 du tableau 1). Cette analyse se fait au moyen d'un modèle «simple» à trois niveaux: mesures emboîtées dans les répondants, eux-mêmes emboîtés dans les intervieweurs. La structure hiérarchique est sans ambiguïté. Ce modèle est semblable à l'exemple du chapitre 8 du traité de Bryk et Raudenbush (1992). Dans cet exemple, les auteurs analysent l'évolution du rendement scolaire des élèves dans les écoles.

Notre variable dépendante est le nombre de réponses «Ne sait pas» pour le répondant i au moment t, interviewé par l'intervieweur j (Y_{tij}). Comme notre analyse ne porte que sur deux mesures, le degré du polynôme ne peut être supérieur à 1. Nous obtenons ainsi l'équation de niveau 1 suivante:

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}\,\text{ANNÉE} + e_{tij}.$$

Notre variable de temps (t), qui est l'année où a eu lieu l'interview, peut prendre la valeur 0 (1992) ou 3 (1995). Nous appliquons un test pour déterminer si π_{1ij} est un terme significatif. S'il ne l'est pas, le modèle devient un modèle nul ou «modèle naïf» (voir Snijders 1996, 411) selon lequel le nombre de réponses «Ne sait pas» ne varie pas au cours du temps et nous pouvons alors considérer les deux mesures comme des tests répétés d'une même valeur constante. Les coefficients de l'équation de niveau 1 sont propres au répondant et à l'intervieweur.

Au niveau des répondants, nous incluons trois variables, à savoir le sexe, le niveau de scolarité et les deux variables de consultation de la presse. Donc, l'équation de niveau 2 contient quatre variables:

$$\pi_{0ij} = \pi_{0j} + \beta_{01j}\,\text{SEXE}_i + \beta_{02j}\,\text{SCOLARITÉ}_i + \beta_{03j}\,\text{PRESSE1}_i + \beta_{04j}\,\text{PRESSE2}_i + r_{0ij}.$$

Si le paramètre estimatif associé à l'année est significatif, nous aurons une équation similaire pour π_{1ij}.

Au troisième niveau (intervieweurs), nous n'incluons aucune variable supplémentaire, mais nous ajustons une coordonnée à l'origine aléatoire et des pentes aléatoires. Nous obtenons ainsi l'équation de niveau 3 suivante:

$$\pi_{0j} = \pi_0 + u_{0j} \text{ et } \beta_{01j} = \beta_{01} + u_{01j}, \ldots$$

Si nous entrons ces spécifications du modèle dans le programme *MLn*, nous obtenons les résultat qui suivent.

Le modèle a du tableau 2 est le modèle nul. Ce modèle ne contient des variables indépendantes ni au niveau des mesures ni à celui des répondants. Dans le cas de ce modèle, le nombre de réponses «Ne sait pas» est constant. Toutefois, la variance de la variable dépendante comprend une composante liée aux mesures, une composante liée au répondant et une composante liée à l'intervieweur. Toutes les variances sont significatives. Autrement dit, une variation s'est produite d'un cycle à l'autre, certains répondants ont utilisé la réponse «Ne sait pas» plus que d'autres et certains intervieweurs ont obtenu plus de réponses «Ne sait pas» que d'autres.

L'inclusion de la variable Année ne produit pas un modèle mieux ajusté. La diminution de l'écart à la moyenne (-2 Log L) n'est pas significative et le paramètre de la variable n'est pas significatif non plus (modèle b). Nous pouvons conclure qu'il n'y a eu aucun changement global significatif du nombre de réponses «Ne sait pas». Nous pouvons donc poursuivre avec un modèle sans variable de temps.

Tableau 2

Analyse des répondants qui ont été interviewés deux fois par le même intervieweur (é.-t. entre parenthèses)

	Modèle a	Modèle b	Modèle c	Modèle d
Fixe				
Niveau des mesures				
constante	4,136 (0,322)	4,028 (0,358)	3 ,749 (0,442)	3,754 (0,523)
année		0,072 (0,089)		
Niveau des répondants				
sexe			2 ,393 (0,434)	2,458 (0,414)
niveau de scolarité			-1,675 (0,425)	-1,778 (0,446)
presse1			0,911 (0,263)	0,887 (0,233)
presse2			1,483 (0,236)	1,426 (0,234)
Aléatoire				
Niveau des intervieweurs				
$\sigma^2_{\text{constante}}$	2,249 (1,040)	2,251 (1,043)	2,666 (0,969)	6,090 (2,109)
$\sigma_{\text{scolarité/constante}}$				-4,099 (1,816)
$\sigma^2_{\text{scolarité}}$				1,396 (1,819)
Niveau des répondants				
$\sigma^2_{\text{constante}}$	14,470 (1,714)	14,480 (1,714)	8,939 (1,308)	8,692 (1,332)
Niveau des mesures				
σ^2_e	13,320 (0,974)	13,300 (0,974)	13,270 (0,969)	13,250 (0,969)
-2 LL	4519,35	4518,62	4414,52	4395,68
Δ df *		1	4	6

Nota: * comparativement au modèle a

Les variables caractéristiques du répondant améliorent considérablement l'ajustement du modèle. La diminution de la valeur de -2 Log L est importante et nettement significative ($p<0,001$). Selon l'analyse, les femmes choisissent l'option «Ne sait pas» plus fréquemment que les hommes et les répondants très instruits, moins fréquemment que ceux dont le niveau de scolarité est faible. Suivre les nouvelles politiques dans la presse réduit les chances de répondre «Ne sait pas». Les effets de presse1 et de presse2 sont tous deux significatifs (modèle c). L'inclusion des variables caractérisant les répondants produit aussi une diminution importante de la variance au niveau des répondants.

Nous avons également essayé d'ajuster des pentes aléatoires au niveau des intervieweurs (modèle b). Notre analyse montre une certaine variation du paramètre associé au niveau de scolarité des répondants. Cette variable indépendante est la seule dont le coefficient est variable au troisième niveau. La valeur de $\sigma^2_{\text{scolarité}}$ n'est pas significative, mais la covariance entre le résidu pour la constante et le résidu pour le niveau de scolarité est importante ($\sigma_{\text{scolarité/constante}} = -4,099$). La covariance est négative, autrement dit les intervieweurs pour lesquels la valeur de la constante est élevée ont un coefficient plus faible pour le niveau de scolarité. Puisque le paramètre fixé pour le niveau de scolarité est négatif, il le sera encore plus pour ces intervieweurs et aura donc une valeur absolue plus grande. Par conséquent, pour les intervieweurs qui recueillent un plus grand nombre de réponses «Ne sait pas», l'écart entre les répondants les moins instruits et les plus instruits sera plus important. Dans le modèle d, la valeur de $\sigma^2_{\text{constante}}$ au niveau de l'intervieweur augmente considérablement comparativement au modèle c. Dans ce modèle, la variance au niveau de l'intervieweur dépend de la valeur de la variable explicative Niveau de scolarité et sera plus importante si cette valeur est nulle. Ceci représente une autre interprétation du modèle d: la variance liée à l'intervieweur est beaucoup plus importante pour les répondants peu instruits que pour les répondants très instruits. Ce modèle, dont la structure de la variance au niveau 3 est plus complexe, est mieux ajusté que les précédents.

L'introduction de la variable ANNÉE dans le modèle c ou dans le modèle d, n'a aucun effet significatif non plus. En outre, nos modèles finals ne donnent aucune preuve d'une évolution du nombre de réponses «Ne sait pas» d'un cycle à l'autre. Tous les modèles prouvent qu'il existe un effet d'intervieweur important. Toutefois, la grandeur relative de la variance montre que la variation est plus forte entre répondants qu'entre intervieweurs.

7. DEUXIÈME ANALYSE: ENSEMBLE DES RÉPONDANTS

Cette analyse porte sur l'ensemble des répondants: les membres du panel qui ont été interviewés deux fois par le même intervieweur, les autres membres du panel et les répondants qui n'ont été interviewés qu'une seule fois. Dans cette deuxième analyse, la structure hiérarchique est rompue. Les mesures sont encore emboîtées dans les répondants et ceux-ci sont encore emboîtés dans les intervieweurs. Par contre, il n'existe aucune structure hiérarchique globale, puisque l'intervieweur peut changer (et, dans la plupart des cas, change) d'un cycle à l'autre (voir section 3). La variable dépendante demeure le nombre de réponses «Ne sait pas» que donnent le répondant i interviewé par l'intervieweur j au moment t (Y_{tij}). Cependant, le modèle change. L'équation de niveau 1 reste la même:

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}\,\text{ANNÉE} + e_{tij}.$$

Ici, nous utilisons π comme notation, puisque le modèle de niveau 1 représente aussi une courbe de croissance. Cependant, cette équation correspond au niveau 1 du modèle à classification croisée (équation (8), section 4.3). En outre, nous continuons d'utiliser l'indice i pour le répondant et l'indice j pour l'intervieweur. Mais il est important de souligner qu'il ne s'agit pas du même modèle que celui utilisé pour la première analyse. Ces indices correspondent aux indices j_1 et j_2 des équations (8) et (9).

Il n'existe aucun troisième niveau «réel». Pour ajuster le modèle à classification croisée dans le programme *MLn*, nous devons définir un troisième niveau; mais, conceptuellement, le répondant et l'intervieweur se situent au même niveau dans ce modèle. Nous obtenons ainsi l'équation de niveau 2 suivante:

$$\pi_{0ij} = \pi_0 + \beta_{01}\,\text{SEXE}_i + \beta_{02}\,\text{SCOLARITÉ}_i + \beta_{03}\,\text{PRESSE1}_i + \beta_{04}\,\text{PRESSE2}_i + r_{0i} + r_{0j}.$$

La partie propre à l'intervieweur (r_{0j}) est incluse dans le deuxième niveau, si bien qu'il n'y a aucune interaction entre la variance liée à l'intervieweur et les variables qui caractérisent le répondant. Cette différence par rapport à la première analyse est la plus importante.

Le modèle à classification croisée demande une capacité de traitement informatique énorme. Nous avons 3 026 répondants et 275 intervieweurs. Nous devons donc créer 275 variables fictives dont tous les coefficients varient au troisième niveau fictif. À l'heure actuelle, il est impossible d'ajuster un tel modèle. Le chiffrier demande beaucoup trop de mémoire (voir Goldstein 1995, p. 118 et Rasbash et Woodhouse 1996, 85-86 pour plus de précision). Il est possible de réduire la mémoire nécessaire et d'améliorer la vitesse d'estimation du modèle en séparant l'ensemble de données en sous-ensembles pour lesquels la classification croisée produit un nombre «nettement» moins élevé de cellules. Ici, nous avons recherché des groupes distincts de mesures, classifiées selon un moins grand nombre de répondants et d'intervieweurs. L'analyse d'un seul groupe de 1 000 mesures classifiées selon 500 répondants et 100

intervieweurs demande plus de calculs que l'analyse d'un ensemble de données contenant 10 groupes de 100 mesures classifiées selon 50 répondants et 10 intervieweurs. Parfois, il est utile d'omettre certaines observations (combinaisons de mesures à des répondants et à des intervieweurs que l'on n'observe pratiquement jamais) pour rendre la séparation plus efficace.

Le logiciel *MLn/MLwiN* offre certaines procédures (grâce aux commandes XSEArch et BXSEarch) conçues pour procéder à cette séparation (Rasbash et Woodhouse 1996, 89-93). Nous avons utilisé la commande BXSEarch qui lance une procédure améliorée visant à produire la séparation maximale en réduisant au minimum la suppression des données: au départ, nous avions 4 790 mesures, 3 026 répondants et 275 intervieweurs. Après omission des observations indiquées par la commande BXSEarch, il nous restait 4 597 mesures sur 3 026 répondants interviewés par 275 intervieweurs. Aucune unité de niveau supérieur (ni répondants ni intervieweurs) n'est laissée de côté. La procédure n'a donné que sept sous-ensembles pour lesquels la classification croisée des répondants et des intervieweurs a donné, au plus 44 cellules. Appliqué de cette façon, le modèle a convergé suffisamment rapidement.

