Nº 12-001-XIF au catalogue

Techniques d'enquête

Juin 2006

Statistique Canada
Statistics Canada
Canada

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : 613-951-1136).

Pour obtenir des renseignements sur l'ensemble des données de Statistique Canada qui sont disponibles, veuillez composer l'un des numéros sans frais suivants. Vous pouvez également communiquer avec nous par courriel ou visiter notre site Web à www.statcan.ca.

Service national de renseignements	1-800-263-1136
Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
Renseignements concernant le Programme des services de dépôt	1-800-700-1033
Télécopieur pour le Programme des services de dépôt	1-800-889-9734
Renseignements par courriel	infostats@statcan.ca
Site Web	www.statcan.ca

Renseignements pour accéder ou commander le produit

Le produit nº 12-001-XIF au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique Publications.

Ce produit nº 12-001-XPF au catalogue est aussi disponible en version imprimée standard au prix de 23 $CAN l'exemplaire et de 44 $CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 $CAN	12 $CAN
Autres pays	10 $CAN	20 $CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par

- Téléphone (Canada et États-Unis) 1-800-267-6677
- Télécopieur (Canada et États-Unis) 1-877-287-4369
- Courriel infostats@statcan.ca
- Poste Statistique Canada
  Division des finances
  Immeuble R.-H.-Coats, 6e étage
  100, promenade Tunney's Pasture
  Ottawa (Ontario) K1A 0T6
- En personne auprès des agents et librairies autorisés.

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois, et ce, dans la langue officielle de leur choix. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui doivent être observées par les employés lorsqu'ils offrent des services à la clientèle. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées dans le site www.statcan.ca sous À propos de nous > Offrir des services aux Canadiens.

Statistique Canada
Division des méthodes d'enquêtes auprès des entreprises

Techniques d'enquête

Juin 2006

Publication autorisée par le ministre responsable de Statistique Canada

Juillet 2006

Nº 12-001-XIF au catalogue, vol. 32, nº 1
ISSN 1712-5685

Nº 12-001-XPF au catalogue, vol. 32, nº 1
ISSN 0714-0045

Périodicité : semestriel

Ottawa

This publication is available in English upon request (Catalogue no. 12-001-XIE).

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population, les entreprises, les administrations canadiennes et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques précises et actuelles.

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.

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 le faire parvenir en français ou en anglais en format électronique et préférablement en Word au rédacteur en chef, (rte@statcan.ca, Statistique Canada, 150 Promenade du Pré Tunney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue.

Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 $ 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, 30 $ CA (15 $ × 2 exemplaires). Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division de la diffusion, Gestion de la circulation, 150 Promenade du Pré Tunney, 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 statisticiens du Québec. Des versions électroniques sont disponibles sur le site internet de Statistique Canada : www.statcan.ca.

Dédiée à la famille de M.P. Singh : Son épouse, Savitri, et ses enfants, Mala, Mamta et Rahul


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 32, numéro 1, juin 2006

Table des matières


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, juin 2006
Vol. 32, N° 1, pp. 1-2
Statistique Canada, N° 12-001-XIF au catalogue

Dans ce numéro

Ce numéro de la revue *Techniques d'enquête* débute par un article spécial à la mémoire de M.P. Singh, rédacteur en chef fondateur, qui a dirigé la revue pendant 30 ans et en a fait la source internationalement reconnue d'information sur les derniers progrès concernant les techniques d'enquête et les méthodes de production de statistiques officielles qu'elle est aujourd'hui. Un grand nombre de collègues et amis proches qu'a comptés M.P. au fil des ans y partagent leurs souvenirs et évoquent sa carrière et ses contributions.

Dans le premier article ordinaire du présent numéro, Thompson discute de l'utilisation des plans de sondage à marche aléatoire pour l'échantillonnage d'une population réseautée. Il montre comment cette approche peut mener à des échantillons en réseau où les probabilités d'inclusion peuvent être estimées indépendamment de la façon dont l'échantillon initial de noeuds est choisi, ce qui donne des méthodes valides d'inférence fondée sur le plan de sondage. La sélection préférentielle de certains types de noeuds ou caractéristiques des graphes est possible grâce au choix du mécanisme de marche aléatoire. Il décrit des plans de sondage à marche aléatoire uniforme ainsi que ciblée, et présente certains exemples.

Durrant et Skinner examinent le recours à l'imputation et à la pondération pour corriger l'erreur de mesure dans l'estimation d'une fonction de répartition. Ils étudient diverses méthodes d'imputation par le plus proche voisin et d'imputation hot-deck, ainsi que la pondération par le score de propension à répondre sous divers modèles de réponse. Ils discutent des propriétés théoriques de ces méthodes et les comparent au moyen de simulations afin d'estimer la distribution de la rémunération horaire au Royaume-Uni d'après des données provenant de l'Enquête sur la population active. Ils concluent qu'une approche fondée sur l'imputation fractionnaire semble être celle qui, dans l'ensemble, est la plus efficace et la plus robuste.

Harms et Duchesne étudient le problème de l'estimation des quantiles en utilisant des données d'enquête. Ils calent une estimation interpolée d'une fonction de répartition sur des quantiles donnés d'une variable auxiliaire, puis inversent l'estimateur interpolé calé résultant de la fonction de répartition de la variable d'intérêt. Enfin, ils réalisent une étude par simulation afin de comparer leur approche à d'autres méthodes.

Dans leur article, Haziza et Rao proposent une nouvelle méthode d'imputation par la régression avec utilisation des probabilités de réponse. Cette nouvelle méthode mène à des estimateurs valides sous l'approche du modèle de non-réponse ou sous celle du modèle d'imputation. Sous la première approche, le mécanisme de réponse est modélisé paramétriquement et n'est pas limité au modèle de non-réponse uniforme, tandis que sous la seconde, les variables d'intérêt sont modélisées et la non-réponse est considérée comme étant ignorable. Les auteurs fournissent aussi des estimateurs de la variance sous leur méthode d'imputation. Ils présentent, pour l'estimation ponctuelle ainsi que l'estimation de la variance, des résultats de simulation qui témoignent des bonnes propriétés de la méthode proposée d'imputation par la régression.

L'article de Zanutto et Zaslavsky traite du problème de l'estimation dans le cas du recensement décennal de la population des États-Unis sous échantillonnage pour le suivi des non-répondants. Au lieu d'essayer d'obtenir l'information auprès de tous les non-répondants, un échantillon est tiré pour le suivi, ce qui pose un problème d'estimation pour petits domaines. La stratégie proposée consiste à prédire le nombre de ménages non répondants dans diverses catégories au moyen d'un modèle hiérarchique loglinéaire, puis à imputer des renseignements détaillés sur les personnes et les ménages selon la méthode d'imputation par donneur. L'idée, à la première étape, est de modéliser les caractéristiques du ménage en utilisant des covariables peu détaillées à des niveaux détaillés de géographie et des covariables plus détaillées à des niveaux plus élevés d'agrégation géographique. Une étude par simulation indique que les propriétés du modèle proposé se comparent favorablement à celles d'autres modèles.

Dans l'article de Théberge, on propose une nouvelle approche pour répartir l'échantillon de la Contre-vérification des dossiers (CVD) de 2006 qui vise à mesurer le sous-dénombrement du recensement et une partie du sur-dénombrement. Les estimations de la CVD sont utilisées conjointement avec les chiffres du recensement pour produire des estimations démographiques, lesquelles servent à établir les paiements de péréquation du gouvernement fédéral canadien aux provinces. L'approche proposée permet d'établir une répartition qui fournit un équilibre entre quatre objectifs. Elle consiste d'abord à calculer une répartition distincte pour chaque objectif. On prend ensuite pour chaque province la taille d'échantillon maximale sur chacune des répartitions. La répartition infraprovinciale de l'échantillon de la CVD est obtenue en utilisant la technique du calage pour effectuer un lissage de paramètres définis au niveau des strates.

Dans son article, Longford discute de la façon de concevoir une enquête lorsque l'on doit produire des estimations pour plusieurs petits domaines, pour lesquels les priorités varient éventuellement, par minimisation d'une somme pondérée des variances espérées. Il commence par développer ses idées dans le contexte de l'estimation directe, puis les étend à l'estimation composite qui combine l'estimateur direct à un estimateur synthétique. Pour illustrer les méthodes, il présente les résultats, sous diverses hypothèses, de la répartition de l'échantillon d'une enquête auprès des ménages entre les divers cantons suisses.

You et Chapman proposent une approche hiérarchique bayésienne de l'estimation pour petits domaines lorsque les erreurs d'échantillonnage des estimateurs directs sont estimées. Ils démontrent leur approche en produisant des estimations pour petits domaines à partir de deux ensembles de données et étudient sa sensibilité aux hypothèses de modélisation.

Khoshgooyanfard et Monazzah comparent des méthodes d'estimation pour petits domaines basées sur un estimateur synthétique, un estimateur composite et un estimateur empirique bayésien en vue de produire des estimations intercensitaires des taux provinciaux de chômage en Iran. Ils constatent que l'estimateur composite et l'estimateur empirique bayésien produisent l'un et l'autre des résultats satisfaisants.

La brève note de Gabler, Häder et Lynn, qui conclut ce numéro, constitue une extension intéressante de l'article publié antérieurement par Gabler, Häder et Lahiri dans *Techniques d'enquête* (1999). Elle offre une solution pratique au problème de la détermination de l'effet de plan lorsque des échantillons différents sont utilisés pour différents domaines exclusifs.

Enfin, nous tenons à souligner que *Techniques d'enquête* est maintenant disponible en ligne dans un format PDF entièrement interrogeable. Tous les articles publiés dans la revue peuvent désormais être consultés gratuitement en direct sur le site Web de Statistique Canada dès leur diffusion. Nous prévoyons également inclure les numéros antérieurs. Tous les articles parus dans les sept derniers numéros sont déjà mis en ligne et les travaux se poursuivent en vue d'ajouter ceux qui ont été publiés au cours des dix années antérieures. Une version imprimée de la revue continue d'être produite pour les abandonnés. Les anciens numéros peuvent être obtenus sur demande en version imprimée ou en format PDF scanné. La revue peut être consultée sur le site Web de Statistique Canada à l'adresse http://www.statcan.ca/bsolc/francais/bsolc?catno=12-001-X.

Harold Mantel, Rédacteur en chef délégué

Techniques d'enquête, juin 2006
Vol. 32, N° 1, pp. 3-10
Statistique Canada, N° 12-001-XIF

À la mémoire de M.P. Singh

Introduction

Don Royce
Statistique Canada

En août 2005, le monde de la méthodologie d'enquête a perdu l'une de ses figures dominantes avec le décès de M. M.P. Singh à l'âge de 63 ans, quelques mois avant sa retraite planifiée. M.P. et moi avions discuté brièvement de sa retraite à venir, mais il était clair pour nous deux qu'il continuerait d'être le rédacteur en chef de *Techniques d'enquête* même après son départ de Statistique Canada. *Techniques d'enquête* faisait partie de sa vie et j'étais très heureux de lui offrir l'occasion de travailler à temps partiel à partir de sa résidence familiale à Toronto, ce qui lui permettrait de continuer à s'occuper de la revue qu'il avait dirigée depuis plus de 30 ans. Malheureusement, cela n'a jamais eu lieu.

Dans la série d'articles qui suivent, nombre de collègues et amis (les deux sont synonymes) les plus proches de M.P. se rappellent de lui comme statisticien, rédacteur en chef, collaborateur, leader et être humain. Je suis profondément reconnaissant à Eric Rancourt de Statistique Canada d'avoir proposé cette série d'articles, et à tous les auteurs qui, par leur temps et leur talent, ont partagé leurs souvenirs de M.P. Singh. Même si les mots ne peuvent jamais traduire complètement l'essence d'une personne, les articles qui suivent décrivent merveilleusement bien la vie de M.P. Singh et nous rappellent l'héritage qu'il laisse à tous ceux qui ont eu la chance de le connaître. Nous espérons que M.P. en aurait été heureux.

Quelques souvenirs

J.N.K. Rao
Université Carleton, Ottawa

Ma première rencontre avec Mangala Prasad Singh (amicalement connu de beaucoup comme M.P.) remonte à 1968 lorsque j'étais professeur invité à l'Indian Statistical Institute (ISI), à Calcutta. M.P. poursuivait son doctorat à l'ISI sous la supervision de M.N. Murthy. Pendant son doctorat, il a aussi travaillé au National Sample Survey (NSS) en Inde. Le NSS était situé sur le campus de l'ISI et M.P. y a travaillé sous la direction de statisticiens d'enquête renommés au NSS et à l'ISI, dont P.C. Mahalanobis, D.B. Lahiri et M.N. Murthy. Il a ainsi reçu une solide formation en conception et en théorie d'enquêtes par sondage. M.P. a fait un bon usage de cette solide formation pendant toute sa célèbre carrière en appliquant les principes d'une conception efficace sous réserve des questions de coût et d'opération, et en insistant sur l'importance d'une théorie robuste avant de mettre en œuvre de nouveaux plans d'enquête ou de réviser des enquêtes permanentes comme l'Enquête sur la population active du Canada (EPA).