Les résultats de l'analyse sont présentés au tableau 3.

Tableau 3
Analyse l'ensemble des répondants
(é.-t. entre parenthèses)

	modèle a	modèle b	modèle c
Fixe			
Niveau des mesures			
constante	3,894 (0,136)	3,967 (0,155)	3,864 (0,165)
année		-0,053 (0,055)	
Niveau des répondants			
sexe			1,808 (0,153)
scolarité			-1,914 (0,148)
presse1			1,185 (0,090)
presse2			1,197 (0,102)
Aléatoire			
Niveau 2			
Intervieweurs			
$\sigma^2_{constante}$	2,777 (0,373)	2,716 (0,368)	2,844 (0,363)
Répondants			
$\sigma^2_{constante}$	11,810 (0,635)	11,800 (0,635)	7,017 (0,527)
σ^2_e	13,130 (0,475)	13,150 (0,476)	13,460 (0,480)
-2 LL	27717,1	27716,3	27042,1
Δ df*			

Nota: * comparativement au modèle a

Ce tableau ressemble fort au tableau 2, mais présente une différence importante. À la partie aléatoire, nous avons noté le niveau 2, qui englobe les intervieweurs et les répondants, pour indiquer clairement que les intervieweurs ne représentent pas un troisième niveau dans l'analyse.

Le modèle a est le modèle nul: il ne contient pas de variables explicatives, mais la variance de la variable dépendante comprend une composante liée aux mesures, une composante liée au répondant et une composante liée à l'intervieweur. La variance liée à l'intervieweur est significative. Donc, ce modèle nous donne de nouveau la preuve qu'il existe un effet d'intervieweur. Néanmoins, il faut interpréter avec prudence l'importance relative des variances, lorsqu'une classification compte un nombre nettement moins grand d'unités que l'autre (Goldstein 1995, p. 117-118). Il n'est pas tout à fait correct de dire que la variation entre répondants est cinq fois plus grande que la variation entre intervieweurs, mais, une fois de plus, nous constatons que la variabilité entre répondants est nettement plus forte qu'entre intervieweurs.

Dans le modèle suivant (modèle b), nous avons inclus la variable de temps (ANNÉE). De nouveau, l'effet de cette variable n'est pas significatif et son ajout dans le modèle n'améliore pas l'ajustement de celui-ci. Nous pouvons donc conclure que, dans l'ensemble, le nombre de réponses «Ne sait pas» n'évolue pas de façon significative au cours du temps.

Le modèle c est le modèle qui contient les variables caractérisant les répondants. Elles ont toutes un effet significatif et ce modèle est nettement mieux ajusté que les précédents. Pour ce qui est du fond, l'interprétation des paramètres est la même que pour la première analyse. Les femmes recourent à la réponse «Ne sait pas» plus souvent que les hommes et le nombre de ces réponses est plus faible si le niveau de scolarité est élevé. La mesure dans laquelle les répondants suivent les nouvelles politiques dans la presse est également un prédicteur de l'utilisation de la réponse «Ne sait pas». Le nombre de réponses «Ne sait pas» est d'autant plus élevé que l'intérêt pour la politique est faible.

8. CONCLUSION ET DISCUSSION

Les conclusions générales de l'étude portent sur la méthodologie ainsi que sur le fond.

Notre analyse corrobore les résultats d'études antérieures concernant l'utilisation de la réponse «Ne sait pas». Le choix de cette réponse est lié au niveau de scolarité et au sexe du répondant, ainsi qu'à son intérêt pour le sujet. De surcroît, le choix varie vraisemblablement selon l'intervieweur. Toutes nos analyses témoignent d'un effet d'intervieweur important. Nous ne constatons aucune évolution significative du recours à la réponse «Ne sait pas» d'un cycle à l'autre de l'enquête. L'effet d'intervieweur prouve que le choix de la réponse «Ne sait pas» ne résulte pas simplement du processus de réponse du répondant et indique qu'il faut donner une formation aux intervieweurs, y compris des instructions sur la façon de poser les questions difficiles et de traiter les réponses «Ne sait pas».

Comme dans la plupart des enquêtes par panel, la non-réponse lors du deuxième cycle de l'enquête n'est pas entièrement aléatoire. Elle dépend des modalités de logement du répondant, de son intérêt pour la politique et de quelques variables sociodémographiques (Loosveldt et Carton 1997). Cet abandon sélectif limite la généralisabilité des résultats en ce qui concerne l'évolution de la variable dépendante, mais nos analyses n'indiquent toutefois aucune évolution générale de l'utilisation de la réponse «Ne sait pas». Il n'est pas improbable non plus que la non-réponse sélective lors du deuxième cycle ait un effet sur la grandeur de l'effet d'intervieweur car, comme le montre la première analyse, les caractéristiques du répondant peuvent interagir avec l'effet d'intervieweur. Il est toutefois peu probable que ces interactions influencent les conclusions de fond concernant l'effet d'intervieweur. Étant donné les résultats de la première analyse et les conclusions de l'article de Loosveldt et Carton, on pourrait s'attendre à ce que l'effet d'intervieweur qui se dégage de la deuxième analyse et, par conséquent, l'effet global d'intervieweur, soit quelque peu sous-estimé. Loosveldt et Carton (1997, p. 1021) montrent que les répondants les moins instruits sont plus susceptibles de se retirer de l'enquête que ceux dont le niveau de scolarité est plus élevé et la première analyse montre que la variance liée à l'intervieweur est plus forte pour les répondants dont le niveau de scolarité est faible.

Du point de vue méthodologique, les conclusions tiennent compte de l'utilisation de divers modèles pour analyser l'effet d'intervieweur dans le cadre des enquêtes par panel. L'étude présentée ici montre que l'on peut analyser les données à structure compliquée résultant de plans de sondage assez complexes en spécifiant le modèle multiniveaux approprié. Le premier modèle (première analyse) ne s'adapte qu'à un petit nombre de cas. Il n'est pas si courant, ni toujours possible, d'affecter le même intervieweur au même répondant pour divers cycles d'une enquête par panel.

Le deuxième modèle (deuxième analyse) est un outil approprié, mais qui exige une capacité de traitement énorme. Le logiciel *MLn* est assez puissant et permet de diminuer la quantité de mémoire requise, au prix d'une légère perte d'information. Toutefois, le deuxième modèle a lui aussi ses limites. Il ne permet pas de modéliser les interactions entre les variables qui caractérisent le répondant et la variance liée à l'intervieweur, comme nous l'avons fait lors de la première analyse, ni celles entre les variables caractérisant le répondant et l'intervieweur. Cependant, l'analyse montre que ce modèle pourrait être un outil fort utile et fort souple. Le modèle à classification croisée convient également si le nombre de mesures augmente. Une enquête par panel comptant trois ou quatre cycles, ou même davantage, où l'on retient certains intervieweurs et où l'on recrute de nouveaux intervieweurs lors de chaque cycle nécessiterait exactement la même analyse. Le modèle multiniveaux permet aussi de traiter les répondants pour lesquels une ou plusieurs mesures manquent, comme le montre notre analyse. La souplesse de ce modèle plus que compense l'impossibilité d'y inclure les interactions entre répondant et intervieweur. On pourrait inclure ces interactions si l'on analysait séparément chaque cycle de l'enquête par panel. Toutefois, ces analyses ne permettraient pas de modéliser une évolution éventuelle de la variable dépendante, un autre avantage important de l'analyse conjointe de tous les cycles de l'enquête par panel.

REMERCIEMENTS

Nous remercions le centre de recherches électorales ISPO-PIOP de nous avoir fourni les données. Jacques Billiet, Marc Swyngedouw, Ann Carton et Roeland Beerten ont recueilli les données originales sur la région flamande. L'ISPO-PIOP est financé par les Services fédéraux des affaires scientifiques, techniques et culturelles. Ni les personnes qui ont recueilli les données au départ ni le centre de recherche ne sont responsables des analyses ou des interprétations présentées ici. Nous remercions également Jon Rasbash (Institute of Education, University of London) pour certains commentaires fort utiles concernant l'exploitation du logiciel *MLn*. Enfin, nous remercions les examinateurs pour leur suggestions et commentaires constructifs concernant une version antérieure de l'article.

ANNEXE 1

La question était la suivante: «On dit que les partis politiques sont «catholiques» ou «non catholiques». Veuillez placer la carte de chaque parti sur la case de la carte numéro 20 qui correspond le mieux à la mesure dans laquelle le parti en question est «catholique» ou «non catholique». Si, selon vous, deux ou plusieurs partis sont aussi «catholiques» ou aussi «non catholiques», placez les cartes sur la même case. Si vous ne savez pas dans quelle mesure un parti est «catholique» ou «non catholique», mettez simplement la carte de côté.» Avec la carte:

Catholique 0 1 2 3 4 5 6 7 8 9 10 non catholique

La question sur le suivi de la politique dans la presse n'était pas la même pour les deux cycles de l'enquête. Pour le premier cycle, la question concernant le suivi dans la presse était: «À quelle fréquence lisez-vous les nouvelles politiques dans les journaux?» Les catégories de réponses étaient les suivantes:

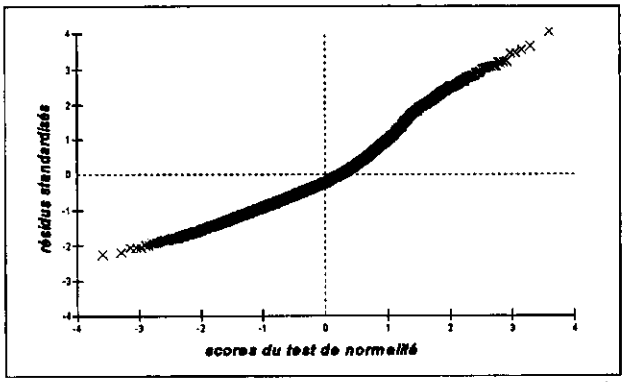
1=(presque) toujours; 2=souvent, 3=de temps en temps; 4=rarement, 5=jamais.

Lors du deuxième cycle, la question est devenue: «À quelle fréquence suivez-vous les nouvelles politiques à la radio, à la télévision ou dans les journaux?».

Les catégories de réponses sont restées les mêmes.

ANNEXE 2

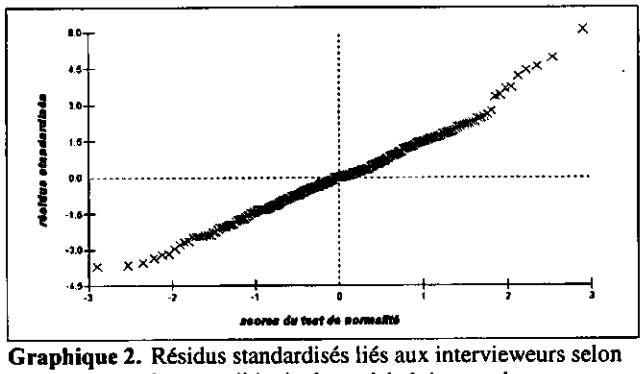
À la quatrième section, nous décrivons les hypothèses sur lesquelles s'appuient les divers modèles utilisés. Pour le dernier, les hypothèses les plus importantes ont trait aux effets aléatoires associés au répondant et à l'intervieweur. On peut évaluer l'hypothèse selon laquelle les valeurs de $\sigma^2_{\text{constante}}$ obtenues pour le répondant et pour l'intervieweur suivent la loi normale en examinant le graphique des probabilités normales pour les résidus. Le graphique 1 présente la courbe obtenue pour les résidus standardisés liés aux répondants et le graphique 2, pour les résidus standardisés liés aux intervieweurs.



Graphique 1. Résidus standardisés liés aux répondants selon le score d'équivalence à la loi normale

Dans ce graphique, les écarts par rapport à la diagonale sont assez limités et ne permettent d'inférer aucune violation de la loi normale. Par ailleurs, il convient de souligner que le nombre d'observations est plus élevé dans le cadran supérieur droit que dans le cadran inférieur gauche.

Sur le graphique 2, on n'observe pas non plus d'écart net par rapport à la diagonale Toutefois, dans ce graphique, certaines valeurs extrêmes attirent l'attention, particulièrement celles situées dans le cadran supérieur droit, qui semblent tomber en-dehors de la fourchette des autres résidus liés aux intervieweurs. De surcroît, le graphique contient lui aussi un plus grand nombre d'observations dans le cadran supérieur droit que dans le cadran inférieur gauche.



Graphique 2. Résidus standardisés liés aux intervieweurs selon le score d'équivalence à la loi normale.

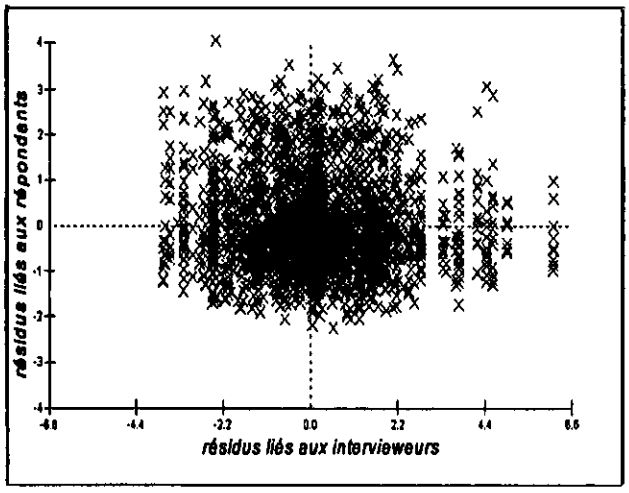
Les conclusions que l'on peut tirer de ces graphiques sont les suivantes: les résidus ne présentent aucune anomalie manifeste, mais il pourrait être utile de poursuivre l'étude pour déterminer la raison des déviations plus nombreuses vers le haut et des valeurs extrêmes pour les résidus relatifs à l'intervieweur. À l'heure actuelle, on ne dispose pas de méthodes efficaces pour faire ces vérifications dans le cas des modèles multiniveaux (Goldstein 1995 p. 29). Mais on peut évidemment analyser l'ensemble de données sans les valeurs extrêmes. C'est ce que nous avons fait au tableau 4.

Tableau 4
Analyse de l'ensemble de données sans les valeurs extrêmes liées aux intervieweurs
(é.-t. entre parenthèses)

Fixe		
Niveau des mesures		
constante	3,853	(0,162)
Niveau des répondants		
sexe	1,820	(0,153)
niveau de scolarité	-1,929	(0,149)
presse1	1,160	(0,090)
presse2	1,217	(0,102)
Aléatoire		
Niveau 2		
Intervieweurs		
$\sigma^2_{\text{constant}}$	2,495	(0,333)
Répondants		
$\sigma^2_{\text{constant}}$	7,109	(0,530)
Niveau des mesures		
σ^2_e	13,420	(0,481)
-2 LL		26850,2

Pour procéder à l'analyse présentée au tableau 4, nous avons exclu deux intervieweurs, à savoir celui qui correspond à la valeur résiduelle la plus faible et celui qui correspond à la valeur résiduelle la plus élevée. Les coefficients qui figurent dans ce tableau sont fort semblables à ceux présentés pour le modèle c au tableau 3. La variance liée à l'intervieweur a diminué légèrement, grâce à l'exclusion des deux valeurs extrêmes, mais les résultats ne donnent aucune preuve que les valeurs extrêmes ont un effet important sur les résultats.

L'autre hypothèse concernant les effets aléatoires associés à l'intervieweur et au répondant est celle qui a trait à leur indépendance réciproque. Les résidus liés à l'intervieweur et au répondant ne devraient pas être corrélés. Naturellement, cette hypothèse est plus difficile à vérifier, parce que les résidus sont chacun lié à leur unité respective et que ces unités ne correspondent pas. On obtient 3 028 résidus liés aux répondants et 275 liés aux intervieweurs. Il est possible de vérifier indirectement cette hypothèse en attribuant aux répondants les résidus liés aux intervieweurs. C'est ce qu'illustre le graphique 3.



Graphique 3. Résidus standardisés liés aux répondants selon les résidus standardisés liés aux intervieweurs.

De nouveau, ce graphique attire l'attention sur le plus grand nombre d'écarts vers le haut et sur les valeurs extrêmes liées aux intervieweurs. Ces déviations mises à part, on n'observe aucune régularité particulière. À cause des valeurs extrêmes liées aux intervieweurs, le nombre d'observations est plus faible du côté droite que du côté gauche du graphique. Cependant, les résidus liés aux répondants n'ont pas vraiment tendance à être plus faibles si les résidus liés aux intervieweurs sont plus élevés. Le graphique ne permet pas non plus de supposer que l'inverse est vrai.

La vérification présentée au graphique 3 est imparfaite, car elle consiste à attribuer aux répondants les résidus liés aux intervieweurs. Une meilleure solution consisterait à ajuster un modèle plus complexe tenant compte d'une interaction entre les deux effets aléatoires. Goldstein (1995, 119) propose ce genre de modèle. L'exécution d'un test en vue de déterminer l'amélioration du modèle grâce au terme d'interaction donne une idée de l'existence d'une corrélation éventuelle entre les résidus. Une autre solution consisterait à ajouter un niveau supplémentaire (la région) aux niveaux des intervieweurs et des répondants. Ce modèle inclurait un terme pour la variation régionale qui entraînerait une corrélation entre les résidus liés à l'intervieweur et au répondant. Snijders et Bosker (1999, p. 159-160) décrivent ce modèle. Toutefois, les deux modèles nécessitent un paramètrage différent avec création d'ensembles distincts de variables fictives. Leur explication ferait, en soi, l'objet d'un rapport et dépasse donc le cadre du présent article.

BIBLIOGRAPHIE

BAILAR, B., BAILEY, L., et STEVENS, J. (1977). Measures of interviewer bias and variance. *Journal of Marketing Research*, 14, 337-343.

BEERTEN, R., BILLIET, J., CARTON, A., et SWYNGEDOUW, M. (1997). *1995 General Election Study Flanders-Belgium. Codebook and Questionnaire.* Leuven: ISPO/Departement Sociologie K.U.Leuven.

BRYK, A.S., et RAUDENBUSH, S. (1992). *Hierarchical Linear Models Applications and Data Analysis Methods.* Newbury Park - London: Sage.

CARTON, A., SWYNGEDOUW, M., BILLIET, J., et BEERTEN, R. (1993). *Source Book of the Voters' Study in Connection with the 1991 General Election.* Leuven: Sociologisch Onderzoeksinstituut/ ISPO.

DIPRETE, T.A., et FORRISTAL, J.D. (1994). Multilevel models: Methods and substance. *Annual Review of Sociology*, 20, 331-357.

GOLDSTEIN, H. (1995). *Multilevel Statistical Models.* London: Edward Arnold.

GROVES, R. M. (1989). *Survey Error and Survey Costs.* New York: Wiley.

HANSON, R.H., et MARKS, E.S. (1958). Influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association*, 53, 635-655.

HOX, J.J. (1994). Hierarchical regression models for interviewer and respondent effects. *Sociological Methods and Research*, 22, 300-318.

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

ISPO/PIOP (1995). *1991 General Election Study Belgium. Codebook and Questionnaire.* Leuven: ISPO.

KREFT, I.G., et DE LEEUW, J. (1998). *Introducing Multilevel Modeling.* London: Sage Publications.

KROSNICK, J.A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213-236.

LOOSVELDT, G., et CARTON, A. (1997). Evaluation of nonresponse in the Belgian Election Panel Study '91 - '95. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1017-1022.

RASBASH, J., et GOLDSTEIN, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational Statistics*, 19, 337-350.

RASBASH, J., et WOODHOUSE, G. (1996). *MLn Command Reference.* Version 1.0a. London: Multilevel Models Project. Institute of Education, University of London.

RAUDENBUSH, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, 18, 321-349.

SÄRNDAL, C., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.

SCHUMAN, H., et PRESSER, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context*. New York: Academic Press.

SCHWARZ, N., et SUDMAN, S. (1995). *Answering Questions*. San Francisco: Jossey-Bass.

SNIJDERS, T. (1996). Analysis of longitudinal data using the hierarchical linear model. *Quality & Quantity*, 30, 405-426.

SNIJDERS, T., et BOSKER, R. (1999). *Multilevel Analysis. An introduction to basic and advanced multilevel modeling*. London: Sage Publications.

SUDMAN, S., et BRADBURN, N. (1974). *Response Effects in Surveys*. Chicago: Aldine.

YANG, M., et GOLDSTEIN, H. (1996). Multilevel models for longitudinal data. Dans *Analysis of Change. Advanced Techniques in Panel Data Analysis*. (Éds. U. Engel, et J. Reinecke). Berlijn - New York: Walter de Gruiter, 191-220.

Techniques d'enquête, décembre 2000
Vol. 26, n° 2, pp. 225-234
Statistique Canada

La conception des écrans et l'ordre des questions dans un module IAO Résultats d'une expérience de facilité d'utilisation menée sur le terrain

MAREK FUCHS[1]

RÉSUMÉ

La conception des écrans et la conception des questionnaires ont une influence sur le comportement des intervieweurs dans un environnement IAO. Des recherches antérieures ont montré que les intervieweurs peuvent travailler de façon plus convenable et efficace si des fonctions et caractéristiques appropriées sont incorporées dans le module IAO. Des expériences de facilité d'utilisation fondées sur la liste des membres des ménages de deux importantes enquêtes gouvernementales ont montré que le recours à des grilles et à des tableaux facilite le travail de l'intervieweur. Ces expériences ont été menées dans des conditions de laboratoire, mais nous avons des résultats d'une première expérience sur le terrain. En mars 1998, une enquête de type ITAO a été menée auprès d'immigrants en Allemagne (taux de réponse 84%, $n = 501$). Dans cette étude de production, quatre versions d'une liste des membres du ménage ont été comparées afin de vérifier deux conceptions des écrans et deux ordres des questions dans une conception factorielle 2×2. Les quatre versions ont été attribuées de façon aléatoire à des intervieweurs et à des répondants. Des mesures du temps ont été intégrées au programme ITAO, et 234 interviews sélectionnées au hasard ont été enregistrées sur bande magnétoscopique, puis analysées en fonction d'un schéma de codage. À l'aide des données, nous avons évalué la facilité d'utilisation des différentes caractéristiques de conception de l'IAO. Les résultats indiquent que la conception des écrans de même que l'ordre des questions ont une influence appréciable sur la durée de l'interview et le comportement des intervieweurs. En particulier, la version quadrillée et thématique donne les résultats les plus rapides pour ce qui est du temps mis à remplir le questionnaire. Les résultats tirés des données de codage indiquent que les différences entre les versions sont attribuables à des comportements particuliers des intervieweurs et des répondants. Les données indiquent que la version thématique quadrillée facilite chez l'intervieweur un comportement axé sur le répondant et donne lieu au meilleur rendement des intervieweurs pour ce qui est de la durée.

MOTS CLÉS: Interview assistée par ordinateur; essai de facilité d'utilisation; expérience sur le terrain; conception des écrans; ordre des questions.