Une grande partie de la thèse de M.P. Singh portait sur l'utilisation efficace des renseignements auxiliaires. Il a étudié deux variables auxiliaires, une en corrélation positive et l'autre en corrélation négative avec la variable d'intérêt, et il a conçu des estimateurs de totaux ratio-produit (ratio-cum-product). Murthy (1967), dans son ouvrage bien connu sur l'échantillonnage, a consacré une section aux estimateurs ratio-produit. M.P. a publié plusieurs documents sur l'utilisation efficace des renseignements auxiliaires développée dans sa thèse : les estimateurs ratio-produit (*Metrika* 1967; *Sankhyā* 1969), l'estimation multidimensionnelle de produits (*Journal of the Indian Society of Agricultural Statistics* 1967) et l'échantillonnage systématique dans l'estimation des ratios et produits (*Metrika* 1967). Il a également publié un important article dans les *Annals of Statistics* (1967) sur l'efficacité relative des stratégies d'échantillonnage à deux phases dans un modèle de superpopulation. La première phase consistait en un échantillonnage aléatoire simple servant à recueillir des données sur une variable auxiliaire x utilisée dans la deuxième phase pour choisir un échantillon PPT sans remplacement et collecter des données sur la variable d'intérêt y.

Au moment de ma visite à l'ISI, M.P. explorait aussi des questions d'inférence en échantillonnage et il faisait face à des problèmes techniques pour démontrer l'admissibilité de certains estimateurs. En effet, un estimateur est admissible dans une classe d'estimateurs non biaisés si aucun autre estimateur de la classe n'est uniformément plus efficace. Malheureusement, le critère d'admissibilité n'est pas suffisamment sélectif et, pour cette raison, la documentation statistique proposait comme choix unique d'autres critères liés à l'admissibilité. Comme je m'intéressais aussi aux questions d'inférence à ce moment-là, nous avons commencé à travailler ensemble sur l'admissibilité des estimateurs. Il a intégré à sa thèse de doctorat le résultat de notre travail. À la fin, nous avons publié un document reposant sur ce travail dans l'*Australian Journal of Statistics* (1973) qui était fondé

sur notre rapport technique de 1969 à l'ISI. Nos résultats démontraient qu'il n'était pas pratique d'utiliser un critère appelé hyperadmissibilité, qui mène à l'estimateur Horvitz-Thompson (HT) du total comme choix *unique* dans *n'importe quel* plan d'échantillonnage. D. Basu obtiendra ultérieurement des résultats similaires de son côté dans son document fondamental de 1971 sur les questions d'inférence, et son fameux exemple des éléphants de cirque a mis un terme à la recherche sur les critères irréalistes qui mènent à un plan d'échantillonnage basé sur un choix unique. M.P. a aussi démontré que l'application de l'hyperadmissibilité à l'estimation de la variance donnait comme choix unique un « mauvais » estimateur de la variance.

Peu de temps après avoir joint les rangs de Statistique Canada en 1970 à titre de méthodologiste, M.P. a participé activement à la révision de l'EPA qui a débouché sur plusieurs innovations. M.P. a proposé l'utilisation systématique de l'échantillonnage PPT sans remplacement avec randomisation initiale pour la sélection des unités primaires à partir des unités non autoreprésentatives (UNAR) et la méthode des groupes aléatoires avec une unité primaire provenant de chaque groupe aléatoire prélevé par échantillonnage PPT auprès des unités autoreprésentatives (UAR). Dans les années 60, j'avais étudié la théorie de ces méthodes du point de vue de leur efficacité et de l'estimation de la variance. De son côté, M.P. a reconnu leurs avantages pratiques dans le contexte de l'EPA. L'échantillonnage systématique PPT et la méthode des groupes aléatoires autorisaient une expansion de l'échantillon de même qu'un renouvellement plus facile des unités primaires d'échantillonnage dans le temps, tandis que la méthode des groupes aléatoires permettait d'adapter la méthode ingénieuse de Keyfitz pour modifier les mesures de taille périmées dans chaque groupe aléatoire. Il publia une communication dans *Metrika* (1975) conjointement avec Dick Platek sur la mise à jour des mesures de taille. Sous l'habile direction de M.P., le groupe de l'EPA apporta plusieurs améliorations méthodologiques à l'efficacité du plan et à l'estimation. Étant donné que M.P. s'était intéressé à l'utilisation efficace des renseignements auxiliaires, l'EPA adopta l'estimation par régression généralisée pour tenir compte de plusieurs variables pour la stratification a posteriori. Le groupe de l'EPA a aussi été le premier à admettre les mérites de l'estimation de la variance par ré-échantillonnage, et on adopta la méthode du jackknife pour l'estimation de la variance. Plus récemment, sous la direction de M.P., on a instauré dans l'EPA l'estimation composite par régression en s'inspirant d'une méthode proposée par Wayne Fuller et moi-même, qui est utile à la fois pour l'estimation du changement et du niveau. Cette méthode de même qu'une méthode antérieure d'Avi Singh s'harmonisent bien avec le système actuel d'estimations de l'EPA qui repose sur la régression généralisée. Trois communications sur l'estimation composite par régression pour l'EPA, notamment une communication de M.P., Jack Gambino et Brian Kennedy, ont paru dans le numéro de juin 2001 de *Techniques d'enquête*.

Depuis 1976, M.P. s'intéressait aussi vivement à l'estimation des données régionales. Son équipe a fait d'importantes contributions méthodologiques aux estimations des données régionales. M.P. et ses collègues proposèrent des estimateurs synthétiques simples de même qu'un nouvel estimateur appelé l'estimateur dépendant de l'échantillon. Cet estimateur est un estimateur composite simple dont les coefficients de pondération s'appliquent aux tailles d'échantillon réalisées qui sont inférieures aux tailles d'échantillon prévues dans les régions. Les estimateurs dépendants de l'échantillon sont alors devenus assez connus et bon nombre d'organismes dans le monde les ont utilisés. En 1994, M.P. publia dans *Techniques d'enquête*, avec Jack Gambino et Harold Mantel, un document expliquant plusieurs questions pratiques liées à l'estimation des données régionales. J'aime tout particulièrement la section sur les questions de plan de sondage. Elle illustre à merveille le compromis qu'a fait l'EPA en ce qui concerne la répartition de l'échantillon pour satisfaire aux critères de fiabilité tant à l'échelle provinciale qu'infraprovinciale. Un chapitre de mon ouvrage sur l'estimation des données régionales (Rao 2003) est consacré aux questions de planification, lesquelles reposent fortement sur ce document de 1994. M.P. participa activement à l'organisation d'une conférence internationale fort réussie en 1985 sur les estimations des données régionales et il rédigea en collaboration en 1987 chez Wiley un ouvrage intitulé *Small Area Statistics*, qui s'inspire des communications sollicitées à la conférence.

M.P. adorait son travail de rédacteur en chef à *Techniques d'enquête*. Il a maintenu des liens étroits avec son équipe de rédacteurs associés, proposant beaucoup de nouvelles idées dont des exposés thématiques tant sur la théorie que sur la pratique, de même que la série d'articles Waksberg. Les déjeuners-causeries organisés chaque année par M.P. aux *Joint Statistical Meetings* ont toujours connu un grand succès auprès des rédacteurs associés! À titre de rédacteur associé à Ottawa et de consultant pour Statistique Canada, j'ai souvent abordé avec M.P. les problèmes auxquels faisait face la revue pendant les 25 dernières années. M.P. a aussi joué un rôle actif à la Société statistique du Canada (SSC) et il a fait la promotion de la théorie de l'échantillonnage aux assemblées générales annuelles de la SSC.

M.P. était un chiromancien d'une remarquable précision; en 1999, il m'avait mis en garde au sujet d'éventuels problèmes de santé. De fait un problème de santé imprévu est survenu en 2001 en raison de complications à la suite d'une appendicite. Quelques mois avant son décès, Avi Singh m'a

dit que M.P. avait lu dans sa propre main la fin de ses graves problèmes de santé. Avi et moi étions certains de revoir M.P. au travail. Toutefois, c'est une croyance populaire en Inde que ceux qui lisent dans leur propre main ne peuvent prévoir leur avenir avec précision. Malheureusement, cette croyance s'est avérée juste dans ce cas-ci.

M.P. était vraiment un grand ami et il me manquera beaucoup. Il convient que ses cendres aient été répandues dans la fleuve sacré Gange, dans la ville la plus sainte pour les Hindous, Varanasi (aussi appelée Bénarès), où était né M.P. Son âme est au Ciel mais son héritage demeurera avec nous.

M.P. du temps où il était chercheur

T.J. Rao
Indian Statistical Institute, Kolkata

J'ai rencontré M.P. pour la première fois lorsqu'il a assisté au quatrième cours d'été (avancé) pour statisticiens organisé par la Research and Training School (RTS) de l'Indian Statistical Institute (ISI) en mai et en juin 1964 à l'Université de Kerala située dans la ville de Trivandrum (maintenant appelée Thiruvananthapuram) dans le sud de l'Inde. Ce cours était destiné aux chercheurs et aux professeurs débutants de l'ISI et d'autres universités. M.P. venait de l'Université Banaras Hindu (BHU) où il était un chargé de cours temporaire. Il a obtenu un baccalauréat en statistique de la même université (BHU) et une maîtrise de l'Université de Poona. J'étais parmi les chercheurs qui ont été sélectionnés par l'ISI pour participer à ce cours. Nous n'avons pas beaucoup interagi pendant le cours.

Un peu plus tard, M.P. a reçu une offre d'emploi de la Division de l'échantillonnage du département National Sample Survey (NSS), qui faisait partie de l'ISI à l'époque. Les professeurs D.B. Lahiri, S. Rajarao et M.N. Murthy dirigeaient déjà alors plusieurs divisions du NSS. En plus de travailler à la conception d'enquêtes par sondage à grande échelle réalisées par le NSS, M.P. passait son temps libre à examiner des problèmes de recherche liés aux enquêtes par sondage. Lahiri et Murthy encourageaient la recherche méthodologique au NSS et ont commencé à organiser une série de séminaires ainsi qu'à diffuser des rapports techniques semblables aux rapports techniques de la RTS de l'ISI. M.P. et moi avons discuté de notre recherche sur les problèmes d'échantillonnage au cours des séminaires organisés par le NSS et la RTS. La majeure partie des travaux de M.P., qu'il a converti en rapports techniques pour la série du NSS, ont été publiés plus tard dans des revues scientifiques réputées.

En s'appuyant sur l'expertise qu'il avait acquise au NSS en travaillant sur des enquêtes polyvalentes, il s'est intéressé aux problèmes liés à l'utilisation de données auxiliaires dans des enquêtes par sondage. Ses premiers travaux ont porté sur l'estimation par les méthodes du quotient et du produit. M.P. a étudié intelligemment et avec succès le cas de multiples variables auxiliaires dont certaines étaient corrélées positivement et certaines étaient corrélées négativement avec la variable étudiée. Il a utilisé les estimateurs par quotient pour les variables corrélées positivement et les estimateurs par produit pour les variables corrélés négativement et a rédigé le *Ratio cum product estimator* (Singh 1967). Ce document est souvent cité et plusieurs chercheurs, en particulier de l'Inde, ont publié des suppléments. Conjointement avec M.N. Murthy, il a élaboré des concepts intéressants sur l'admissibilité des estimateurs (Murthy et Singh 1969). Au cours de l'année 1968, le professeur J.N.K. Rao a visité l'ISI et nous avons été très chanceux d'interagir avec lui.

M.P. aimait beaucoup assister à des conférences. Il n'en a jamais manqué une à son alma mater BHU ou au Indian Science Congress. Il a entrepris la rédaction de sa thèse avec beaucoup de sérieux et il aimait discuter avec les professeurs M.N. Murthy, J.N.K. Rao et D. Basu. Il a présenté ses travaux de recherche comme thèse (Singh 1969) pour obtenir un doctorat en philosophie (Ph.D.) de l'Indian Statistical Institute en 1969 sous la direction de M.N. Murthy. Il a quitté le NSS et l'ISI en 1970 pour se joindre à Statistique Canada.

Il manque beaucoup à tous les chercheurs qui étaient à l'ISI entre 1965 et 1970 et à ses collègues du NSS.

Bibliographie

Singh, M.P. (1967). Ratio cum product method of estimation. *Metrika*, 12, 34-42.

Singh, M.P. (1969). *Some aspects of estimation in sampling from finite populations*. Thèse de doctorat, soumise à l'Indian Statistical Institute.

Murthy, M.N., et Singh, M.P. (1969). On the concepts of best and admissible estimators in sampling theory. *Sankhyā*, 31, 343-354.

M.P. Singh

Nanjamma Chinnappa
Statistique Canada (retraitée)

Comme beaucoup d'entre vous connaissent M.P., le statisticien, ainsi que ses réalisations en statistique, je vais essayer de vous parler de M.P., l'homme.

Je n'avais jamais rencontré M.P. avant mon arrivée au Canada, même si j'avais entendu dire qu'il était le jeune homme qui avait été nommé à mon poste lorsque j'ai démissionné de mon emploi au département National Sample Survey (NSS) de l'Indian Statistical Institute de Kolkata, en Inde. J'ai appris que, lorsque M.N. Murthy (alors chef du secteur de la méthodologie du NSS) m'a envoyé l'ébauche de son livre *Sampling Theory and Methods* pour que je la lise, M.P. est celui qui a lu mes commentaires et en a discuté avec M.N. Murthy. Beaucoup plus tard, lorsque M.N. Murthy a appris que j'avais été embauché par Statistique Canada, il m'a

donné le numéro de téléphone de M.P. à Ottawa. Alors, quand nous sommes arrivés à Ottawa, j'ai appelé M.P. de l'hôtel où nous restions et, à ma surprise, il est venu me chercher par un matin froid et humide de la fin de septembre et m'a emmené à Statistique Canada. Ce geste chaleureux et amical a égayé ma journée et mon arrivée à Statistique Canada.

M.P. était originaire de la ville ancienne de Bénarès, en Inde, et il semble que certaines des qualités qui ont rendu cette ville célèbre avaient déteint sur lui. Il était doux, gentil, imperturbable, tenace et sage. Beaucoup de gens m'ont dit qu'il n'était jamais trop occupé pour écouter leurs problèmes et qu'il les aidait toujours avec des paroles bienveillantes et des suggestions. Beaucoup de jeunes statisticiens ont profité de ses conseils concernant leurs recherches et leur carrière.

M.P. aimait la musique et la danse indienne classique. Sa famille comptait beaucoup pour lui. Il était le pilier sur lequel s'appuyaient sa femme et ses enfants lorsqu'ils éprouvaient des difficultés. Lors des rencontres sociales, il avait toujours beaucoup de plaisir et riait énormément. Et lorsqu'il est tombé gravement malade il y a quelques années, il m'a dit que c'était sa foi en Dieu et en lui-même qui l'avait aidé à se rétablir. On se souviendra longtemps de lui, non seulement comme un statisticien de renom, mais aussi comme d'un brave homme qui s'est lié d'amitié avec beaucoup de gens et qui en a aidé plus d'un.

Une carrière en méthodologie d'enquête

Gordon Brackstone
Statistique Canada (retraité)

Presque toute la carrière de M.P. Singh s'est déroulée dans le secteur de la méthodologie de Statistique Canada. Il a joint l'organisme en 1970, après avoir obtenu un doctorat en échantillonnage de l'Indian Statistical Institute. Au moment de son décès, il était directeur de la Division des méthodes d'enquêtes auprès des ménages à la Direction de la méthodologie. Sa montée dans l'organisme a été constante plutôt que vertigineuse : chef de section en 1973, directeur adjoint en 1982 puis directeur en 1994. Cette progression constante reflète bien son approche de la méthodologie d'enquête qui privilégiait le souci du détail dans la recherche et les essais afin d'établir de solides fondations pour la mise en œuvre et les améliorations futures.

Nos carrières à Statistique Canada ont coïncidé, à une année près au début ou à la fin, et elles se sont souvent croisées, particulièrement à partir de 1982. Au début des années 80, lorsque nous avons senti la nécessité d'améliorer l'intégration et la surveillance de la recherche méthodologique à Statistique Canada, j'étais à peu près certain qu'on demanderait à M.P. de diriger ce travail, et il a été dûment désigné comme président du Comité de recherche sur la méthodologie. À ce poste jusqu'en 1987, il a mis en place les processus de planification et les critères de déclaration qui, améliorés par ses successeurs, ont régi la gestion de la recherche méthodologique pendant deux décennies. C'est au cours de cette même période que Statistique Canada inaugura les symposiums sur la méthodologie, M.P. jouant un rôle clé dans plusieurs des premiers symposiums (et dans beaucoup d'autres par la suite).

Au cours de sa longue carrière à Statistique Canada, M.P. a participé à une vaste gamme de travaux méthodologiques, mais on associera toujours son nom de façon plus immédiate à deux projets : la conception de l'Enquête sur la population active du Canada (EPA) et la fonction de rédacteur en chef de la revue *Techniques d'enquête*.

L'EPA est la base du programme des enquêtes auprès des ménages de Statistique Canada. Non seulement est-elle la source des estimations mensuelles sur les conditions du marché du travail au Canada, mais sa base de sondage est aussi la base d'échantillonnage de nombreuses autres enquêtes auprès des ménages, dont plusieurs enquêtes longitudinales des années 90. Sa conception efficace est donc cruciale à la rentabilité du programme de statistiques sociales du Canada. D'abord inaugurée en 1945, l'EPA a typiquement subi au moins une révision de son échantillon après chaque recensement décennal. M.P. a joint Statistique Canada juste à temps pour la révision majeure qui a suivi le recensement de 1971. Cette révision incluait non seulement le plan de sondage mais aussi le questionnaire, les méthodes de collecte et les systèmes de traitement. Une révision si importante nécessitait une vaste collaboration interdisciplinaire et M.P. a été un intervenant clé dans les aspects méthodologiques de cette révision. Ses articles de cette période traitent surtout de l'optimisation du plan à plusieurs degrés et de la mise à jour de l'échantillon. Il est coauteur de la description officielle de la méthodologie de l'Enquête sur la population active (Platek et Singh 1976).

À la suite de cette révision, les pressions pour la production d'estimations régionales du marché du travail s'accrurent et amenèrent M.P. à élaborer des méthodes d'estimation de données régionales à partir de l'EPA (Drew, Singh et Choudhry 1982). Au moment de la révision prévue après le recensement de 1981, M.P. était devenu le président du comité chargé de superviser le processus complet de révision. En plus des objectifs habituels d'efficacité de l'échantillonnage, cette révision avait pour but de produire de meilleures données infraprovinciales et d'améliorer le rôle de l'EPA comme véhicule pour la réalisation d'autres enquêtes auprès des ménages. Naturellement, M.P. a encore été un des auteurs principaux derrière la description du nouveau plan de sondage (Statistique Canada 1990).

Les efforts déployés pour faire de l'EPA la base d'autres enquêtes auprès des ménages connurent un tel succès, qu'un problème de surcharge survint à la fin des années 90. Avec l'ajout des enquêtes longitudinales et des enquêtes sur la santé au programme d'enquêtes régulier, on se préoccupa davantage du fardeau de réponse. De plus, on sentait le besoin d'avoir des bases de sondage plus ciblées pour certaines sous-populations. M.P., chercha alors d'autres solutions dont certaines axées sur le registre des adresses élaboré aux fins de recensement. Quelques-unes de ces approches ont été intégrées dans la révision de l'EPA après le recensement de 2001, révision amorcée au moment de son décès; d'autres idées plus ambitieuses pour une nouvelle base de sondage aux enquêtes auprès des ménages sont toujours à l'étude par ses successeurs.

Pendant plus de 30 ans, M.P. a orienté les travaux méthodologique à l'EPA. Ses nombreux articles, souvent rédigés en collaboration avec son personnel, témoignent de son influence permanente sur la conception de cette enquête phare et sur l'évolution qu'ont connue de nombreux jeunes statisticiens au début de leur carrière.

Au cours de cette même période, M.P. a aussi assumé une autre lourde responsabilité soit celle de rédacteur en chef de *Techniques d'enquête*. L'évolution de cette revue, de sa naissance en 1975 jusqu'à son 25e anniversaire, a été décrite par son fondateur, Richard Platek (1999), qui avait eu la clairvoyance de nommer M.P. comme son premier rédacteur en chef.

Sous le leadership de M.P., la revue a franchi beaucoup d'étapes importantes. En 1982, elle est devenue une publication officielle de Statistique Canada – entièrement bilingue et tarifée. On invita des auteurs de l'extérieur de Statistique Canada; un panel hautement qualifié de rédacteurs associés fut recruté; on présenta des numéros thématiques, qui attirèrent souvent les meilleurs articles d'une récente conférence ou symposium; on institua la rubrique *Dans ce numéro* où le rédacteur en chef présentait une vue d'ensemble du contenu; des numéros spéciaux du 25e anniversaire parurent en 1999–2000, accompagnés d'un index des volumes 1 à 26. Pendant cette période, on offrit, d'abord à l'Association internationale de statisticiens d'enquêtes puis d'autres organismes statistiques, des abonnements à prix réduit. Plus récemment, des versions électroniques de la revue sont devenues disponibles.

Au cours de cette période, M.P. a tenu la barre, planifiant les numéros à venir, à l'affût de travaux intéressants dignes d'être inclus dans la revue, encourageant des auteurs potentiels, recrutant et harcelant les rédacteurs associés au cours du processus d'examen, travaillant avec le personnel de la publication et du marketing de Statistique Canada pour améliorer la revue et en faire la promotion. En tant que membre du Conseil de gestion de la revue de 1987 à 2004, j'ai été à même de constater et d'admirer son enthousiasme et sa persévérance face à de nombreuses difficultés. C'était pour lui, je crois, une œuvre d'amour.

Ces brèves descriptions de seulement deux des multiples contributions de M.P. à Statistique Canada et à la profession statistique ne peuvent rendre pleinement justice à sa carrière. J'espère qu'elles donnent l'image d'un professionnel sur qui on pouvait toujours compter, qui savait allier une capacité de compréhension profonde et de recherche des méthodes statistiques à une connaissance des contraintes pratiques que comporte l'application des méthodes statistiques aux enquêtes. Son style s'appuyait sur la raison et la persistance, sans éclat et dans l'acceptation des confrontations, auxquelles s'ajoutait une préoccupation innée pour les sentiments des autres. J'ai toujours pris plaisir à travailler avec M.P. et je suis honoré d'être associé à ses réalisations.

Bibliographie

Drew, D., Singh, M.P. et Choudhry, H. (1982). Évaluation des techniques d'estimation pour les petites regions dans l'enquête sur la population active du Canada. *Techniques d'enquête*, 8, 19-52.

Platek, R., et Singh, M.P. (1976). *Methodology of the Canadian Labour Force Survey*, Statistique Canada, numéro de catalogue 71-526.

Platek, R. (1999). Techniques d'enquête – 25 ans d'histoire. *Techniques d'enquête*, 25, 123-125.

Statistique Canada (1990). *Methodology of the Canadian Labour Force Survey 1984-1990*, Statistique Canada, numéro de catalogue 71-526.

À la mémoire de M.P. Singh

Fritz Scheuren
président de
l'American Statistical Association pour 2005

Avec le décès l'été dernier de M.P. Singh, la communauté statistique entière perd un érudit, un homme d'honneur et un homme d'action. C'est en ces termes que j'ai parlé de lui au Symposium sur la méthodologie de Statistique Canada à l'automne 2005.

Toutefois, je serai bref, donnant seulement un exemple de ce qui pourrait être dit. D'autres parlant aussi de lui. Ils en diront plus.

Mes souvenirs de M.P. remontent à plus de 20 ans. Je ne me souviens pas exactement à quel moment je l'ai rencontré pour la première fois, mais j'ai été un de ses rédacteurs associés à *Techniques d'enquête* pendant au moins tout ce temps.

Il aimait me faire lire des articles sur le couplage d'enregistrements, quelquefois sur la pondération ou l'estimation, et, moins souvent, sur des sujets reliés aux données manquantes. Ses choix étaient toujours pour moi une occasion d'apprendre. De façon générale, après son premier

examen, la qualité était excellente et, sous sa direction, mon travail consistait à faire en sorte que les numéros de la revue à paraître soient encore meilleurs.

Les remises en question faisaient partie de son travail de rédacteur en chef de *Techniques d'enquête*. La revue devait offrir des formulations statistiques mathématiques très bien soutenues, mais elles devaient aussi pouvoir être mises en application. En d'autres mots, les idées devaient être très bonnes, mais aussi éminemment utiles. Et elles l'ont toujours été. Ce n'est pas un mince exploit.

Beaucoup de jeunes professionnels hors pair, dans leur premier article, démontrent uniquement un de ces aspects, habituellement la dimension mathématique de leur sujet. Selon moi, l'objectif que fixait M.P. à ses rédacteurs associés qui recevaient des articles présentant au moins un de ces aspects était d'aider les auteurs, grâce à l'examen et à nos commentaires, à atteindre le deuxième objectif. C'est tout un journal que sa vision a créé!

À propos, il m'avait confié que j'avais peut-être tendance à en faire trop dans mon rôle de soutien aux auteurs, mais je pense que, secrètement, il était heureux de mon approche de ne jamais abandonner un article qui pouvait devenir extraordinaire, si on était suffisamment patient. Plusieurs articles dont je me suis occupé ont éprouvé sa patience mais l'ont récompensé en bout de ligne.

M.P. avait une ténacité qui complétait son inépuisable bonté. Sa direction ferme et sûre de *Techniques d'enquête* nous obligeait tous à respecter des normes élevées. Même quand sa santé a commencé à décliner, son esprit est toujours demeuré manifeste.

Le mot que j'ai utilisé pour caractériser M.P. à la conférence de l'automne dernier était celui de « Mensch ». Ce mot allemand désignant une « personne » peut être familier pour nombre d'entre vous dans son sens yiddish d'un être humain complet ou entier. Mais en réalité, le mot « Mensch » ne peut vraiment pas se traduire. C'est pourquoi je l'ai laissé en yiddish ici (même si je n'ai pas utilisé de caractères hébreux, ce qui aurait été approprié). Il n'y a certainement pas de définition simple qui puisse rendre justice soit au mot, soit à la personne qu'a été M.P.

Il nous manque tous énormément. Il était un bon ami, un homme de famille tendre, ouvert aux idées nouvelles, prudent dans ses conseils pratiques et rigoureux dans sa pensée. Il sera toujours un modèle de ce qu'est un statisticien d'enquête.

Quelques souvenirs de M.P. Singh

David A. Binder
Statistique Canada (retraité)

Je garde de très bons souvenirs de M.P. Singh, que j'ai pendant de nombreuses années. Ses forces, comme statisticien d'enquête exceptionnel et comme personne attentive et aimable, le définissaient dans une classe à part.

J'ai rencontré M.P. Singh pour la première fois à l'été 1970. Je travaillais comme étudiant à la Division de l'agriculture de Statistique Canada, où M.P. Singh et J.C. (John) Koop étaient alors méthodologistes. Je partageais un bureau avec Jack Graham, en congé sabbatique de l'Université Carleton. À ce moment-là, Jack m'avait confié combien Statistique Canada était privilégié d'avoir M.P. et John comme méthodologistes d'enquête, car ils étaient deux des meilleurs statisticiens d'enquête au monde. Des talents si exceptionnels à Statistique Canada m'ont incité à choisir cet endroit pour amorcer ma carrière.

La plupart des gens connaissaient M.P. par les douzaines d'articles qu'il avait publiés, par ses fonctions à la revue *Techniques d'enquête* et par ses interventions aux conférences statistiques. Ses publications comprenaient des articles sur la conception et la révision des enquêtes auprès des ménages, sur l'estimation (dont l'estimation composite et l'estimation par domaine), sur l'estimation des données régionales et sur les ajustements pour la non-réponse. Ses questions et suggestions lors de conférences et de réunions reflétaient la profondeur de sa pensée sur les nombreuses complexités des méthodes d'enquête.