1. INTRODUCTION

Les interviews assistées par ordinateur sont en train de devenir une technique d'enquête standard (Couper, Baker, Bethlehem, Clark, Martin, Nicholls et O'Reilly 1998). Pour les enquêtes téléphoniques aussi bien que pour les interviews personnelles, de plus en plus d'études sont menées à l'aide de techniques d'interview assistée par ordinateur (IAO). Plusieurs enquêtes gouvernementales importantes aux États-Unis sont en train de passer à l'IAO, ou cette transition est déjà terminée. Même en Europe, on observe une évolution vers l'interview assistée par ordinateur (Schneid 1991; Fuchs 1994, 1995; Laurie et Moon 1997; Projektgruppe SOEP 1998) – même si, pour le moment, les aspects méthodologiques de ce phénomène ne représentent pas le point central de la recherche européenne.

Les chercheurs et les responsables des enquêtes se fient à l'interview assistée par ordinateur pour plusieurs raisons (toutefois, il semble parfois que de solides arguments sont moins importants que la simple popularité de l'IAO):

- ils espèrent recueillir des données de qualité supérieure grâce à des vérifications incorporées de la cohérence et des intervalles au cours de l'interview;
- l'IAO permet d'enchaîner les questions automatiquement et de concevoir des questionnaires plus complexes sans imposer un fardeau indu à l'intervieweur;
- ils espèrent consacrer moins de temps et d'argent à l'interview et au post-traitement et réduire les budgets d'enquête une fois les coûts initiaux de matériel et de logiciel absorbés;
- ils espèrent bénéficier de la capacité de l'IAO d'intégrer des données externes à l'interview, fonction particulièrement intéressante pour les études par panel.

La transition générale à l'IAO est vue d'un bon œil. Les chercheurs et les directeurs sur le terrain en tirent des avantages (Nicholls et deLeeuw 1996) et les intervieweurs (Couper et Burt 1994) de même que les répondants (Baker 1992) manifestent beaucoup de sympathie ou du moins de l'acceptation. Par contre, les interviews assistées par ordinateur ont introduit certains problèmes supplémentaires dans le déroulement des interviews: au début, la recherche méthodologique a porté surtout sur des problèmes de matériel et de logiciel (on trouvera des synthèses dans Couper, Groves et Kosary 1989 et dans Weeks 1992). Des études plus récentes ont traité de l'acceptation par les intervieweurs et les répondants, de la

[1] Dr. Marek Fuchs, Catholic University of Eichstätt, Department of Sociology, Ostenstrasse 26, 85071 Eichstätt, Germany. Courrier électronique: marek.fuchs@ku-eichstaett.de.

durée des interviews et de la facilité d'utilisation (on trouvera une synthèse dans Couper et coll. 1998). Le présent document fait partie de cette récente discussion des «effets de la technologie» (Fuchs, Couper et Hansen 2000).

2. CONTEXTE THÉORIQUE

Dans la présente analyse, le point de convergence théorique touche surtout à deux aspects de la facilité d'utilisation: 1) la segmentation du cheminement de l'interview et 2) le manque de souplesse pour l'intervieweur.

1. Segmentation: Dans un environnement IPAO, l'intervieweur se voit imposer un fardeau supplémentaire: le processus de saisie a lieu durant l'interview. Normalement, un intervieweur lit une question, reçoit une réponse, entre les données, appuie sur [enter] et voit ensuite apparaître le prochain écran et la question suivante. Comparativement à l'IPC, les intervieweurs ne peuvent pas regarder en avant et anticiper la question suivante tout en inscrivant les réponses à la question précédente, et ils ne peuvent pas commencer à lire la question suivante avant d'appuyer sur [enter] – ils ne peuvent pas s'occuper des deux tâches en même temps. Dans ce genre de procédure, l'interaction entre l'intervieweur et le répondant est segmentée par des touches [enter]. Il n'existe encore aucune indication quantitative que ce genre de segmentation nuit aux données ou au déroulement de l'interview. Par contre, on fait remarquer que l'intervieweur n'a pas de «vue d'ensemble» et que la pertinence des questions et les relations entre elles ne sont peut-être pas claires (House 1985; Groves et Mathiowetz 1984).

Nos résultats, obtenus de plusieurs séries d'essais de facilité d'utilisation menées en laboratoire sur la configuration des écrans d'une liste de membres de ménage (Couper et coll. 1997; Hansen, Couper and Fuchs 1998), ont mis en valeur une configuration d'écran particulière qui permet à l'intervieweur de mieux comprendre le questionnaire, tout en conservant l'interaction avec le répondant et tout en inscrivant les données. Deux versions d'une série de questions ont été mises à l'essai en laboratoire en fonction du temps mis à remplir le questionnaire et de la facilité d'utilisation. Nous avons comparé une conception dite thématique et une conception dite quadrillée. House et Nicholls (1988) ont décrit trois stratégies de conception des écrans pour des modules assistés par ordinateur: axée sur l'élément, axée sur l'écran et axée sur la forme. Un module axé sur l'élément fait afficher une question et un champ de saisie à la fois, et les opérations logiques ont lieu durant la transition d'un écran à l'autre. Facile à programmer, une telle conception dirige l'attention de l'intervieweur sur la question même. Un module axé sur l'écran regroupe plusieurs éléments supposant des réponses séquentielles. Toutes les opérations logiques sont exécutées après chaque élément. Pour ce qui est d'un écran axé sur la forme, plusieurs éléments sont présentés en même temps dans un tableau ou dans une grille, et l'intervieweur peut appuyer sur des touches de curseur pour passer d'un champ à l'autre et les remplir dans n'importe quel ordre.

La version axée sur l'élément vérifiée dans notre expérience correspond aux caractéristiques décrites par House et Nicholls (1988) pour une stratégie fondée sur l'écran. Pour sa part, la conception fondée sur une grille est en réalité un module axé sur la forme. Elle permet aux intervieweurs d'enregistrer l'information dans l'ordre choisi par le répondant; elle donne à l'intervieweur une meilleure vue d'ensemble du module, et les mises à jour ainsi que les reculs en arrière sont plus faciles (on trouvera des détails dans Couper et coll. 1997). De plus, elle permet à l'intervieweur d'avoir plus d'une question à un même écran (avantages: vitesse d'exécution et informations contextuelles). Le graphique ci-dessous illustre la conception des écrans fondée sur l'élément et celle fondée sur la grille pour les IAO.

Nous avons observé que la conception fondée sur la grille permet de réduire la segmentation. Les intervieweurs peuvent commencer à lire la question suivante tout en inscrivant les données de la question précédente. Même un recul en arrière semble plus facile dans la conception fondée sur une grille. Par contre, nous n'avons constaté qu'un faible appui pour la conception fondée sur une grille relativement au temps mis à exécuter la tâche (on trouvera des détails dans Couper et coll. 1997). On peut donc poser la question: Que faire pour réduire la segmentation et améliorer davantage l'efficacité d'une liste de membres de ménage pour ce qui est de la durée?

2. Manque de souplesse: La deuxième caractéristique qui risque de poser des problèmes dans une interview assistée par ordinateur est le manque de souplesse. Un des avantages du module IAO est que l'intervieweur ne peut guère sauter de questions. Même si les modules IAO peuvent faire amplement usage de l'enchaînement des questions et de filtres, l'ordre des questions est déterminé d'avance. Normalement, il faut appuyer sur la touche [enter] pour chaque question avant que le système ne passe à l'écran suivant. On trouve cette rigidité de l'ordre des questions avantageuse car elle évite toute difficulté de la part de l'intervieweur de suivre le déroulement, les questions destinées à des répondants particuliers, les filtres, l'enchaînement des questions, *etc.* L'intervieweur n'a pas à s'en soucier et peut s'en tenir à l'exécution des éléments. Par contre, il en résulte un ordre très rigide des questions et très peu de souplesse, pour l'intervieweur, quant à l'ordre des questions. Voici un exemple qui illustre bien cet effet: la plupart des modules IAO appliquent un ordre des questions à la liste des membres du ménage, toutes les questions adressées à une personne étant posées avant que l'intervieweur ne pose les mêmes questions à la personne suivante (au sujet de la «conception axée

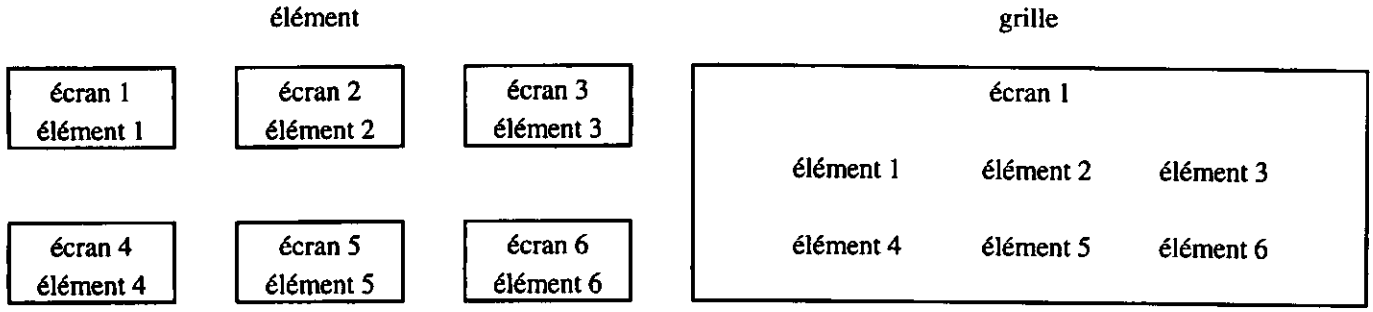


Figure 1. Conception fondée sur l'élément et conception fondée sur la grille

sur la personne», voir Couper et coll. 1997 et Fuchs 2001; au sujet des «questions groupées», voir Moyer 1996). Ainsi, le module IAO peut exiger que l'on demande l'âge du répondant, son niveau de scolarité et d'autres questions avant de passer à l'âge de son épouse. (Cela s'explique en partie par le fonctionnement des programmes informatiques et des bases de données: les ménages représentent les principales catégories, les personnes ou les autres entités étant traitées comme des sous-catégories.) Lorsqu'on remplit le questionnaire pour une liste de membres du ménage, il arrive (et même assez souvent, comme on le verra ci-dessous) que le répondant fournisse non seulement la réponse à la question posée (par exemple «j'ai 34 ans»), mais également à une question connexe: «j'ai 34 ans et mon épouse en a 32»; le répondant pourra aussi déclarer: «nous sommes tous noirs» lorsqu'il est question de race (Oksenberg, Beebe, Blixt et Cannell 1992).

En présence d'un questionnaire imprimé, il est facile pour un intervieweur d'utiliser immédiatement l'information supplémentaire fournie par le répondant. Pour une réponse du type «nous sommes tous noirs», par exemple, l'intervieweur peut facilement cocher les cases appropriées pour tous les membres du ménage en même temps. Une personne qui s'intéresse à la conception des questionnaires peut alors se poser la question suivante: vu le manque de souplesse d'un environnement assisté par ordinateur, quel est le meilleur ordre des questions si l'on veut recueillir des renseignements pour tous les membres du ménage?

Moore et Moyer ont présenté les résultats d'une expérience menée sur deux ordres des questions servant à recueillir des informations sur toutes les personnes admissibles d'un ménage (Moore et Moyer 1998a, 1998b). Selon le premier modèle, on pose toutes les questions pour la première personne admissible du ménage, puis on passe à la personne suivante. Un tel ordre des questions est ce que l'on appelle la stratégie axée sur la personne. Selon le deuxième modèle, appelé la stratégie thématique, la première question est posée pour toutes les personnes admissibles, puis on pose la deuxième question pour toutes les personnes, et ainsi de suite. Les résultats de Moore et Moyer appuient fortement une conception thématique: le modèle thématique comporte une non-réponse moins élevée, moins d'interruptions et de refus, et il prend appréciablement moins de temps. Les intervieweurs préfèrent nettement ce modèle.

Dans l'expérience que nous présentons ici, nous avons fait appel aux avantages de la stratégie thématique et de la conception des écrans fondée sur une grille: nous avons combiné les deux modèles d'écran (conception fondée sur l'élément et conception fondée sur la grille) et les deux ordres des questions (ordre axé sur la personne et ordre axé sur le thème) de façon à vérifier les quatre versions résultantes dans une expérience sur le terrain. Pour ce faire, nous avons adopté l'hypothèse qui suit: la facilité d'utilisation d'un module IAO ne relève pas seulement de la programmation, mais elle est liée à la conception du questionnaire et au contexte social de l'interview. Les deux aspects d'un module assisté par ordinateur, c'est-à-dire la conception des écrans et l'ordre des questions, peuvent favoriser ou détruire la bonne marche d'une interview. Les résultats de la recherche antérieure avaient donné lieu à l'hypothèse suivante: la combinaison d'une conception des écrans axée sur une grille et d'un ordre des questions axé sur un thème suscite l'interaction la plus efficace entre l'intervieweur et le répondant.

3. MÉTHODE

L'expérience s'est déroulée en Allemagne en mars 1998. On a interrogé des immigrants d'origine allemande provenant de la Pologne, de la Roumanie et de l'ancienne Union soviétique. À compter du 28 février 1998, et jusqu'au 20 mars 1998, les intervieweurs ont mené $n = 501$ interviews. On a fait parvenir aux répondants une lettre d'introduction à l'avance, et on a communiqué avec eux par téléphone jusqu'à 15 fois. Le taux de réponse a été de 84%, et la non-réponse à une question a été assez faible. Les interviews ont été menées dans le cadre du programme ITAO CI3. Quelque 95 questions ont été posées sur divers thèmes. En moyenne, les interviews ont duré 23 minutes.

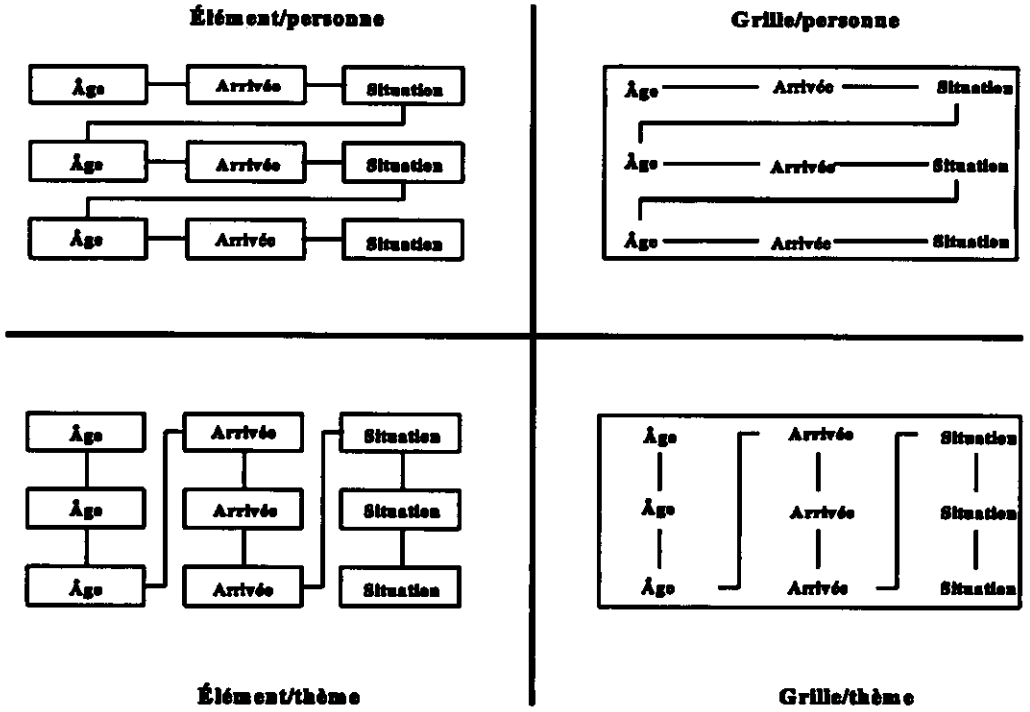


Figure 2. Quatre versions vérifiées dans l'expérience (chaque cases représente un écran)

Quatre versions d'une liste des membres d'un petit ménage, avec trois éléments par personne, ont été incluses dans le module: une version élément/personne, une version grille/personne, une version élément/thème et une version grille/thème. Toutes les versions ont fait appel au même libellé des questions et aux mêmes instructions pour l'intervieweur, mais nous avons modifié la conception des écrans et l'ordre des questions en fonction de la stratégie théorique décrite ci-dessus (figure 2). La version élément/personne est considérée comme la version standard – elle représente la conception des questionnaires appliquée habituellement aux volets socio-démographiques des enquêtes IAO. Une des quatre versions a été attribuée de façon aléatoire à chaque interview – et donc aux intervieweurs et aux répondants. Nous avons mesuré le temps total requis pour la liste des membres du ménage, ainsi que le temps consacré à chaque élément de cette section des 501 interviews. De plus, 234 interviews ont été choisies au hasard, et nous avons enregistré sur bande magnétoscopique l'intervieweur en train de parcourir la section de la liste des membres du ménage. Les segments vidéo ont été codés en fonction du comportement de l'intervieweur et du répondant, et les données résultantes ont été combinées aux mesures du temps.

4. RÉSULTATS

La durée a été appréciablement différente pour chacune des quatre versions: les intervieweurs ont consacré 6,6 secondes par élément dans la version élément/personne (considérée comme la version standard). Par contre, chaque élément a supposé 5,5 secondes dans la version grille/thème. C'est là une réduction de 17% environ pour la version grille/thème. Les deux autres versions se situaient entre les deux premières.

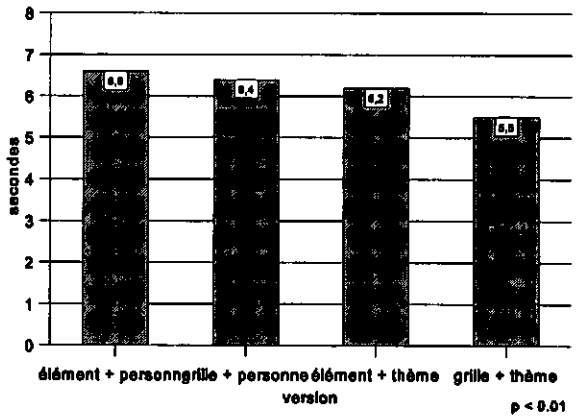


Figure 3. Durée par élement selon la version

Il importe de mentionner que les deux facteurs semblent jouer un rôle dans la diminution du temps mis à accomplir la tâche. Si nous distinguons les deux facteurs, nous obtenons les résultats suivants: les deux versions axées sur le thème sont appréciablement plus courtes que les deux versions fondées sur la personne, et les deux versions fondées sur la grille supposent un temps appréciablement moindre que les deux versions axées sur la personne. L'effet combiné s'applique à la version grille/thème, et

donne une valeur de 5,5 secondes par élément. (Une analyse de la variance indique que les deux facteurs – c'est-à-dire la conception des écrans et l'ordre des questions – favorisent indépendamment la diminution du temps (conception des écrans: $p < 0,01$, un tiers de l'effet total; ordre des questions: $p < 0,001$, deux tiers de l'effet total, aucune interaction appréciable).)

Pourquoi la version grille/thème est-elle plus rapide? Une analyse détaillée indique que cette version est particulièrement rapide pour la collecte des renseignements sur la deuxième personne et les personnes subséquentes du ménage – effet appréciable que l'on appelle l'effet de boucle (Fuchs 2001). Cette expression décrit le phénomène suivant: l'intervieweur consacre beaucoup plus de temps à recueillir les renseignements pour la première personne du ménage comparativement aux personnes subséquentes. L'effet de boucle moyen donne au total 3,4 secondes par élément, ce qui représente une réduction de 38% environ relativement à la première personne (tableau 1).

L'effet de boucle n'est pas particulièrement propre à cette expérience. Nous avons observé des effets de boucle dans nos expériences antérieures liées à la liste des membres des ménages NHIS (Couper et coll. 1997). Toutefois, il est intéressant de noter que l'effet de boucle est appréciablement plus grand pour les versions axées sur le thème qu'il ne l'est pour les versions qui suivent un ordre des questions axé sur la personne (tableau 1). Ainsi, les versions fondées sur le thème permettent d'accélérer le processus pour la deuxième personne et les personnes subséquentes d'un ménage, et comportent donc un effet de boucle plus grand. (Il est possible que la conception de notre expérience a eu comme effet que des intervieweurs ne savaient pas quelle version ils utilisaient, ce qui a pu faire baisser leur rendement pour le tout premier élément. Or cet effet devrait être le même pour toutes les versions, de sorte que les résultats ne devraient pas être touchés.)

Tableau 1
Durée et effet de boucle (en secondes)

Élément	Âge	Arrivée	Situation	Tous les éléments
Durée par élément				
Première personne du ménage	9,4	9,9	7,7	9,0
Toutes les autres personnes du ménage	6,6	5,7	4,5	5,6
Toutes les personnes	8,0***	7,8***	6,1***	7,3***
Effet de boucle				
Différence entre la première et toutes les autres personnes du ménage	-2,8	-4,2	-3,2	-3,4
Effet de boucle selon la version				
Grille + thème	-6,8	-6,4	-4,6	-5,9
Élément + thème	-5,2	-8,3	-4,3	-6,0
Grille + personne	-0,5	-2,7	-3,0	-2,1
Élément + personne	0,3	-0,3	-1,3	-0,4
Effet de boucle moyen	-2,8***	-4,2**	-3,2**	-3,4***

**$p < 0,01$; ***$p < 0,001$

L'analyse des bandes vidéo nous a permis de fournir des raisons partielles au moins pour ces différences: dans des conditions axées sur le thème, les intervieweurs et les répondants s'adaptent différemment à l'interview comparativement aux versions fondées sur la personne. Lorsque l'on pose les questions pour toutes les personnes du ménage, le répondant reconnaît très rapidement le caractère logique du processus. Dans une forte proportion des cas (30% environ), la réaction est du type: «nous sommes tous arrivés pendant la même année» (ce qui veut dire: «cessez de me répéter cette question»).

Si le questionnaire suit la conception fondée sur la personne, l'intervieweur doit mémoriser cette information, et au moment d'aborder la personne suivante, il doit se rappeler: «ne répète pas cette question, car le répondant a déjà fourni la bonne réponse!». Parfois l'intervieweur s'en souvient, mais la plupart du temps il pose la question de nouveau. Cela est particulièrement vrai pour une configuration d'écran axée sur l'élément et ne fournissant aucune indication des réponses à la même question pour les autres membres du ménage. Dans un modèle fondé sur le thème, par contre, l'intervieweur peut facilement s'adapter à cette situation. Il suffit alors d'entrer le même code pour toutes les personnes du ménage sans répéter la question. L'intervieweur aussi bien que le répondant s'habituent aux questions, de sorte que le processus question-réponse comporte moins d'interventions verbales de la part de l'intervieweur comme de celle du répondant. L'intervieweur aussi bien que le répondant peuvent anticiper la prochaine question, et l'interview se déroule plus facilement. Cela est particulièrement vrai lorsque le module IAO comporte une grille et fournit d'autres informations contextuelles, par exemple, les réponses fournies pour d'autres membres du ménage à la même question. (Les résultats figurant dans la partie inférieure de la figure 4 permettent de conclure que la version grille/personne ne bénéficie pas dans la même mesure des avantages de la stratégie fondée sur le thème. Toutefois, à cause de la conception de la grille, l'effet de boucle est appréciablement plus grand que pour la version élément/personne). Par conséquent, le temps consacré par élément est appréciablement plus court, et l'intervieweur peut adopter un comportement orienté vers le répondant semblable à celui dont il est question dans Schober et Conrad (1997).