Il a aussi rédigé en collaboration des monographies sur les enquêtes par panel (Kasprzyk *et al*. 1989) et sur les statistiques régionales (Platek *et al*. 1987), et il a écrit un exposé de synthèse sur les *Techniques d'enquête* dans l'*Encyclopedia of Statistical Sciences* (Singh 1988).

Rédacteur en chef de *Techniques d'enquête* depuis sa fondation en 1975, M.P. a supervisé l'évolution de la revue, au début comme l'outil principal de publication des recherches du personnel de Statistique Canada puis comme une revue internationale de pointe à laquelle ont collaboré régulièrement des auteurs de partout dans le monde. La section sur les méthodes de recherche par sondage de l'American Statistical Association puis l'Association internationale des statisticiens d'enquête adoptèrent *Techniques d'enquête* comme publication pour leurs membres. Cela reflète bien les nombreuses années de travail assidu qu'a données M.P. à la revue. On retrouvait sa gentillesse et son attention même dans les commentaires encourageants qu'il faisait lorsqu'il devait écrire une lettre de refus à un auteur!

Au fil des ans, M.P. a été un chef de file dans l'adaptation des enquêtes auprès des ménages à l'évolution de la technologie. Il a toujours cherché des moyens d'améliorer les méthodes de collecte de données. Il a orienté Statistique Canada dans cet univers de l'interview face-à-face, de l'interview téléphonique et des méthodes informatiques. Tout dernièrement, il s'appliquait à élaborer des méthodes plus efficaces, notamment en instaurant le concept d'un échantillon-maître pour la conception d'enquêtes auprès des

ménages à Statistique Canada, et il a aidé à convaincre les gestionnaires de Statistique Canada des mérites potentiels de ce concept.

M.P. a eu une influence majeure à Statistique Canada sur la qualité et le calibre de la recherche en méthodes statistiques. Les réalisations du Bureau dans ce domaine ont été reconnues dans le monde entier et on demande maintenant souvent à Statistique Canada de participer à des activités de recherche, par exemple en présentant des articles à des réunions, en participant à des discussions entre experts et en siégeant à différents comités et groupes consultatifs. Un comité de recherche en méthodologie a été créé en 1982-1983 et M.P. en a été le premier président. C'est là qu'il a aidé à élaborer un programme de recherche et un plan stratégique pour la Direction de la méthodologie. Même si le Programme de recherche a évolué au fil du temps, le Programme de recherche en méthodologie est toujours florissant, grâce à la structure et au soutien de gestion que M.P. a aidé à mettre en place.

Tout au long de ma carrière à Statistique Canada, j'ai pu profiter largement de la présence de M.P. Aux réunions de gestion et aux réunions où il représentait la Direction de la méthodologie, il a toujours veillé à ce que nous conservions nos qualités distinctives comme méthodologistes et à ce que les décisions que nous prenions aient un sens pour notre groupe.

Même avec toutes ses réalisations comme statisticien d'enquête, c'était son tempérament que j'admirais le plus. À mon avis, sa compassion désintéressée pour les autres, quel que soit leur niveau de compétence, était sa plus grande force. Je me souviens d'une occasion où nous étions tous les deux en train d'interviewer à Ottawa un candidat hautement qualifié que nous avions fait venir de très loin. Toutefois, après quelques minutes, il était clair, malgré ses qualifications, que cette personne n'était pas faite pour travailler à la Direction de la méthodologie. Bien que le candidat ait fait un voyage spécial à Ottawa pour l'interview, M.P. a pris le temps de mettre la personne à l'aise en discutant avec elle de sujets familiers, même si M.P. reconnaissait aussi qu'elle n'était pas faite pour la Direction.

M.P. a toujours su faire l'éloge des autres lorsque leurs réalisations étaient dignes de mention. C'est une des nombreuses raisons pour lesquelles beaucoup l'aimaient et pourquoi il manque à plusieurs.

Bibliographie

Kasprzyk, D., Duncan, G., Kalton, G. et Singh, M.P. (Éd.) (1989). *Panel Surveys*. New York : John Wiley & Sons, Inc.

Platek, R., Rao, J.N.K., Särndal, C.-E. et Singh, M.P. (Éd.) (1987). *Small Area Statistics* : *An International Symposium*. New York : John Wiley & Sons, Inc.

Singh, M.P. (1988). Encyclopedia of Statistical Sciences, (Éds. D.L. Banks, Read, B. Campbell et S. Kotz), New York : John Wiley & Sons, Inc. Vol. 9, 109-110.

Gestionnaire et mentor

Jack Gambino
Statistique Canada

D'autres ont écrit sur les contributions importantes et variées de M.P. Singh à la profession statistique et à Statistique Canada. J'ai eu la bonne fortune de travailler étroitement avec M.P. pendant 17 ans et de connaître plusieurs aspects de lui que seuls ceux qui le côtoyaient régulièrement ont vu et apprécié. J'ai vu M.P. dans son rôle de rédacteur en chef à *Techniques d'enquête*, dans ses activités quotidiennes qui menaient à la publication de chaque numéro de la revue, dans son rôle de gestionnaire et dans son rôle de superviseur et de mentor.

Dans les années 80, lorsque j'ai joint Statistique Canada, il était impossible de ne pas rencontrer M.P. Singh. Pendant les premières années, il était pour moi la personne qui posait les questions clés à tous les séminaires de méthodologie auxquels j'assistais. Beaucoup plus tard, lorsque nous siégions ensemble à quelques comités, j'étais toujours fasciné lorsque, pendant les réunions, il posait de bonnes questions sur des sujets qui étaient clairement en dehors du domaine de la méthodologie. Invariablement, ses questions aidaient à clarifier les problèmes, non seulement pour les méthodologistes mais aussi pour tous les participants. Cela m'a fait comprendre que je ne devais pas présumer que j'étais la seule personne à ne pas saisir complètement le sujet de discussion.

M.P., le rédacteur en chef : j'ai commencé à connaître personnellement M.P. lorsque j'ai joint sa sous-division en 1988. Il m'a immédiatement recruté comme rédacteur adjoint de *Techniques d'enquête*. C'était une pratique courante chez M.P. – lorsqu'il rencontrait des personnes avec un bon bagage technique, elles devenaient des rédacteurs adjoints potentiels de la revue. Ceux d'entre nous qui ont été assez chanceux pour occuper un tel poste ont beaucoup appris de l'expérience. Au fil du temps, lorsque M.P. s'est fié à notre jugement, il s'en est remis de plus en plus à notre opinion, par exemple, pour décider du sort d'un article qui avait fait l'objet d'évaluations contradictoires.

M.P., le gestionnaire : son approche envers ses rédacteurs adjoints illustre bien son style de gestion plus général. Il laissait à chacun le soin de faire ses preuves et, sauf de rares exceptions, les capacités de chaque employé se sont développées parallèlement à la confiance que M.P. leur manifestait. Beaucoup de gestionnaires suivent une philosophie de gestion spécifique, quelquefois sautant sur la dernière tendance de gestion, quelle qu'elle soit. M.P. n'était pas ainsi. Gestionnaire intuitif, il avait le don de déceler les futurs « talents » au tout début de leur carrière. Il était aussi un

gestionnaire non autoritaire qui savait encourager son personnel dans son travail. Même si M.P. pratiquait une gestion ouverte et souple, il savait quand faire acte d'autorité, comme beaucoup d'entre nous qui ont travaillé avec lui en ont fait la rude expérience, bien qu'en de rares occasions.

M.P. était un penseur stratégique qui aimait discuter en profondeur de statistique et de gestion. Cela donnait quelquefois lieu à de longues réunions où nous devions tous donner notre point de vue. Au moment où nous pensions que le problème était réglé, M.P. intervenait et la discussion repartait! Évidemment, l'avantage de son approche était qu'à la fin de la réunion, nous comprenions tous les tenants et les aboutissants du sujet et que nous parvenions toujours à un consensus.

Tout au long de sa carrière, M.P. s'est toujours fortement intéressé au perfectionnement des chercheurs et à la recherche à Statistique Canada. Pour lui, un programme de recherche actif était essentiel au succès continu de Statistique Canada. C'est pourquoi il a travaillé à améliorer la visibilité professionnelle des chercheurs et, d'une façon plus générale, des méthodologistes d'enquête, au sein de la Société statistique du Canada et dans d'autres organismes.

M.P., le superviseur et le mentor : après quelques années sous sa direction, j'ai eu la chance de l'avoir comme supérieur direct. Il est impossible de distinguer le superviseur et le mentor chez M.P. Il portait un réel intérêt à la carrière de ses employés immédiats, leur prodiguant des conseils et les orientant vers des choix judicieux ou, beaucoup plus important, les orientant loin des mauvais choix. L'approche qu'il utilisait souvent est intéressante : au lieu d'être direct, il amenait souvent l'employé, d'une façon presque socratique, à comprendre que ce n'était pas une si bonne idée. Une autre technique était celle du « regard » – quiconque l'a bien connu savait d'un simple coup d'œil si M.P. considérait une idée comme particulièrement mauvaise.

M.P. m'a beaucoup appris au sujet des enquêtes mais plus important encore, j'ai appris de lui ce qui fait un bon gestionnaire, un bon motivateur et un bon mentor. Je conçois maintenant que son plus grand rôle était peut-être celui *d'enseignant*. Ceux d'entre nous qui ont travaillé étroitement avec lui au fil des ans continueront à profiter de son exemple pour le reste de leur carrière, et je m'attends à ce que nous léguions à la prochaine génération ce que nous avons appris de lui après l'avoir intégré à notre propre expérience.

En ses propres mots

Eric Rancourt
Statistique Canada

M.P. était un homme doté d'une impressionnante personalité. Beaucoup de ses employés et collègues n'ont pas eu la chance de travailler étroitement avec lui, mais pour ceux dont ce fut le cas, il s'est révélé comme une personne très humaine et très polyvalente. Vous trouverez ci-dessous quelques-unes de ses citations que d'autres et moi-même avons rassemblées. C'était habituellement des phrases encourageantes qui nous incitaient à toujours repartir de son bureau du bon pied!

- Pas besoin d'une réunion, ma porte est toujours ouverte pour discuter de n'importe quoi.
- Il est bon d'avoir un projet de prédilection.
- Nous ne concevons pas les enquêtes pour calculer la variance.
- Je suis certain que cela peut se faire.
- Vous me dites que les deux tiers de vos suggestions ne se retrouvent pas dans l'enquête! Ne vous plaignez pas; si seulement 10 % de vos idées sont mises en application, vous aurez une carrière exceptionnelle.
- Il y a un panneau sur l'autoroute qui indique 100 km/h; cela ne veut pas dire qu'il faut rouler à 100 km/h.
- Ne vous inquiétez pas, il y a encore du temps.
- Après tout le travail que nous mettons à concevoir des enquêtes, ce que nous nous rappelons et apprécions le plus ne sont pas les méthodes ni les résultats; ce sont les personnes avec lesquelles nous avons travaillé.

(Traductions libres).

Techniques d'enquête, juin 2006
Vol. 32, N° 1, pp. 11-26
Statistique Canada, N° 12-001-XIF au catalogue

Plans de sondage à marche aléatoire ciblée

Steven K. Thompson [1]

Résumé

Les populations humaines cachées, Internet et d'autres structures en réseau conceptualisées mathématiquement sous forme de graphes sont intrinsèquement difficiles à échantillonner par les moyens conventionnels et les plans d'étude les plus efficaces comportent habituellement des procédures de sélection de l'échantillon par suivi adaptatif des liens reliant un nœud à un autre. Les données d'échantillon obtenues dans le cadre de telles études ne sont généralement pas représentatives au pied de la lettre de la population d'intérêt dans son ensemble. Cependant, un certain nombre de méthodes fondées sur le plan de sondage ou sur un modèle sont maintenant disponibles pour faire des inférences efficaces à partir d'échantillons de ce type. Les méthodes fondées sur le plan de sondage ont l'avantage de ne pas s'appuyer sur un modèle de population hypothétique, mais dépendent, en ce qui concerne leur validité, de la mise en œuvre du plan de sondage dans des conditions contrôlées et connues, ce qui est parfois difficile, voire impossible, en pratique. Les méthodes fondées sur un modèle offrent plus de souplesse quant au plan de sondage, mais requièrent que la population soit modélisée au moyen de modèles de graphes stochastiques et que le plan de sondage soit ignorable ou de forme connue, afin qu'il puisse être inclus dans les équations de vraisemblance ou d'inférence bayésienne. Aussi bien pour les méthodes basées sur le plan de sondage que celles fondées sur un modèle, le point faible est souvent le manque de contrôle concernant l'obtention de l'échantillon initial, à partir duquel débute le dépistage des liens. Les plans de sondage décrits dans le présent article offrent une troisième méthode, dans laquelle les probabilités de sélection de l'échantillon deviennent pas-à-pas moins dépendantes de la sélection de l'échantillon initial. Un modèle de « marche aléatoire » markovienne idéalise au moyen d'un graphe, les tendances d'un plan d'échantillonnage naturel d'une séquence de sélections par dépistage de liens à suivre. Le présent article présente des plans de sondage à marche uniforme ou ciblée dans lesquels la marche aléatoire est ajustée à chaque pas afin de produire un plan de sondage ayant les probabilités stationnaires souhaitées. On obtient ainsi un échantillon qui, à d'importants égards, est représentatif au pied de la lettre de la population d'intérêt dans son ensemble, ou qui ne nécessite que de simples facteurs de pondération pour qu'il en soit ainsi.