L'inclusion de remarques par l'intervieweur sert parfois de signal que la réponse à la question précédente a été inscrite de façon à aiguiller le répondant, celui-ci devinant la question suivante et fournissant la réponse appropriée même sans stimulation supplémentaire. Dans un cas extrême, cela peut mener le répondant à fournir l'information au sujet de toutes les personnes du ménage en même temps: «nous sommes tous venus pendant la même année». Les diverses versions mises à l'essai dans cette expérience favorisent ce genre de comportement à différents degrés. Les résultats nous permettent de conclure que la version grille/thème pousse les intervieweurs et les répondants à

dévier de l'interview prévue de façon plus marquée que les autres versions. Pour ce qui est de la durée, cette version permet à l'intervieweur d'utiliser de façon efficace l'information fournie pour tous les membres du ménage en même temps. L'analyse du codage vidéo appuie notre interprétation de l'occurrence propre à une version de comportements des intervieweurs (1) et de comportements des répondants (2) qui permettent de gagner du temps:

1. L'analyse des bandes vidéo permet d'observer des comportements d'intervieweur ne correspondant pas aux procédures standard: quelque 78% des éléments sont lus tels quels, et 9,3% des éléments ne sont pas présentés au répondant par l'intervieweur. De plus, dans 5% des cas, l'intervieweur ne lit pas la question, mais fournit plutôt un autre stimulus comportant le lien entre la personne suivante et le répondant (par exemple «... et votre épouse?»). (Il est intéressant de noter que les intervieweurs choisissent les mêmes expressions verbales, spontanément, que Moore et Moyer 1998a, 1998b ont utilisées dans leurs expériences sur l'ordre des questions.) Dans 5,5% des cas, l'intervieweur ne lit pas la question, mais vérifie plutôt la réponse («... et votre épouse a 32 ans?».) On observe également quelques questions incomplètes et des réponses inscrites incorrectement. Au total, 22% environ des éléments sont touchés par au moins un comportement d'intervieweur qui ne correspond pas à une séquence d'interview standard – ce qui représente une valeur étonnamment élevée puisque tous les intervieweurs savaient que les interviews étaient enregistrées sur bande magnétoscopique! Toutefois, comparativement à d'autres études sur le comportement des intervieweurs, les valeurs sont appréciablement moins élevées. Oksenberg, Cannell et Blixt (1996), par exemple, ont appliqué un codage du comportement à la National Medical Expenditure Survey, observant de 37% à 41% de comportements d'intervieweur de ce genre. Il sera question ci-dessous de la mesure dans laquelle ce genre de comportement aide à obtenir des mesures valables.

De ces résultats particuliers, nous tirons la conclusion suivante: la plupart de ces comportements représentent une espèce de raccourci, par exemple l'intervieweur ne lit pas le texte exact de la question, tente de faciliter la conversation ou de la rendre plus conforme aux règles de la conversation. À notre avis, cela indique que l'intervieweur ne veut pas poser une question à laquelle une réponse a déjà été fournie. L'intervieweur ne veut pas rester indifférent devant l'information fournie par le répondant, cherchant plutôt à respecter les règles de la conversation. Ces comportements prennent donc moins de temps que des comportements d'intervieweur standard. Selon nous, la priorité ne consiste donc pas à épargner du temps, mais bien à personnaliser le processus question-réponse en fonction de comportements du répondant non prévus et ne pouvant pas être intégrés par le module assisté par ordinateur.

Table 2
Comportement de l'intervieweur et comportement du répondant selon la version

	Grille + thème	Élément + thème	Grille + personne	Élément + personne	Total
Proportion moyenne d'éléments touchés par un comportement d'intervieweur ne correspondant pas à l'interview consignée selon le cas	0,48	0,43	0,34	0,21	0,36***
Le répondant fournit l'information pour toutes les personnes du ménage en même temps	38,2%	44,4%	29,0%	10,8%	29,7%***

***p < 0,001

Afin de distinguer la proportion de cas touchés par un certain comportement du répondant et la proportion moyenne d'éléments par cas (!) touchés par un certain comportement de l'intervieweur, nous avons utilisé des pourcentages pour ceux-là et des décimales pour ceux-ci.

Afin de pouvoir comparer les quatre versions du modèle d'écran en fonction des comportements d'intervieweur ne correspondant pas à l'interview standard, nous avons calculé la proportion des éléments par cas touchés par ce genre de comportement. Les quatre versions manifestent des écarts fort différents relativement à la séquence prévue des interviews: l'application de la version grille/thème à une interview entraîne plus de deux fois plus de comportements de ce genre (la proportion moyenne des éléments touchés est de 0,48) que la version élément/personne (0,21) qui représente la norme pour la plupart des études menées jusqu'à présent. (Une analyse de la variance indique que les deux facteurs influencent l'effet global indépendamment (conception des écrans: $p < 0,001$; ordre des questions: $p < 0,001$; aucun effet appréciable d'interaction). On peut attribuer 25% environ de l'effet global à la conception des écrans et 75% environ à l'ordre des questions.) Cela influence aussi le temps consacré à l'interview: les éléments touchés par un comportement d'intervieweur qui diffère de la séquence prévue de l'interview prennent appréciablement moins de temps (4,0 secondes) que les éléments qui respectent la séquence ordinaire (6,8 secondes; $p < 0,001$).

2. De plus, une analyse du comportement des répondants indique que le modèle thématique suscite une plus forte proportion de cas (42,3% comparativement à 19,7% pour la stratégie personnelle; $p < 0,001$) dans lesquels le répondant fournit au moins une fois des renseignements pour toutes les personnes ou pour un groupe de personnes du ménage en même temps (par exemple «nous sommes tous venus pendant la même année»; «nous avons tous le même statut juridique»). Par contre, la différence entre le modèle

axé sur la grille et le modèle axé sur l'élément est appréciablement plus faible (33,6% contre 26,1%), sans toutefois devenir significative. L'analyse révèle cependant un effet d'interaction appréciable ($p < 0,05$): le recours à un ordre des questions axé sur le thème n'entraîne pas une différence appréciable. Toutefois, une conception fondée sur une grille jumelée à un ordre des questions axé sur un thème fait augmenter le nombre de cas dans lesquels le répondant fournit des renseignements pour tous les membres du ménage en même temps.

Il est remarquable que les résultats diffèrent même pour les deux conceptions des écrans lorsqu'on utilise un ordre des questions axé sur la personne. L'étude a été menée par téléphone, les répondants n'étant pas du tout au courant de la conception des écrans. La seule explication possible est que l'intervieweur modifie son comportement en fonction de la conception des écrans, aiguillant le répondant de façon différente. Les répondants aussi bien que les intervieweurs réagissent donc à la conception des écrans et à l'ordre des questions dans le modèle de type grille/personne d'une façon qui facilite l'interaction intervieweur-répondant, ce qui permet à l'interview de se dérouler plus doucement. (Comme nous l'avons vu, les intervieweurs modifient leur comportement même pour un modèle de type grille/personne (tableau 2), mais l'ordre des questions ne suscite pas chez le répondant un comportement équivalent.)

Un inconvénient possible de ces comportements intervieweur-répondant pourrait être une baisse de la qualité des données à cause de modifications du processus question-réponse prédéterminé; le répondant considère alors la réponse de façon moins intense et rigoureuse. Nous n'observons que de très rares valeurs manquantes pour une question, de sorte que l'analyse de cet indicateur standard de la qualité des données n'est pas efficace. En réalité, nous ne nous attendons pas à une plus forte proportion de valeurs manquantes pour une question dans l'une ou l'autre version. Toutefois, on pourrait s'interroger sur l'homogénéité des réponses fournies par le répondant. Dans une forte proportion des cas, le répondant entend le texte complet de la question une seule fois et il est possible que le répondant soit moins rigoureux au moment de répondre à la même question pour des membres subséquents du ménage. De plus, le fait de répondre pour tous les membres du ménage en même temps («nous sommes tous arrivés pendant la même année») risque d'accroître l'homogénéité de la réponse et donc de faire baisser la qualité des données.

Afin de pouvoir évaluer cet inconvénient possible, nous avons calculé le nombre de catégories de réponse distinctes choisies par le répondant pour un élément particulier et pour tous les membres du ménage (par exemple pour l'année d'arrivée: le répondant 1985, le partenaire 1987, la fille 1987, le fils 1988 = trois catégories de réponse distinctes). Cela devrait indiquer si les répondants qui adoptent le raccourci («nous sommes tous venus pendant la même année») sont seulement ceux pour qui c'est réellement le cas, ou bien si même les autres répondants ont fourni une réponse pour tous les membres du ménage même s'ils auraient dû choisir des catégories de réponse distinctes à cause de la situation dans leur ménage particulier (malheureusement nous n'avons aucune validation externe des réponses fournies). En examinant le nombre moyen de catégories de réponse distinctes (tableau 3), nous ne remarquons aucune différence pour ce qui est de l'homogénéité des données. Pour l'année d'arrivée comme pour le statut juridique (à titre d'Allemand ou à titre d'étranger), il n'existe aucune différence visible entre les versions. Pour toutes les versions, le nombre moyen de catégories de réponse distinctes choisies (une pour chaque personne d'un ménage) ne manifeste aucune différence appréciable

Tableau 3
Nombre moyen de catégories distinctes (homogénéité) par ménage selon la version

Variable	Grille + thème	Élément + thème	Grille + personne	Élément + personne	Total
Année d'arrivée (19 catégories)	1,2	1,2	1,2	1,3	1,2
Statut (4 catégories)	1,3	1,3	1,3	1,3	1,3

Aucune différence appréciable

Ces résultats ne fournissent qu'une faible indication que le modèle fondé sur une grille ne nuit pas à la qualité des données. Il faudra utiliser d'autres indicateurs standard de la qualité des données comportant des ensembles de données plus grands si l'on veut déterminer dans quelle mesure la qualité des données est affectée. Toutefois, d'après les données disponibles, nous ne pouvons déceler un effet sur la validité des réponses.

5. DISCUSSION ET CONCLUSION

Les résultats de notre comparaison de quatre versions d'une liste des membres d'un ménage (le même libellé des questions étant utilisé pour toutes les versions) montrent que les intervieweurs tout comme les répondants fonctionnent de façon plus efficace dans le cadre d'un modèle de type grille/thème que pour l'une ou l'autre des trois autres versions. Le fait de combiner un schéma des écrans fondée sur une grille et un ordre des questions fondé sur un thème permet de réduire la durée moyenne de 17% environ. Deux tiers de cette réduction peuvent être attribués à l'ordre des questions, et un tiers environ à la configuration des écrans. Il importe de mentionner que l'effet de schéma des écrans est moins marqué que celui de l'ordre des questions et encore plus faible, comparativement à l'effet sur la durée, pour le comportement de l'intervieweur et pour le comportement du répondant.

Même si l'effet de la conception quadrillée sur le comportement de l'intervieweur et sur celui du répondant n'est pas considérable, il aide à cerner deux raisons du meilleur rendement de la version grille/thème pour ce qui est de la durée de l'interview: 1) pour la version grille/thème, l'intervieweur aussi bien que le répondant s'adaptent mieux à la logique du processus question-réponse, les deux pouvant anticiper la prochaine question plus aisément et le processus question-réponse se déroulant plus doucement; 2) cette version entraîne plus de cas dans lesquels le répondant fournit des renseignements plus rapidement pour les personnes du ménage, et fournit plus souvent des renseignements pour tous les membres du ménage ou du moins pour plusieurs membres en même temps. Les résultats ne sont pas tout à fait uniformes, mais cette version particulière permet à l'intervieweur de s'adapter plus facilement à la situation, d'inscrire les renseignements et d'aider le répondant à fournir la prochaine réponse appropriée sans répétition du texte complet de la question.