Mots clés : Échantillonnage adaptatif; échantillonnage déterminé selon les répondants (Respondent-driven sampling); échantillonnage d'une population cachée; échantillonnage en réseau; échantillonnage par graphes; marche aléatoire; méthode de Monte Carlo par chaîne de Markov; plans d'échantillonnage par dépistage de liens.

1. Introduction

Les populations comportant des liens ou une structure en réseau sont conceptualisées sous forme de graphes dans lesquels les nœuds (ou sommets) représentent les unités de la population et les arêtes ou les arcs, les relations ou liens entre ces unités. L'un des grands problèmes des études par établissement de graphes est qu'il est difficile, voire impossible, pour de nombreuses populations d'intérêt, d'obtenir des échantillons au moyen des plans de sondage conventionnels et que les échantillons sélectionnés peuvent être, tels qu'ils sont obtenus, fortement non représentatifs de la population d'intérêt dans son ensemble. En pratique, les seules méthodes d'échantillonnage applicables consistent souvent à suivre les liens à partir des nœuds sélectionnés, afin d'y ajouter des nœuds et des liens supplémentaires. Par exemple, lors de l'étude de populations humaines cachées, telles que les utilisateurs de drogues injectables, les travailleurs du sexe et d'autres populations courant le risque de contracter ou de transmettre le VIH/Sida ou l'hépatite C, les liens sociaux sont suivis en partant des répondants identifiés au départ, afin d'accroître l'échantillon de participants à l'étude. De même, dans les études des caractéristiques d'Internet, la procédure habituelle consiste à obtenir un échantillon de sites Web en suivant les liens allant des sites initiaux vers d'autres sites.

Klovdahl (1989) a utilisé l'expression « marche aléatoire » pour décrire une procédure conçue afin d'obtenir un échantillon à partir d'une population cachée en demandant à un répondant d'identifier plusieurs contacts, dont un est sélectionné au hasard pour être le répondant suivant, et en répétant le scénario pendant un certain nombre de pas. Heckathorn (1997) a décrit des méthodes d'échantillonnage déterminé selon les répondants « respondent-driven sampling » en appliquant des procédures de ce genre. En pratique, la raison qui motive l'utilisation de plans de sondage de ce type est de pénétrer plus en profondeur dans la population cachée afin d'obtenir des répondants plus « représentatifs » de la population que ne le sont peut-être les personnes plus visibles sélectionnées initialement. Dans les études d'Internet, l'idée parallèle est que l'« internaute aléatoire », qui choisit une page Web au hasard, clique ensuite au hasard sur l'un des liens figurant sur cette page, passant ainsi à une autre page, et ainsi de suite (Brin et Page 1998).

1. Steven K. Thompson, Département de statistique et de science actuarielle, Université Simon Fraser, Burnaby (Colombie-Britannique), Canada, V5A 1S6. Courriel : Thompson@stat.sfu.ca.

Le plan de sondage à marche aléatoire peut être conceptualisé comme une chaîne de Markov (Heckathorn 1997, 2002; Henzinger et coll. 2000; Salganik et Heckathorn 2004). Dans le présent article, nous décrivons certaines modifications apportées à ces plans de sondage à chaîne de Markov, dans le but d'obtenir des probabilités stationnaires de valeur égale ou spécifiée afin d'obtenir des estimations simples des caractéristiques du graphe de la population d'intérêt.

Les approches de l'inférence à partir d'échantillons provenant d'un graphe comprennent les méthodes fondées sur le plan de sondage, les méthodes fondées sur un modèle et les méthodes mixtes fondées sur une combinaison des deux. Dans l'approche fondée sur le plan de sondage, toutes les valeurs des variables de nœud et de lien du graphe sont considérées comme étant fixes ou données, et l'inférence est basée sur les probabilités induites par le plan de sondage intervenant dans la sélection de l'échantillon. Dans l'approche fondée sur un modèle, la population proprement dite est considérée comme une réalisation d'un modèle de graphe stochastique, qui fournit la loi de probabilité conjointe de toutes les variables de nœud et de lien. Les approches fondées sur le plan de sondage décrites antérieurement comprennent les méthodes d'échantillonnage en réseau ou basé sur la multiplicité (Birnbaum et Sirken 1965), l'échantillonnage en grappes adaptatif appliqué à un graphe (Thompson et Collins 2002), ainsi que quelques-unes des méthodes décrites dans la littérature sur l'échantillonnage en boule de neige (Frank 1977, 1978; Frank et Snijders 1994). Une méthode combinant les approches fondées sur le plan de sondage et sur un modèle est utilisée dans Felix-Medina et Thompson (2004) pour étudier une population cachée dans laquelle un échantillonnage par dépistage de liens est réalisé à partir d'un échantillon d'enquête probabiliste tiré d'une base de sondage couvrant uniquement une partie de la population.

L'avantage des méthodes fondées sur le plan de sondage, dans le cas de populations humaines cachées qui ont leur propre réseau social et sont difficiles à modéliser de façon réaliste, est que certaines propriétés des inférence, comme l'absence de biais et la convergence des estimateurs, ne dépendent pas d'hypothèses de modélisation. Par contre, elles dépendent de la mise en œuvre du plan de sondage comme il a été prévu; or, l'application exacte d'un plan de sondage particulier peut constituer un très grand défi dans les études de populations humaines cachées. C'est ce qui a motivé l'élaboration d'une gamme de méthodes fondées sur un modèle pour l'inférence à partir d'échantillons de graphe, y compris les techniques du maximum de vraisemblance et les techniques bayésiennes (Thompson et Frank 2000; Chow et Thompson 2003). Fondées sur l'hypothèse que l'échantillon de départ est « ignorable » au sens de la vraisemblance (Rubin 1976) ou que le plan de sondage est de forme connue de sorte qu'il peut être inclus dans les équations de la vraisemblance et de l'inférence bayésienne, ces méthodes conviennent à une très grande gamme de procédures d'échantillonnage par dépistage de liens, y compris la plupart des variantes des méthodes d'échantillonnage en boule de neige et en réseau. Toutefois, en pratique, il se peut que l'échantillon initial soit sélectionné d'une façon loin d'être ignorable, avec probabilités de sélection dépendant de la valeur de nœud, du degré de nœud et d'autres facteurs. L'omniprésence du problème de la sélection de l'échantillon initial dans les études par dépistage de liens a été soulignée par Spreen (1992), entre autres.

L'approche poursuivie dans le présent article ne repose pas sur l'hypothèse d'un contrôle total sur toutes les possibilités de plan de sondage, mais vise plutôt à tirer parti de la façon dont les échantillons ont naturellement tendance à être sélectionnés dans les populations en réseau par les ethnographes ou d'autres spécialistes des sciences sociales, les membres de la population proprement dits ou les moteurs de recherche Web automatisés. Partant de ces processus naturels de sélection, nous introduisons des modifications itératives afin d'obtenir des procédures d'échantillonnage qui, pas à pas, s'approchent des probabilités de sélection souhaitées.

Quoique la structure sous-jacente des plans de sondage décrits dans l'article dépende de chaînes de Markov, les estimateurs et les paramètres présentant le plus d'intérêt pour les chercheurs pourraient en fait ne pas être markoviens. Par exemple, alors que la séquence de sélection d'unités d'échantillonnage peut ne dépendre, à chaque pas, que de l'unité sélectionnée le plus récemment, la séquence selon laquelle des unités distinctes sont ajoutées à l'échantillon dépend de toutes les unités sélectionnées jusqu'à ce moment-là. Par conséquent, nous étudions par simulation les propriétés de plusieurs estimateurs conjugués à divers plans de sondage, en sélectionnant répétitivement des échantillons à partir de réalisations d'un graphe stochastique et à partir d'une population empirique provenant d'une étude sur des personnes courant un grand risque de transmission du VIH/Sida.

À la section 2, nous décrivons les plans à marche aléatoire. Aux sections 3 et 4, nous présentons les plans à marche uniforme et ciblée, respectivement. À la section 5, nous donnons un exemple illustratif en prenant pour population une réalisation d'un modèle de graphe stochastique et un exemple empirique en utilisant des données provenant d'une étude portant sur une population présentant un risque élevé de transmission du VIH/Sida.

2. Marche aléatoire

La population d'intérêt est un graphe, donné par un ensemble de N nœuds portant les étiquettes $U = \{1, 2, \ldots, N\}$ et ayant les valeurs $\mathbf{y} = \{y_1, \ldots, y_N\}$, et une matrice **A** de dimensions $N \times N$ indiquant les relations ou les liens entre les nœuds. Un élément a_{ij} de **A** a la valeur 1 s'il existe un lien allant de i au nœud j et la valeur 0, autrement. Nous supposons que les éléments diagonaux a_{ii} sont nuls. Pour le nœud i, la somme de ligne $a_{\bullet i}$ est le « degré sortant » ou nombre de nœuds vers lesquels i possède un lien (successeurs) et la somme de colonne $a_{\bullet i}$ est le « degré entrant » ou nombre de nœuds qui ont un lien vers i (prédécesseurs). Dans le cas d'un graphe non orienté, la matrice **A** est symétrique et le degré entrant de tout nœud est égal à son degré sortant.

Soit W_k l'unité ou le nœud du graphique qui est sélectionné lors de la k^{e} vague. Si i est le nœud sélectionné à la k^{e} vague, alors à la vague $k+1$, l'un des nœuds reliés en partant de i est sélectionné au hasard. Donc, $\{W_0, W_1, W_2, \ldots\}$ est une chaîne de Markov avec

$$P(W_{k+1} = j \mid W_k = i) = a_{ij} / a_{i\bullet}. \qquad (1)$$

Soit **Q** la matrice de transition de la chaîne avec les éléments $q_{ij} = P(W_{k+1} = j \mid W_k = i)$. La chaîne est une marche aléatoire en ce sens qu'à chaque pas, l'un des états voisins de l'état courant est sélectionné au hasard.

Si le graphe est constitué d'une seule composante connectée, c'est-à-dire si chaque nœud du graphe peut être atteint à partir de chaque autre nœud selon un certain chemin, alors la chaîne est irréductible et ses probabilités stationnaires $(\pi_1, \ldots, \pi_N)$ satisfont $\pi_j = \sum \pi_i q_{ij}$ pour $j = 1, \ldots, N$. En fait, dans le cas du plan d'échantillonnage à marche aléatoire simple dans un graphe non orienté connecté, on peut montrer que les probabilités stationnaires (Salganik et Heckathorn 2004) sont

$$\pi_j \propto a_{\bullet j}.$$

Autrement dit, dans un graphe non orienté ne comportant qu'une seule composante connectée, la fréquence de sélection de long terme de tout nœud est proportionnelle à son degré entrant, qui est égal au degré sortant, puisque le graphe n'est pas orienté.

Supposons que l'on veuille estimer une caractéristique du graphe de population, telle que la moyenne de population des valeurs de nœud $\mu_y = \sum_{i=1}^{N} y_i / N$ en utilisant des données provenant d'un échantillon sélectionné par marche aléatoire. La moyenne d'échantillon $\bar{y} = \sum_{i \in s} y_i$ n'est généralement pas sans biais, parce que la valeur y_i d'un nœud peut être reliée au degré de celui-ci et, donc, à sa probabilité d'être sélectionné. Cependant, on peut obtenir une estimation approximativement sans biais en pondérant chaque valeur y de l'échantillon par l'inverse de son degré entrant, en supposant que cette information puisse être extraite des données (Salganik et Heckathorn 2004).

2.1 Marche aléatoire avec sauts aléatoires

Dans un graphe avec composantes distinctes ou avec nœuds non connectés, la marche aléatoire simple que nous venons de décrire n'a pas la propriété que chaque nœud peut, en dernière analyse, être atteint à partir de chaque autre nœud. Sans cette propriété, la loi limite de la marche aléatoire est sensible à la loi initiale, puisque la probabilité limite de sélection d'un nœud dépend de la probabilité initiale de démarrer dans la composante qui contient ce nœud. Une modification du plan d'échantillonnage qui permet de surmonter ce problème consiste à autoriser un saut avec faible probabilité vers un nœud choisi au hasard dans l'ensemble du graphe. À chaque pas, cette marche aléatoire suit un lien sélectionné au hasard avec la probabilité d et, avec la probabilité $1-d$, saute à un autre nœud du graphe au hasard ou avec une probabilité spécifiée. Dans la littérature traitant de la recherche au sujet d'Internet, d est appelé « facteur d'amortissement », puisqu'une valeur de d inférieure à 1 amortit l'effet du degré sortant d'un nœud donné (Brin et Page 1998).

Les probabilités de transition pour la marche aléatoire avec sauts sont

$$q_{ij} = \begin{cases} (1-d)/N + d a_{ij} / a_{i\bullet} & \text{si} \quad a_{i\bullet} > 0 \\ 1/N & \text{si} \quad a_{i\bullet} = 0. \end{cases} \qquad (2)$$

Dans le cas de la faible probabilité $1-d$ d'un saut aléatoire à n'importe quel pas, la marche aléatoire markovienne peut, en principe, atteindre tout nœud du graphe à partir de tout autre nœud, de sorte que la chaîne est irréductible. En outre, les sauts aléatoires, qui comprennent la possibilité d'aller du nœud i au nœud i, assurent que la chaîne soit apériodique de sorte que les probabilités stationnaires concordent avec les probabilités limites. Si $d < 1$, la probabilité stationnaire du nœud i n'est pas une fonction simple de son propre degré entrant et dépend aussi des probabilités stationnaires des nœuds qui s'y relient.