Nos résultats se rapportent à la discussion sur la conception des questionnaires d'enquête pour la collecte des données assistée par ordinateur avec la participation d'un intervieweur. Les résultats décrits dans le présent exposé permettent de conclure que le recours à des grilles facilite l'interaction intervieweur-répondant et accélère la collecte des données. Les expériences que nous avons menées, dans le laboratoire de facilité d'utilisation du Survey Research Center de l'Université du Michigan sur des modèles quadrillés et thématiques, indiquent qu'il est possible d'améliorer le rendement des intervieweurs en leur fournissant des grilles (Couper et coll. 1997). Moore et Moyer (1998a; 1998b) ont montré qu'il est possible d'améliorer l'efficacité des interviews en adoptant également un ordre des questions fondé sur un thème. Le présent exposé indique qu'une combinaison des deux fonctions est encore plus avantageuse pour les interviews.

Le recours à des grilles et à un ordre des questions fondé sur un thème suscite un nombre accru de cas dans lesquels l'intervieweur s'éloigne de l'interview standard. D'un point de vue strictement méthodologique, cela peut représenter un sérieux inconvénient, surtout si l'intervieweur s'éloigne de la séquence d'interview standard en ayant recours à des questions globales pour tous les membres du ménage. Ainsi, Martin (1999) a décelé un accroissement appréciable du nombre de personnes dénombrées dans un ménage lorsque l'on pose d'autres questions. De plus, Kindermann et coll. (1997) ont montré, pour des questions fondées sur une liste hors-ménage, que des signaux supplémentaires sur la victimisation entraînent une augmentation appréciable des crimes dénombrés. Si l'on généralise ces résultats pour des questions globales d'une personne à l'autre, on s'attendrait à une diminution de la qualité des données lorsque les intervieweurs ont recours à des comportements non prévus consistant à appliquer des questions globales. Toutefois, les résultats signalés dans le présent exposé n'indiquent pas que les intervieweurs ont recours à des questions globales, et l'auteur ne recommande pas d'avoir recours à des questions globales lors de la conception d'un questionnaire d'enquête. Les résultats favorisent plutôt une conception des écrans permettant à l'intervieweur d'utiliser les renseignements fournis par le répondant lorsque celui-ci adopte un mode global et fournit des renseignements pour plusieurs membres du ménage en même temps. Il ne s'agit donc pas d'encourager les chercheurs à utiliser des questions globales, et nous ne souhaitons pas que les intervieweurs modifient les questions prévues de façon à poser des questions globales. Toutefois, lorsqu'un répondant fournit plus de renseignements qu'on ne lui en demande, la conception des écrans d'un questionnaire IAO ne devrait pas empêcher l'intervieweur d'en bénéficier.

Il a été démontré qu'un modèle fondé sur une grille facilite la tâche des intervieweurs à cet égard, car il permet à ceux-ci de rajuster leur comportement en fonction des règles générales de la conversation. Les résultats du codage des comportements indiquent que les intervieweurs s'éloignent fréquemment des procédures d'interview standard. Oksenberg et coll. (1992, 1993) ont affirmé que ces changements reflètent souvent des rajustements apportés par les intervieweurs devant les modalités de la situation de façon à mieux enchaîner avec ce qui précède immédiatement ou de respecter la situation particulière du répondant. Cela s'impose plus précisément lorsque le répondant ne se limite pas à la question demandée, mais fournit des renseignements supplémentaires. Comme l'ont souligné Oksenberg et coll. (1992, 1995), afin de ne pas donner l'impression de ne pas porter attention à la réponse, les intervieweurs inscrivent souvent la réponse eux-mêmes sans poser la question ou ne posent la question qu'en partie. Ils adoptent donc un comportement axé davantage sur le répondant afin de ne pas paraître insensibles. Une conception des écrans fondée sur une grille et un ordre des questions fondé sur un thème permettent à l'intervieweur de respecter ces règles de la conversation et de réagir en fonction de la situation. Cela pourrait être acceptable ou même préférable, en présence de questions factuelles, pourvu que le comportement de l'intervieweur ne nuise pas à la qualité des données (par exemple questions tendancieuses).

Ce qu'il faut faire afin d'améliorer un questionnaire assisté par ordinateur du point de vue de l'interaction intervieweur-répondant: nos données indiquent que la version grille/thème suscite un déroulement particulier de l'interview auquel l'intervieweur et le répondant peuvent s'adapter facilement. Jeff Moore (1996; Moore et Moyer 1998a, 1998b) a indiqué que les intervieweurs préfèrent la version axée sur un thème. Par contre, nous connaissons mal la satisfaction des répondants à l'égard de cet ordre des questions. Il est donc important d'évaluer leur opinion au sujet de la version différente. De plus, nous ne savons pas si cette version correspond à la façon dont l'information est mémorisée par les répondants. Il se peut que les répondants puissent facilement s'adapter à cette version, mais que d'un

point de vue cognitif et social (fardeau) ou du point de vue de l'exactitude des réponses, ce ne soit pas la bonne méthode. De plus, il y a lieu de déterminer si nos résultats se laissent transposer d'une liste des membres d'un ménage à d'autres segments d'un questionnaire. Nous menons actuellement des essais sur le terrain afin de comparer différents modèles en fonction d'informations factuelles dans un contexte autre que celui d'une liste des membres d'un ménage et pour des catégories d'attitudes. Les versions étudiées dans la présente expérience comportent différents degrés d'information contextuelle fournis à l'intervieweur dans l'exécution d'un questionnaire (questions précédentes, questions suivantes, *etc.*). La question générale est celle-ci: Que se passe-t-il si nous avons davantage recours à des grilles ou à des écrans axés sur la forme? Dans quelles circonstances est-il utile d'améliorer l'efficacité des interviews et quelles sont les limites d'une telle stratégie? Il est cependant trop tôt pour présenter des résultats.

Il y a également lieu d'aborder des questions restées sans réponse à l'aide de recherches plus poussées. Personnellement, j'aimerais proposer une stratégie particulière servant à évaluer ce genre de question à supposer que la conception de questionnaires assistés par ordinateur ait de l'importance pour différents clients: chercheurs, intervieweurs et répondants. Bien entendu, il est important qu'un questionnaire IAO réponde aux besoins du chercheur et que le processus question-réponse soit bien conçu dans chaque cas. Toutefois, à mon avis, il est également important de tenir compte de la dimension sociale de l'interaction intervieweur-répondant et des comportements qui surviennent entre les questions. Si un questionnaire IAO perturbe la dimension sociale du processus de mesure, il risque de nuire à la qualité même des données. Nous ne savons toujours pas quelle stratégie entraîne les meilleurs compromis entre un processus de mesure valable et fiable d'une part, et une interview qui se déroule doucement, brièvement et sans embarras d'autre part. Afin de déterminer dans quelle mesure une conception particulière des écrans IAO pourrait nuire à la qualité des données et comment elle pourrait aider à épargner du temps, de l'argent et des efforts de la part de l'intervieweur, nous devons mener d'autres études de facilité d'utilisation.

Afin d'évaluer les questions mentionnées ci-dessus, nous avons vraiment besoin d'autres expériences sur le terrain. Puisque nous voulons analyser la dimension sociale de l'interview et ses effets sur le comportement de l'intervieweur de même que sur la durée de l'interview, les expériences de laboratoire ne répondent pas à tous nos besoins. Bien sûr, les expériences de laboratoire se font dans un contexte plus contrôlé, fournissent des renseignements plus détaillés au sujet des participants et, par conséquent, exigent un moins grand nombre de cas. Et pourtant, sans essais sur le terrain, nous ne confronterons jamais nos prototypes et nos solutions en fonction de la véritable interaction intervieweur-répondant et aussi du processus question-réponse. On devrait donc étudier la facilité d'utilisation à l'aide d'une combinaison d'expériences de laboratoire et d'essais sur le terrain.

REMERCIEMENTS

Une partie des résultats a été présentée à un colloque tenu à l'Institute for Social Research de l'Université du Michigan le 21 mai 1998, et à l'occasion de la 54e assemblée annuelle de l'American Association for Public Opinion Research le 16 mai 1999. Nous remercions les intervieweurs qui ont participé à la présente expérience. Mick Couper, Siegfried Lamnek, Jeffrey Moore et deux examinateurs anonymes ont fourni des suggestions utiles au sujet de versions antérieures du présent exposé.

BIBLIOGRAPHIE

BAKER, R.P. (1992). New technology in survey research: computer-assisted personal interviewing (CAPI). *Social Science Computer Review*, 10, 145-157.

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

COUPER, M.P., et BURT, G. (1994). Interviewer attitudes toward computer-assisted personal interviewing (CAPI). *Social Science Computer Review*, 12, 38-54.

COUPER, M.P., GROVES, R.M., et KOSARY, C. (1989). Methodological issues in CAPI. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 349-354.

COUPER, M.P., FUCHS, M., HANSEN, S.E., et SPARKS, P.(1997). CAPI Instrument Design for the Consumer Expenditure (CE) Quarterly Interview Survey. Rapport final. University of Michigan.

FUCHS, M. (1994) *Umfrageforschung mit Telefon und Computer*. Einführung in die computergestützte telefonische Befragung. Weinheim: Psychologie Verlags Union.

FUCHS, M. (1995). Die computergestützte telefonische Befragung. Einige Antworten auf Probleme der Umfrageforschung. *Zeitschrift für Soziologie*, 24, 284-299.

FUCHS, M. (2001). The impact of technology on interaction in computer-assisted interviews. (Éd. D. W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, H. van der Zouwen). *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*. Wiley (à paraître).

FUCHS, M., COUPER, M., et HANSEN, S. (2000). Technology effects: Do CAPI or PAPI interviews take longer? *Journal of Official Statistics* (sous press).

GROVES, R.M., et MATHIOWETZ, N.A. (1984). Computer assisted telephone interviewing: effect on interviewers and respondents. *Public Opinion Quarterly*, 48, 356-369.

HANSEN, S.E., COUPER, M.P., et FUCHS, M. (1998). Usability Evaluation of the NHIS Instrument. Article présenté à l' Annual Meeting of the AAPOR, St. Louis, MO.

HOUSE, C.C. (1985). Questionnaire design with computer assisted telephone interviewing. *Journal of Official Statistics*, 1, 209-219.

HOUSE, C.C., et NICHOLLS, W.L. (1988). Questionnaire design for cati: design objectives and methods. (R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, et J. Waksberg, Éds.). *Telephone Survey Methodology*. New York: Wiley. 421-436.

LAURIE, H., et MOON, N. (1997). Converting to CAPI in a Longitudinal Panel Study. Document de travail de l' ESRC Research Centre on Micro-Social Change, 97-11, Essex.

MARTIN, E. (1999). Who knows who lives here? Within-household disagreements as a source of survey coverage error. *Public Opinion Quarterly*, 63, 220-236.

MOORE, J.C. (1996). Person- vs. Topic-based Design for Computer-Assisted Household Survey Instruments. Article présenté à la InterCASIC '96, International Conference on Computer-Assisted Survey Information Collection, San Antonio, TX.

MOORE, J.C., et MOYER, H.L. (1998a). ACS/CATI Person-Based/Topic-Based Field Experiment – Final Report. Center for Survey Methods Research, U.S. Bureau of the Census.

MOORE, J.C., et MOYER, H.L. (1998b). Questionnaire Design Effects on Interview Outcomes. Article présenté à l'Annual Meeting of the AAPOR, St. Louis, MO.

MOYER, L.H. (1996). Which is better: grid listing or grouped questions design for data collection in establishment surveys? *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 986-990.