De façon plus générale, les sauts peuvent être faits avec n'importe quelle probabilité spécifiée $\mathbf{p} = (p_1, \ldots, p_N)$ et la probabilité d'un saut peut dépendre de l'état courant, de sorte que les probabilités de transition sont

$$q_{ij} = \begin{cases} (1-d_i) p_j + d_i a_{ij} / a_{i\bullet} & \text{si} \quad a_{i\bullet} > 0 \\ 1/N & \text{si} \quad a_{i\bullet} = 0. \end{cases}$$

Des estimations des caractéristiques du graphe de population approximativement sans biais par rapport au plan peuvent être obtenues en pondérant les valeurs d'échantillon par des facteurs inversement proportionnels aux probabilités limites de sélection de la chaîne de Markov, mais avec le problème supplémentaire que ces probabilités limites sont inconnues et doivent être estimées d'après les données

d'échantillon (voir Henzinger et coll. 2000, pour une approche de ce problème).

Dans la suite de l'article, les expressions « marche aléatoire » ou « marche aléatoire ordinaire » feront référence à la marche aléatoire avec sauts, sauf indication contraire explicite.

3. Marche uniforme

À la présente section, nous proposons une modification du plan d'échantillonnage à marche aléatoire qui mène à des probabilités stationnaires uniformes $\boldsymbol{\pi} = (\pi, \ldots, \pi)$.

Commençons par considérer le cas du graphe de population constitué d'une seule composante connectée. Soit **Q** la matrice de transition pour la marche aléatoire simple avec les probabilités de transition q_{ij} données par (1). Supposons qu'au pas k, l'état du processus est i. Une sélection provisoire est faite en utilisant les probabilités de transition de la i^{e} ligne de **Q**. Supposons que la sélection provisoire soit le nœud j. Si le degré sortant $a_{j\bullet}$ du nœud j est inférieur au degré sortant $a_{i\bullet}$ du nœud i, alors la sélection pour la vague suivante est le nœud j, c'est-à-dire $W_{k+1} = j$. Si, par contre, le degré sortant du nœud j est supérieur au degré sortant du nœud i, alors un nombre aléatoire uniforme Z est sélectionné dans l'intervalle unitaire. Si $Z < a_{i\bullet} / a_{j\bullet}$, alors $W_{k+1} = j$. Sinon, $W_{k+1} = i$.

En utilisant la méthode de Hastings-Metropolis (Hastings 1970), nous construisons la matrice de transition pour la marche modifiée dans le graphe connecté au moyen des éléments

$$P_{ij} = q_{ij}\alpha_{ij} \qquad \text{pour} \quad i \neq j$$

et

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

où

$$\alpha_{ij} = \min\left\{\frac{a_{i\bullet}}{a_{j\bullet}}, 1\right\}.$$

Dans le cas d'un graphe de population contenant des composantes distinctes ou des nœuds isolés, la marche aléatoire avec sauts, dont la matrice de transition **Q** est donnée par (2), peut être modifiée pour obtenir

$$P_{ij} = q_{ij}\alpha_{ij} \qquad \text{pour} \quad i \neq j$$

et

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

où

$$\alpha_{ij} = \min\left\{\frac{q_{ji}}{q_{ij}}, 1\right\}.$$

Donc, pour deux nœuds mutuellement connectés i et j, la probabilité d'acceptation d'une transition de i à j est

$$\alpha_{ij} = \min\left\{\frac{(1-d)/N + d/a_{j\bullet}}{(1-d)/N + d/a_{i\bullet}}, 1\right\}.$$

Pour une transition d'une unité isolée à une unité dans une composante plus grande qu'un nœud, la probabilité d'acceptation est $\alpha_{ij} = 1 - d$. Pour les autres probabilités d'acceptation, $\alpha_{ij} = 1$. Notons aussi que dans un graphe orienté, la probabilité d'acceptation serait nulle pour le parcours d'un lien asymétrique.

La marche uniforme est appliquée, si l'état courant est i, en sélectionnant un prochain état candidat, disons j, d'après les probabilités de transition figurant sur la i^{e} ligne de **Q**. Un nombre aléatoire uniforme standard Z est sélectionné et, si $Z < \alpha_{ij}$, l'état suivant est j, tandis qu'autrement, la marche reste à l'état i pendant un pas supplémentaire.

La quantité α_{ij}, dans le cas des plans à marche uniforme, dépend des probabilités de transition connues de la marche aléatoire de base, si bien que sa mise en œuvre ne nécessite pas d'estimation.

4. Marche ciblée

La même approche peut être suivie pour construire une marche ayant n'importe quelle probabilité stationnaire spécifiée, comme la sélection des nœuds dont la valeur y est élevée avec de plus grandes probabilités ou la sélection de nœuds de telle façon que les probabilités soient strictement proportionnelles à leur degré, même si le graphe contient des composantes distinctes connectées. Soit $\pi_i(y)$ la probabilité de sélection stationnaire souhaitée pour le i^{e} nœud sous forme d'une fonction de sa valeur de y. Par exemple, lors d'une étude d'une population humaine cachée exposée au risque de transmission du VIH/Sida, supposons que l'on souhaite échantillonner les utilisateurs de drogues injectables $(y_i = 1)$ avec une probabilité double de celle appliquée pour les membres de cette population qui ne prenne pas ce genre de drogues $(y_i = 0)$. Les probabilités de transition pertinentes pour la marche à valeur ciblée, en utilisant de nouveau la méthode de Hastings-Metropolis, sont

$$P_{ij} = q_{ij}\alpha_{ij} \qquad \text{pour} \quad i \neq j$$

et

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

où

$$\alpha_{ij} = \min\left\{\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1\right\}.$$

Soulignons que la probabilité de transition de base est connue, puisqu'elle dépend uniquement du degré sortant des nœuds observés, de la probabilité choisie d et du ratio spécifié π_j / π_i.

Dans le cas d'une marche pour laquelle la probabilité de sélection relative dépend de la valeur de y, le ratio $\pi_j(y_j)/\pi(y_i)$ est spécifié et

$$\alpha_{ij} = \min\left\{\frac{\pi_j(y_j) q_{ji}}{\pi_i(y_i) q_{ij}}, 1\right\}.$$

Un autre exemple de marche ciblée pourrait être celui d'une distribution cible obtenue en sélectionnant les nœuds proportionnellement à leur degré sortant, c'est-à-dire au nombre de liens qui en partent. Puisque le degré d'un nœud isolé est nul, une possibilité que nous nommerons marche ciblée selon le « degré + 1 », consiste simplement à ajouter une unité à chaque degré, de sorte que $\pi_i \propto a_{i\bullet} + 1$ soit la probabilité de sélection cible.

Un choix légèrement différent, appelé simplement marche ciblée selon le degré, consiste à n'ajouter une unité qu'au degré des nœuds isolés, de sorte que $\pi_i \propto \max(a_{i\bullet}, 1)$. Pour une marche ciblée selon le degré de ce type, la probabilité d'acceptation d'une transition entre deux nœuds connectés mutuellement est

$$\alpha_{ij} = \min\left\{\frac{a_{j\bullet}(1-d)/N + 1}{a_{i\bullet}(1-d)/N + 1}, 1\right\}.$$

Pour une transition entre un nœud isolé et un nœud dont le degré est positif, la probabilité est

$$\alpha_{ij} = \min(a_{j\bullet}(1-d), 1).$$

La probabilité de transition entre deux nœuds ayant chacun un degré positif est

$$\alpha_{ij} = \min\left\{\frac{a_{j\bullet}}{a_{i\bullet}}, 1\right\}.$$

Dans ce cas,

$$\alpha_{ij} = \min\left\{\frac{a_{j\bullet} q_{ji}}{a_{i\bullet} q_{ij}}, 1\right\}.$$

Puisque les nœuds isolés, n'ayant aucun lien avec d'autres nœuds, sont de degré nul, afin de leur donner une probabilité de sélection positive, on peut attribuer arbitrairement à leur degré la valeur « 1 » dans le calcul de la marche ciblée selon le degré ou ajouter la valeur 1 au degré de chaque nœud.

5. Plan de sondage à marche sans remise

Les résultats relatifs aux lois limites des sections précédentes s'appliquent exactement au plan de sondage à marche aléatoire avec remise, de sorte que la sélection des nœuds peut se poursuivre indéfiniment au sein de la population finie. Certains estimateurs utilisés dans les exemples qui suivront sont toutefois fondés sur la séquence d'unités distinctes sélectionnées par ce processus. Dans le cas de la séquence d'unités distinctes, qui, en fait, fournit un échantillon à marche aléatoire sans remise, on ne peut ajouter des unités que jusqu'à ce que le nombre de nœuds distincts soit le même dans l'échantillon que dans la population finie, point auquel la moyenne d'échantillon et la moyenne de population coïncident.

Une autre procédure en vue de sélectionner un échantillon par marche aléatoire sans remise consiste à restreindre directement la sélection de l'unité suivante, à n'importe quel pas, à l'ensemble d'unités qui n'ont pas encore été sélectionnées, comme dans la « marche aléatoire auto-évitante » (Lovász 1993). Si l'on utilise une procédure sélection-rejet comme dans les marches ciblées, la sélection suivante est faite d'après l'ensemble d'unités qui n'ont fait l'objet d'aucune sélection provisoire, que l'unité ait été ou non acceptée.

6. Estimateurs fondés sur les valeurs des nœuds acceptés

Sous une marche aléatoire uniforme avec remise, la moyenne d'échantillon tirage par tirage de la série de valeurs acceptées est asymptotiquement sans biais par rapport à la moyenne de population, parce que les probabilités de sélection limites sont toutes égales. La moyenne de l'échantillon tirage par tirage est la moyenne nominale englobant les valeurs répétées, de sorte que la valeur d'un nœud est pondérée par le nombre de fois que le nœud est sélectionné. Si l'on utilise un plan de sondage sans remise, ce même estimateur n'est pas précisément asymptotiquement sans biais, parce que les probabilités limites ne sont pas exactement égales. L'estimateur de variance type fondé sur une variance d'échantillon à marche aléatoire n'est pas sans biais, à cause de l'interdépendance des marches aléatoires. Les estimateurs de la variance sont estimés empiriquement dans les exemples.

Dans le cas d'une marche ciblée dans laquelle la probabilité limite π_i du nœud i est proportionnelle à c_i, un estimateur asymptotiquement convergent, fondé sur les probabilités limites, est fourni par l'estimateur par le ratio généralisé

$$\hat{\mu} = \frac{\sum_{s_a} y_i / c_i}{\sum_{s_a} 1/q_i}.$$

Notons que l'estimateur d'Horvitz-Thompson ne peut être utilisé, parce que la constante de proportionnalité des probabilités d'inclusion est inconnue, tandis que dans l'estimateur par le ratio généralisé, elle s'annule. De nouveau, les probabilités limites sur lesquelles est fondé l'estimateur sont vérifiées exactement pour le plan de sondage avec remise. Pour la variante sans remise, l'estimateur est examiné empiriquement dans les exemples.

7. Exemples

7.1 Graphe stochastique réalisé

La figure 1 illustre, pour commencer, une petite population simulée de 60 nœuds. Les nœuds dont la valeur est $y = 1$ sont de couleur foncée et ceux dont la valeur est $y = 0$ sont clairs. Nous considérons la réalisation entière comme étant notre population d'intérêt. Le modèle qui produit la réalisation est un modèle stochastique en blocs dans lequel la probabilité d'un lien entre deux nœuds quels qu'ils soient dépend des valeurs de nœud. Des liens sont plus susceptibles d'exister entre des nœuds de même type et les nœuds foncés sont plus fortement connectés que les nœuds de couleur claire. Par exemple, on pourrait souhaiter estimer la proportion de nœuds positifs (c'est-à-dire de nœuds dont $y = 1$) dans le graphe. Dans le graphe de population, 24 des 60 nœuds sont positifs, si bien que la proportion réelle est de 0,4. Le même graphe est présenté à droite, mais avec les tailles de nœud proportionnelles aux probabilités de sélection limites de la marche aléatoire. Étant donné que les nœuds positifs ont davantage tendance à former des liens, nombre d'entre eux ont une probabilité de sélection supérieures à la moyenne.

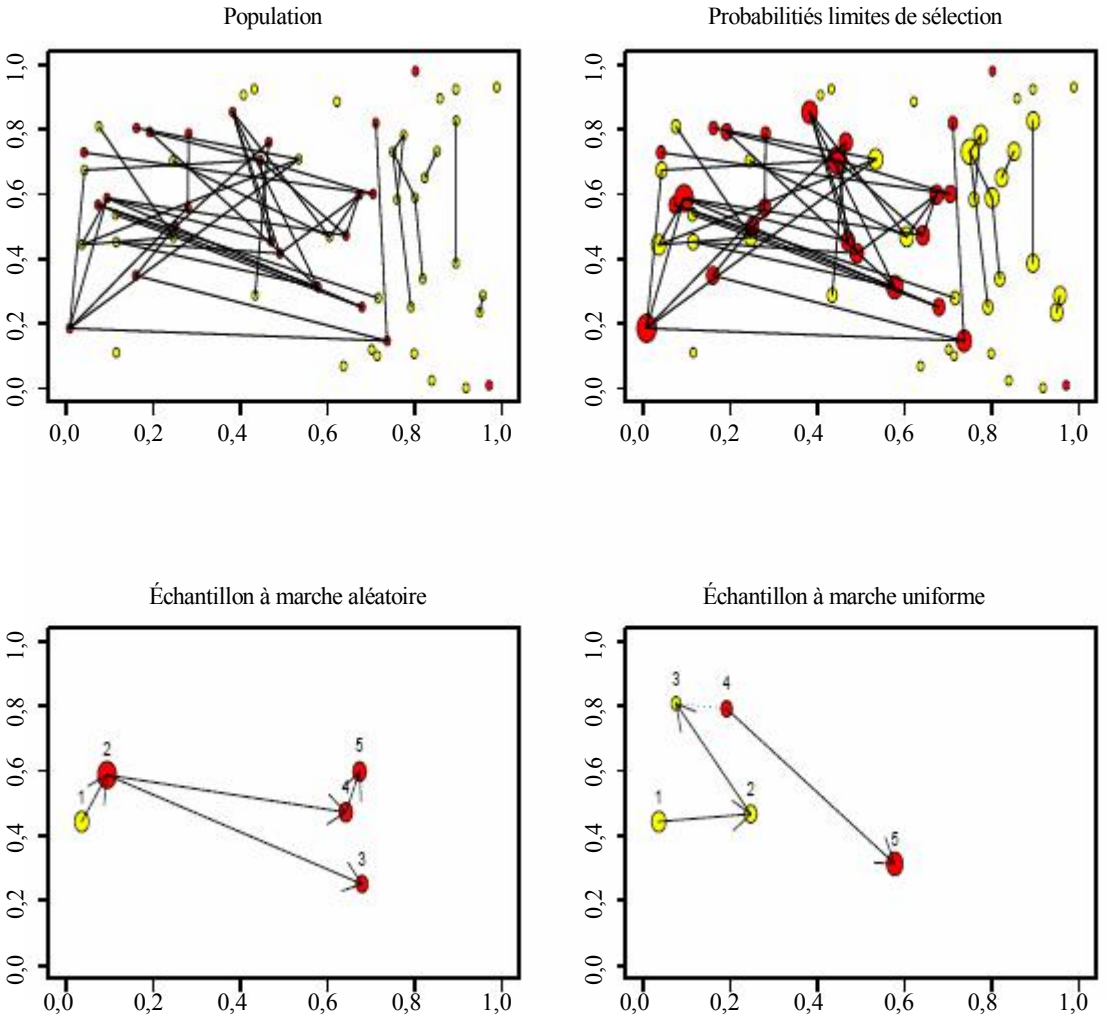


Figure 1. En haut à gauche : La population est la réalisation du modèle de graphe stochastique en blocs. En haut à droite : Probabilités limites des nœuds de la marche aléatoire. En bas à gauche : Marche aléatoire de cinq pas. En bas à droite : Marche uniforme de cinq pas. Des échelles d'axe arbitraires sont fournies comme aides visuelles pour distinguer les nœuds d'échantillon des nœuds de population.

À la rangée inférieure de la figure 1 sont présentées une marche aléatoire et une marche uniforme sélectionnées à partir de la population, comme il est illustré. Chacune a pour point de départ le même nœud sélectionné au hasard, dénoté « 1 », et se poursuit jusqu'à ce que cinq nœuds distincts soient sélectionnés. Les flèches indiquent la direction dans laquelle sont suivis les liens et un saut vers un nouveau nœud sélectionné au hasard dans le graphe est indiqué parune ligne en pointillé. Notons que la marche aléatoire revient en arrière du troisième nœud sélectionné vers le deuxième avant de suivre un nouveau lien vers le quatrième nœud échantillonné. À partir du premier nœud échantillonné, la marche uniforme passe le nœud à probabilité plus élevée sélectionné par la marche aléatoire et accepte à sa place un des nœuds qui y sont liés. Ces marches peuvent l'une et l'autre, à tout moment, faire un saut aléatoire, quoique dans les exemples illustrés, seule la marche uniforme en fasse un, lors de la transition du troisième au quatrième nœud échantillonné.

7.2 Population empirique

Des données provenant d'une étude sur la transmission hétérosexuelle du VIH/Sida dans une population à risque élevé à Colorado Springs (Potterat et coll. 1993; Rothenberg et coll. 1995) sont présentées aux figures 2 et 3. Les 595 membres de la population étudiée qui ont été interviewés sont représentés par les nœuds du graphe, et les relations sexuelles déclarées entre ces personnes sont représentées par des liens (arcs) entre les nœuds. (Les liens d'ordre sexuel supplémentaires de n'importe laquelle de ces 595 personnes avec des personnes qui n'ont pas été interviewées subséquemment ne sont pas présentés.) La population étudiée comprend les personnes à risque, c'est-à-dire les utilisateurs de drogues injectables, les travailleurs du sexe, leurs partenaires sexuels et d'usage de drogues, ainsi que d'autres personnes avec lesquelles ils ont des contacts sociaux étroits. La variable de nœud illustrée est celle de la prostitution, avec une couleur foncée pour une valeur positive ($y = 1$). Seuls les liens de nature sexuelle sont représentés, quoique bon nombre d'entre eux coïncident avec les liens relatifs à la consommation de drogues. La composante sexuellement connectée la plus importante du graphique contient 219 personnes. La composante connectée suivante, par ordre décroissant de taille, contient 12 personnes, et est suivie par plusieurs composantes de 4, 3 et 2 personnes. Les nœuds restants représentent des personnes n'ayant aucun contact sexuel déclaré au sein de la population interviewée.

Le profil observé de cette population, dont une composante connectée est beaucoup plus grande que les autres, a été décrit par divers chercheurs comme n'étant pas atypique des études portant sur des populations cachées à risque. Nous utilisons la population susmentionnée uniquement à titre de population empirique à partir de laquelle nous sélectionnons des échantillons afin de comparer des plans d'échantillonnage et des estimateurs.

La figure 3 représente la même population avec la taille de nœud tirée proportionnellement à la probabilité de sélection limite de la marche aléatoire.

Chaque tracé de la figure 4 montre la moyenne d'échantillon cumulative d'une marche unique poursuivie jusqu'à ce que 120 nœuds distincts soient sélectionnés. La proportion réelle de nœuds positifs (valeur 1) dans la population empirique (0,2235) est représentée par la droite horizontale dans chaque tracé.

Les tracés de la rangée supérieure de la figure 4 représentent une marche aléatoire ordinaire dont le nœud de départ est sélectionné aléatoirement. Le tracé de gauche montre la moyenne d'échantillon cumulative des unités distinctes. Le tracé de droite montre les mêmes données, mais avec la moyenne d'échantillon tirage par tirage, qui comprend les sélections répétées d'un même nœud, de sorte que chaque valeur de nœud soit pondérée par le nombre de fois que le nœud a été sélectionné durant la marche aléatoire.

Dans la rangée inférieure de la figure 4, nous montrons les deux mêmes types de moyenne d'échantillon pour une marche uniforme poursuivie jusqu'à ce que 120 nœuds distincts soient sélectionnés. Notons que, pour la marche aléatoire ordinaire, les fluctuations de la moyenne d'échantillon ont lieu principalement au-dessus de la moyenne réelle, ce qui représente le biais positif résultant de la sélection préférentielle des personnes plus fortement connectées, à plus haut risque, dans la population. Dans le cas de la marche uniforme, la moyenne d'échantillon fluctue plus près de la valeur réelle, sa valeur étant parfois supérieure et parfois inférieure. Chacun de ces tracés donne aussi une idée de l'autocorrélation présente dans une chaîne de Markov unique.

Les tracés de la figure 5 représentent la valeur attendue des nœuds à mesure qu'une marche progresse vague par vague, pour divers types de marches et diverses lois initiales à partir desquelles est sélectionné le premier nœud, pour la population empirique de 595 nœuds. Donc, pour la k^{e} vague, les tracés représentent $\mathrm{E}(Y_k)$, où Y_k est la valeur du nœud sélectionné à la k^{e} vague. La ligne en trait interrompu montre la moyenne réelle pour la population de Colorado Springs (0,2235). Les trois autres courbes représentent trois lois initiales différentes. Dans tous les cas, la courbe qui part du nœud le plus bas est la loi initiale uniforme, puisque la moyenne pour le nœud initial sélectionné au hasard est égale à la moyenne de la population. La loi initiale fondée sur la valeur de nœud, selon laquelle les nœuds positifs ($y = 1$) ont une probabilité initiale de sélection double des nœuds

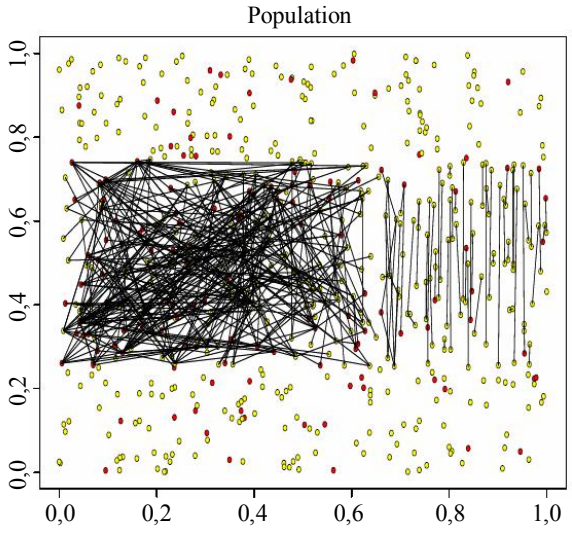


Figure 2. Population à haut risque de l'étude de Colorado Springs sur la transmission hétérosexuelle du VIH/Sida (Potterat et coll. 1993; Rothenberg et coll. 1995, et communications personnelles). Les cercles foncés représentent les individus présentant le risque le plus élevé, ici ceux qui se sont prostitués. Les liens entre les individus sont ceux de relations sexuelles et d'injection de drogues.

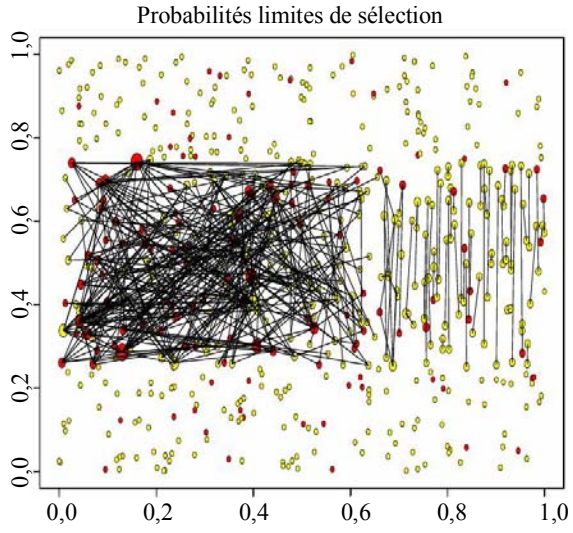


Figure 3. Probabilités limites de sélection par marche aléatoire pour la population de Colorado Springs. Soulignons que, dans la population réelle, un grand nombre d'individus présentant le comportement à risque le plus élevé ont aussi une forte probabilité d'être sélectionnés dans le cas de la marche aléatoire ordinaire et auront donc tendance à être surreprésentés dans l'échantillon.

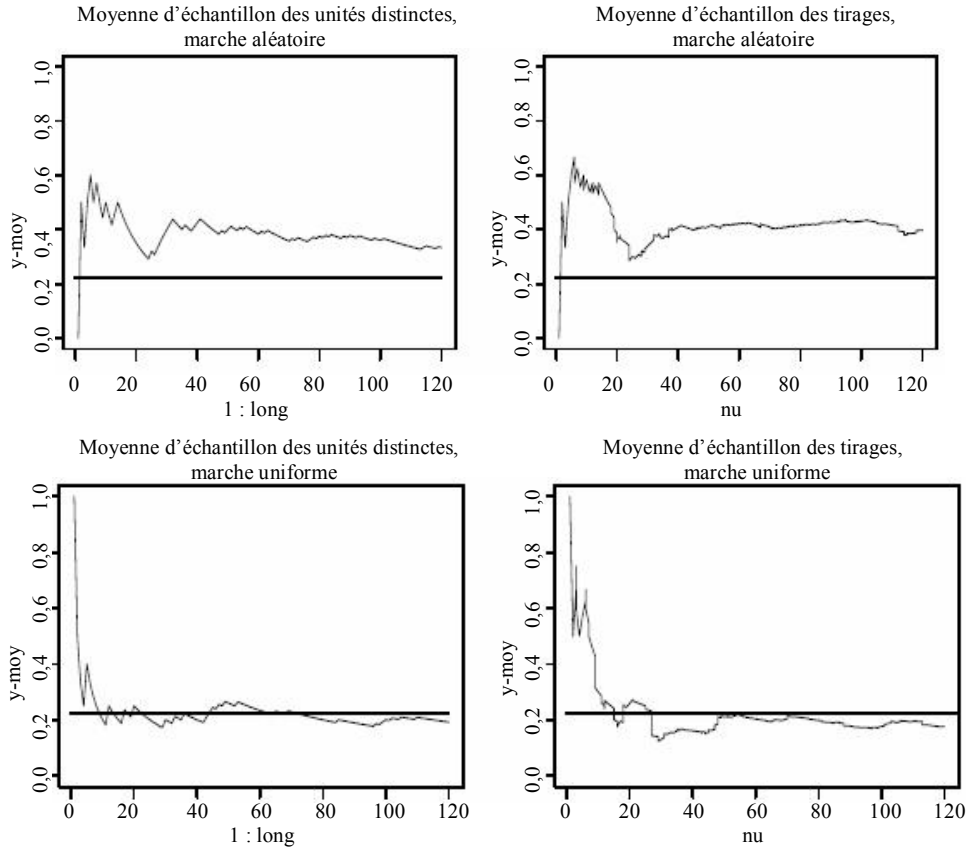


Figure 4. Chemins d'échantillon des moyennes d'échantillon pour une marche aléatoire unique de 120 nœuds de long. Les deux tracés supérieurs correspondent à une marche aléatoire ordinaire, et les deux tracés inférieurs, à une marche uniforme. La moyenne d'échantillon des unités distinctes, jusqu'à la vague donnée par l'axe des x, est représentée à gauche. La moyenne d'échantillon des tirages nominaux est représentée à droite, de sorte que la valeur de nœud soit pondérée par le nombre de fois que le nœud est sélectionné.

nuls ($y = 0$), donne la courbe de l'espérance qui est, dans tous les cas, principalement au milieu au départ et manifeste la tendance la plus forte vers une périodicité initiale. La loi initiale fondée sur le degré, selon laquelle la probabilité initiale de sélection d'un nœud est proportionnelle à son degré (plus une unité, puisque le degré des nœuds isolés est nul), forme la courbe supérieure dans chacun des tracés.