NICHOLLS, W.L., et de LEEUW, E. (1996). Factors in acceptance of computer-assisted interviewing methods: a conceptual and historic review. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 758-763.

OKSENBERG, L., BEEBE, T., BLIXT, S., et CANNELL, C. (1992). *Research on the Design and Conduct of the National Medical Expenditure Survey Interviews*. Rapport final. Survey Research Center, Ann Arbor, USA.

OKSENBERG, L., CANNELL, C., et BLIXT, S. (1996). Analysis of Interviewer and Respondent Behavior in the Household Survey. U.S. Department of Health and Human Services. AHCPR No. 96-N016.

PROJEKTGRUPPE SOEP (1998). Funktion und Design einer Ergänzungsstichprobe für das Sozio-oeconomische Panel. Diskussionspapiere des DIW, 163, Berlin.

SCHNEID, M. (1991). Einsatz computergestutzter Befragungssysteme in der Bundesrepublik Deutschland. Ergebnisse einer Umfrage. ZUMA-Arbeitsbericht 91/20. Mannheim: ZUMA.

SCHOBER, M.F., et CONRAD, F.G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, 61, 576-602.

SUCHMAN, L., et JORDAN, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association*, 85, 45-54.

WEEKS, M.F. (1992). Computer-assisted survey information collection: A review of CASIC methods and their implications for survey operations. *Journal of Official Statistics*, 8, 445-465.

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article durant l'année 2000. Un astérisque indique que la personne a participé plus d'une fois.

J.-F. Beaumont, *Statistique Canada*
W. Bell, *U.S. Bureau of the Census*
* D.R. Bellhouse, *University of Western Ontario*
Y. Berger, *University of Southampton*
P. Biemer, *Research Triangle Institute*
* D.A. Binder, *Statistique Canada*
D. Cantor, *Westat Inc.*
P. J. Cantwell, *U.S. Bureau of the Census*
R.G. Carter, *Statistique Canada*
J. Chen, *University of Waterloo*
C.Z.F. Clark, *U.S. Bureau of the Census*
M. Cohen, *National Center for Education Statistics*
* J.-C. Deville, *Institut national de la statistique et des études économiques*
* P. Dick, *Statistique Canada*
* J.L. Eltinge, *U.S. Bureau of Labor Statistics*
* W.A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
* M.A. Hidiroglou, *Statistique Canada*
D. Holt, *Office for National Statistics, U.K.*
J.-S. Hwang, *Academia Sincia*
* D. Judkins, *Westat, Inc.*
* G. Kalton, *Westat, Inc.*
S. Kaufman, *National Center for Education Statistics*
J. Kim, *Westat Inc.*
* P.S. Kott, *National Agricultural Statistics Service*
* M. Kovačević, *Statistique Canada*
M. Kramer, *U.S. Bureau of the Census*
P. Lahiri, *University of Nebraska - Lincoln*
N. Laniel, *Statistique Canada*
* M. Latouche, *Statistique Canada*
P. Lavallée, *Statistique Canada*
S. Linacre, *Australian Bureau of Statistics*
* S. Lohr, *Arizona State University*
* H. Mantel, *Statistique Canada*
P. Merkouris, *Statistique Canada*
J. Moloney, *Statistique Canada*
G. Nathan, *Central Bureau of Statistics, Israel*
D. Norris, *Statistique Canada*
J.-S. Pischke, *Massachusetts Instituteof Technolory*
D. Pfeffermann, *Hebrew University*
N. Plante, *L'Institut de la Statistique du Québec*
N.G.N. Prasad, *University of Alberta*
* B. Quenneville, *Statistique Canada*
* E. Rancourt, *Statistique Canada*
* J.N.K. Rao, *Carleton University*
* L.-P. Rivest, *Université Laval*
K. Rust, *Westat Inc.*
I. Sande, *Telcordia Technologies*
N. Schenker, *University of California - Los Angeles*
F.J. Scheuren, *The Urban Institute*
A. Scott, *University of Auckland*
J. Sedransk, *Case Western Reserve University*
* J. Shao, *University of Wisconsin - Madison*
* M.P. Singh, *Statistique Canada*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
R.T. Smith, *National Agricultural Statistics Service*
E. Stasny, *Ohio State University*
* D. Stukel, *Statistique Canada*
A. Théberge, *Statistique Canada*
* Y. Tillé, *Ecole Nationale de la Statistique et de l'Analyse de l'Information*
S. Tremblay, *Statistique Canada*
C. Tucker, *U.S. Bureau of Labor Statistics*
* R. Valliant, *Westat, Inc.*
J. Waksberg, *Westat, Inc.*
M. Wendt, *Statistique Canada*
K.M. Wolter, *National Opinion Research Center*
C. Wu, *University of Waterloo*
* W. Yung, *Statistique Canada*
* E. Zanutto, *University of Pennsylvania*
A. Zaslavsky, *Harvard University*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 2000: J. Beauseigle (Division de la diffusion) et L. Perreault (Division des langues officielles et traduction). Finalement on désire exprimer notre reconnaissance à C. Corbeil, C. Ethier, C. Larabie, D. Lemire et G. Ray de la Division des méthodes des enquêtes auprès des ménages, pour leur apport à la coordination, la dactylographie et la rédaction.

The Canadian Journal of Statistics **La Revue Canadienne de Statistique**

CONTENTS TABLE DES MATIÈRES

Volume 28, No. 2, June/juin 2000, 225-448

James O. RAMSAY
Differential equation models for statistical functions 225

Nancy E. HECKMAN and James O. RAMSAY
Penalized regression with model-based penalties 241

Debbie J. DUPUIS and David C. HAMILTON
Regression residuals and test statistics: assessing naive outlier deletion 259

Martin BILODEAU and Pierre DUCHESNE
Robust estimation of teh SUR model 277

Joris PINKSE
Nonparametric two-step regression estimation when regressors and error are dependent 289

Ursula U. MULLER
Nonparametric regression for threshold data 301

Ronald W. BUTLER and Aparna V. HUZURBAZAR
Bayesian prediction of waiting times in stochastic models 311

Caterina CONIGLIANI, J. Ivan CASTRO and Anthony O'HAGAN
Bayesian assessment of goodness of fit against nonparametric alternatives 327

Caterina CONIGLIANI and Anthony O'HAGAN
Sensitivity of the fractional Bayes factor to prior distributions 343

William J. REED
Reconstructing the history of forest fire frequency:
identifying hazard rate change points using the Bayes information criterion 353

Antonio CUEVAS, Manuel FEBRERO and Ricardo FRAIMAN
Estimating the number of clusters 367

Peter T. KIM and Ja-Yong KOO
Directional mixture models and optimal estimation of the mixing density 383

Dominique FOURDRINIER and Idir OUASSOU
Spherically symmetric distribution with constraints on the norm 399

Serge B. PROVOST and Young-Ho CHEONG
On the distribution of linear combinations of the components of a Dirichlet random vector 417

Michael D. deB. EDWARDES
Implications of random cut-points theory for the Mann-Whitney and binomial tests 427

Martin R. PETERSEN and James A. DEDDENS
Effects of omitting a covariate in Poisson models when the data are balanced 439

Konstantinos FOKIANOS, Amy PENG and Jing QIN
A generalized-moments specification test for the logistic link^2 446

Forthcoming Papers/Articles à paraître 447

The Canadian Journal of Statistics **La Revue Canadienne de Statistique**

CONTENTS TABLE DES MATIÈRES

Volume 28, No. 3, June/juin 2000, 449-672

Feifang HU & John D. KALBFLEISCH: The estimating function bootstrap 449

Discussion:

James V. ZIDEK & Steven X. WANG: Comment 1 482

Thomas J. DICICCIO & Robert J. TIBSHIRANI: Comment 2 485

Christian LÉGER: Comment 3 487

Angelo J. CANTY & Anthony C. DAVISON: Comment 4 489

Stephen M. S. LEE: Comment 5 494

Rejoinder:

Feifang HU & John D. KALBFLEISCH 496

Alan M. POLANSKY: Stabilizing bootstrap-*t* confidence intervals for small samples 501

Steven E. STERN & A. H. WELSH: Likelihood inference for small variance components 517

Francesca DOMINICI, Giovanni PARMIGIANI & Merlise CLYDE:
Conjugate analysis of multivariate normal data with incomplete observations 533

Barbara TONG & Kert VIELE: Smooth estimates of normal mixtures 551

Dianliang DENG & Sudhir R. PAUL:
Score tests for zero inflation in generalized linear models 563

Jiti GAO & Thomas YEE: Adaptive estimation in partially linear autoregressive models 571

Thomas S. FERGUSON, Christian GENEST & Marc HALLIN:
Kendall's tau for serial dependence 587

Min CHEN & Gemai CHEN: Geometric ergodicity of
nonlinear autoregressive models with changing conditional variances 605

Jean-Yves DAUXOIS: Inférence par les martingales pour
des processus ponctuels à compensateur discontinu 615

Bing LI: Nonparametric estimating equations based on a penalized information criterion 621

Peter Xue-Kun SONG: Monte Carlo Kalman filter and
smoothing for multivariate discrete state space models 641

Cheikh A. T. DIACK: Sur la convergence des tests de Schlee et de Yatchew 653

Forthcoming Papers/Articles à paraître 671

Volume 29 (2001): Subscription rates/Frais d'abonnement 672

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 1, 2000

Correcting the Bias in the Range of a Statistic Across Small Areas
David R. Judkins and Jun Liu 1

A Procedure for Stratification by an Extended Ekman Rule
Dan Hedlin 15

A Note on Use of Inverse Sampling: Post Estimation Between Successive Infections
Kung-Jong Lui 31

Optimal Weighting of Index Components: An Application to the Employment Cost Index
Michael K. Lettau and Mark A. Loewenstein 39

Random Selection in a National Telephone Survey: A Comparison of the Kish, Next-Birthday, and Last-Birthday Methods
Diane Binson, Jesse A. Canchola, and Joseph A. Catania 53

The Effect of Different Rotation Patterns on the Revisions of Trend Estimates
David G. Steel and Craig H. McLaren 61

Book and Software Reviews 77

In Other Journals 83

Volume 16, Number 2, 2000

Recent Developments for Poverty Measurement in U.S. Official Statistics
David M. Betson, Constance F. Citro, and Robert T. Michael 87

Nearest Neighbor Imputation for Survey Data
Jiahua Chen and Jun Shao 113

A Note on Jackknife Variance Estimation for the General Regression Estimator
Pierre Duchesne 133

Stratification by Size Revisited
Alan H. Dorfman and Richard Valliant 139

An Estimation File that Incorporates Auxiliary Information
Cary T. Isaki, M.M. Ikeda, J.H. Tsay, and Wayne A. Fuller 155

Large Scale Fitting of Regression Models with ARIMA Errors
Björn Fischer and Christophe Planas 173

Book and Software Reviews 185

Volume 16, Number 3, 2000

Model-Based Alternatives to Trimming Survey Weights
Michael R. Elliott and Roderick J.A. Little 191

Permanent and Collocated Random Number Sampling and the Coverage of Births and Deaths
Lawrence R. Ernst, Richard Valliant, and Robert J. Casady 211

Survey Estimation for Highly Skewed Populations in the Presence of Zeroes
Forough Karlberg 229

The General Application of Significance Editing
David Lawrence and Richard McKenzie 243

Developing Usability Guidelines for AudioCasi Respondents with Limited Literacy Skills
Sid J. Schneider and Brad Edwards 255

Technology Effects: Why Do CAPI Interviews Take Longer?
Marek Fuchs, Mick Couper, and Sue Ellen Hansen 273

Book and Software Reviews 287

Editorial Collaborators 291

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.

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 $\exp(\cdot)$ et $\log(\cdot)$ 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.