Les six tracés de la figure 5 montrent les espérances des valeurs de nœud pour six différents types de marches. Pour une marche aléatoire qui suit seulement les liens, sans possibilité de sauts aléatoires, la loi à long terme est fonction du point de départ, lequel dépend de la loi initiale. Les trois lignes séparées du premier tracé reflètent la sensibilité à la loi initiale. Par ailleurs, la marche aléatoire avec sauts permet à n'importe quel nœud d'être atteint par n'importe quel autre de sorte que la loi limite est atteinte assez rapidement peut importe la loi initiale. Avec la marche uniforme, celle qui débute avec la loi uniforme demeure dans la loi uniforme, vague après vague tandis que les marches qui débutent avec les autres lois inégales décrites, tendent assez rapidement vers la loi uniforme. Chacune de marches qui dépendent de la valeur ou du degré atteint sa loi limite assez rapidement, avec une espérance de la valeur du nœud passablement plus élevée que la valeur moyenne dans la population. La marche « degré +1 » atteint une loi dont les probabilités de sélection sont proportionnelles à un plus la valeur de chaque nœud tandis que la marche selon le degré tend vers une loi dont les probabilités limites sont proportionnelles au degré de chaque nœud sauf que les nœuds isolés se voient attribuer une valeur de degré égale à un.

Les tableaux 1 et 2 montrent les valeurs calculées de d'espérance de y pour la population de l'étude de Colorado Springs pour chaque type de marche, vague par vague, et avec diverses lois de départ pour la sélection des nœuds. Les résultats pour les marches aléatoires ordinaires sont présentés au tableau 1 et pour les marches uniformes, au tableau 2. Les espérances sont présentées pour les sélections initiales, les vagues 1, 2, 3, 4, 5, 6, 8, 16 et 32, et pour la limite à mesure que le nombre de vagues tend vers l'infini. Les trois lois initiales considérées pour la sélection du premier nœud d'une marche sont la sélection aléatoire, la sélection avec probabilité deux fois plus élevée pour les nœuds positifs que pour les nœuds à valeur nulle, et la sélection proportionnelle au degré entrant de chaque nœud plus 1. Notons que, dans le cas de k marches indépendantes

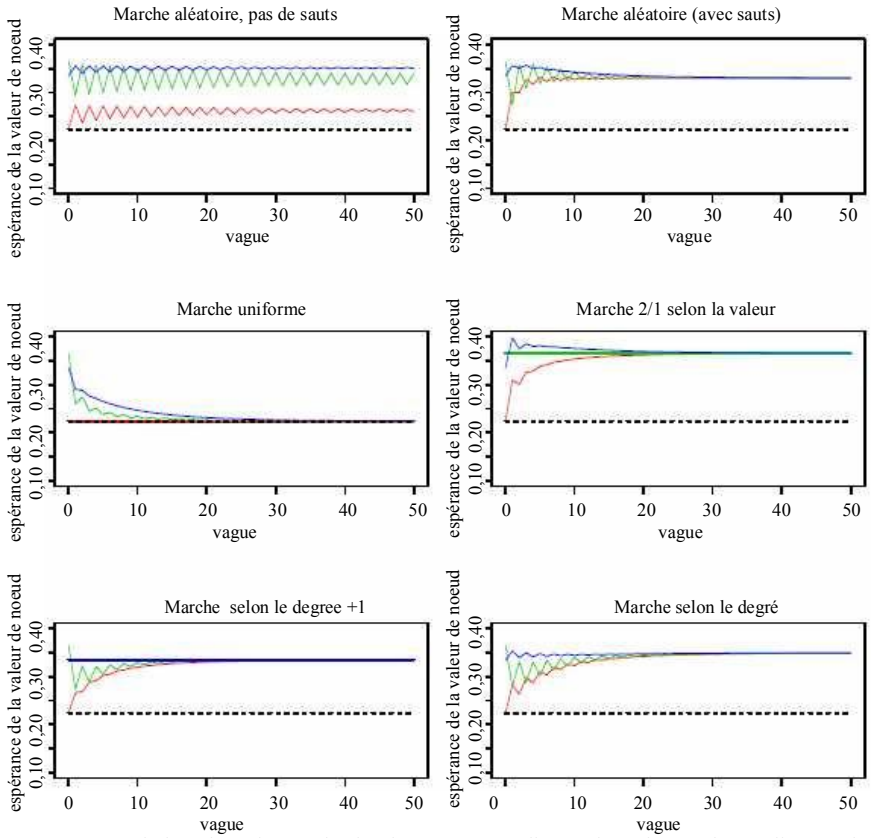


Figure 5. Espérance de la valeur de nœud selon la vague pour divers plans de marche appliqués à la population empirique de Colorado Springs. Chaque tracé illustre un plan de marche. La courbe en trait interrompu représente la moyenne réelle. Les trois autres courbes représentent l'espérance pour les trois distributions initiales examinées. Dans chaque cas, la courbe inférieure débute par la loi uniforme, celle du milieu, par la loi de probabilité 2/1 selon la valeur et la courbe supérieure, par la loi de probabilité selon le degré.

selon un plan donné, les espérances à la vague j s'appliqueraient à la moyenne d'échantillon des k valeurs de y à la vague j provenant de chacune des marches.

Tableau 1
Marches aléatoires : espérance de y pour les vagues 0, 1, 2, 3, 4, 5, 6, 8, 16, 32 et à l'infini. La vague 0 correspond à la sélection initiale. Trois hypothèses de sélection initiale différentes sont appliquées : sélection initiale aléatoire ($\pi_0 = 1/N$ pour tous les nœuds), probabilité de sélection des nœuds de valeur $y = 1$ double de celle des nœuds de valeur $y = 0 (\pi_0 \propto y+1)$, et probabilité de sélection initiale proportionnelle au degré entrant du nœud plus 1 ($\pi_0 \propto a_{\bullet j}+1$). La moyenne réelle des valeurs de nœud pour cette population est égale à 0,2235294

valeur	$\pi_0 = 1/N$	$\pi_0 \propto y+1$	$\pi_0 \propto a_{\bullet j}+1$
0	0,2235294	0,3653846	0,3349894
1	0,2998771	0,2752690	0,3560839
2	0,3005446	0,3587093	0,3507451
3	0,3273606	0,3082865	0,3570490
4	0,3177081	0,3594697	0,3500041
5	0,3320705	0,3179675	0,3528395
6	0,3231213	0,3542086	0,3469835
8	0,3256034	0,3490933	0,3440449
16	0,3291087	0,3372548	0,3363884
32	0,3302606	0,3313908	0,3315119
∞	0,3303787	0,3303787	0,3303787

Tableau 2
Marches uniformes : espérance de y pour les vagues 0, 1, 2, 3, 4, 5, 6, 8, 16, 32 et à l'infini, pour trois hypothèses de sélection initiale différentes

valeur	$\pi_0 = 1/N$	$\pi_0 \propto y+1$	$\pi_0 \propto a_{\bullet j}+1$
0	0,2235294	0,3653846	0,3349894
1	0,2235294	0,2590239	0,2903147
2	0,2235294	0,2741356	0,2877974
3	0,2235294	0,2447258	0,2761270
4	0,2235294	0,2511473	0,2707929
5	0,2235294	0,2372440	0,2646280
6	0,2235294	0,2420866	0,2600923
8	0,2235294	0,2371714	0,2522952
16	0,2235294	0,2285370	0,2352150
32	0,2235294	0,2243635	0,2256228
∞	0,2235294	0,2235294	0,2235294

Dans le cas des marches aléatoires ordinaires, qui ont pour point de départ l'échantillon initial, la valeur observée n'est sans biais par rapport à la valeur de population que pour la sélection initiale, puis le biais augmente rapidement pour atteindre sa valeur limite de 0,3303787–0,223594. Comme les échantillons initiaux présentent un biais en faveur des nœuds positifs, le biais change moins à mesure que la marche progresse.

Dans le cas de la marche uniforme, la sélection aléatoire initiale coïncide avec la distribution stationnaire, de sorte que la marche demeure sans biais vague après vague. Dans le cas de la sélection initiale des nœuds positifs avec probabilité double de celle des nœuds de valeur nulle, le biais est réduit considérablement après chacune des quelques premières vagues et les valeurs des nœuds sélectionnés s'approchent de leur état limite sans biais. Dans le cas de la sélection initiale proportionnelle au degré entrant plus 1, quelques vagues de plus sont nécessaires pour que le biais devienne petit. Le rapprochement initial rapide de l'espérance vers la valeur limite donne à penser qu'il pourrait être souhaitable de considérer une période initiale « de rodage » qui ne sera pas utilisée dans l'estimation. Même un rodage très court d'une à trois vagues pourrait réduire sensiblement le biais des estimateurs fondés sur de courtes marches.

Les figures 6 à 9 illustrent les distributions d'échantillonnage des moyennes d'échantillon et des estimateurs pondérés pour divers plans à marche aléatoire pour l'ensemble de données de Colorado Springs. Chaque histogramme est basé sur 1 000 simulations du plan d'échantillonnage appliqué à la population empirique. Pour les plans illustrés aux figures 6 et 7, chaque échantillon comprend 24 marches, chacune de 5 pas, c'est-à-dire continuant jusqu'à ce que 5 nœuds distincts soient sélectionnés. La figure 5 représente les distributions des moyennes d'échantillon pour les marches aléatoires (rangée supérieure) et pour les marches uniformes (rangée inférieure). La distribution de la moyenne des 24 moyennes d'échantillon des 5 unités distinctes est donnée à gauche. À droite est donnée la moyenne des 24 moyennes tirage par tirage, qui intègre les sélections répétées.

La proportion réelle (0,2235) de nœuds ayant la valeur y dans la population empirique est indiquée par le triangle plein, tandis que la moyenne de la distribution d'échantillonnage est indiquée par le triangle vide. Dans le cas des marches aléatoires, les moyennes d'échantillon présentent un biais par excès, tandis que dans le cas de la marche uniforme, elles sont presque sans biais. La moyenne n'est précisément sans biais ni dans l'un ni dans l'autre cas parce que la marche se poursuit jusqu'à ce qu'un nombre fixe de nœuds distincts soit sélectionné au lieu de se poursuivre pendant un nombre fixe de vagues.

La figure 7 illustre la distribution de l'estimateur par le ratio généralisé pour les marches ciblées dont les probabilités stationnaires sont reliées à la valeur des nœuds et à leur degré (degré du nœud plus 1). Aux fins de comparaison, chacune de ces marches a débuté dans sa propre loi stationnaire, ce qui donne en fait les distributions des estimateurs après le « rodage ». Ces estimateurs ne sont pas dépourvus de biais, parce que la taille effective de l'échantillon est fixe, ce qui affecte les probabilités réelles avec lesquelles les nœuds distincts sont sélectionnés en série et que le dénominateur de l'estimateur est aléatoire, puisqu'il est égal à la somme des poids de sondage.

Les figures 8 et 9 donnent les distributions des mêmes estimateurs et plans de sondage qu'aux figures 6 et 7, mais dans le cas où chaque échantillon consiste en une longue marche de 120 nœuds distincts.

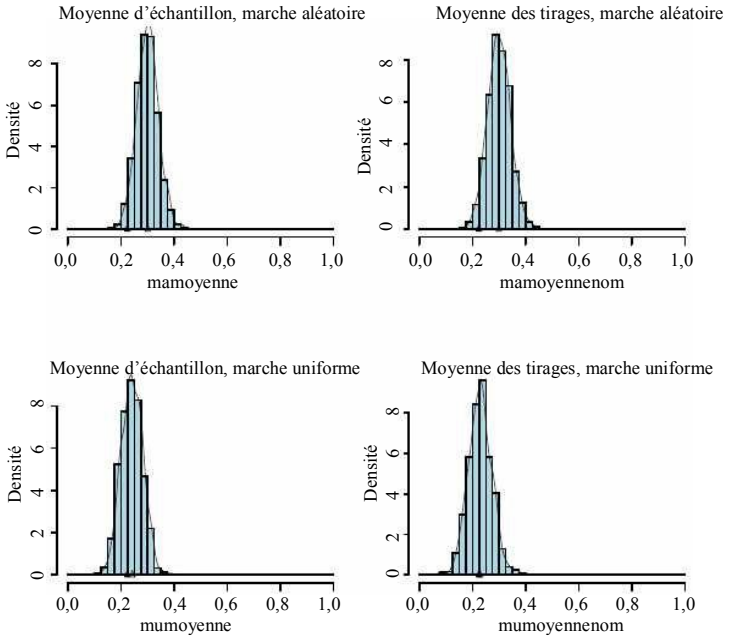


Figure 6. Distributions des moyennes d'échantillon en tant qu'estimateurs de la proportion de personnes qui se sont prostituées dans la population empirique de l'étude de Colorado Springs, sous les marches aléatoires et uniformes. Le triangle plein représente la proportion réelle dans la population. Le triangle vide représente la moyenne de la distribution de l'estimateur. À noter la surestimation dans le cas des moyennes d'échantillon pour les marches aléatoires ordinaires. Les marches aléatoires sont représentées à la partie supérieure de la figure et les marches uniformes, à la partie inférieure. Le plan comporte 24 marches, chacune de 5 pas, et l'ensemble des 120 observations sont utilisées dans l'estimateur. Le nombre de réalisations de la simulation est égal à 1 000.

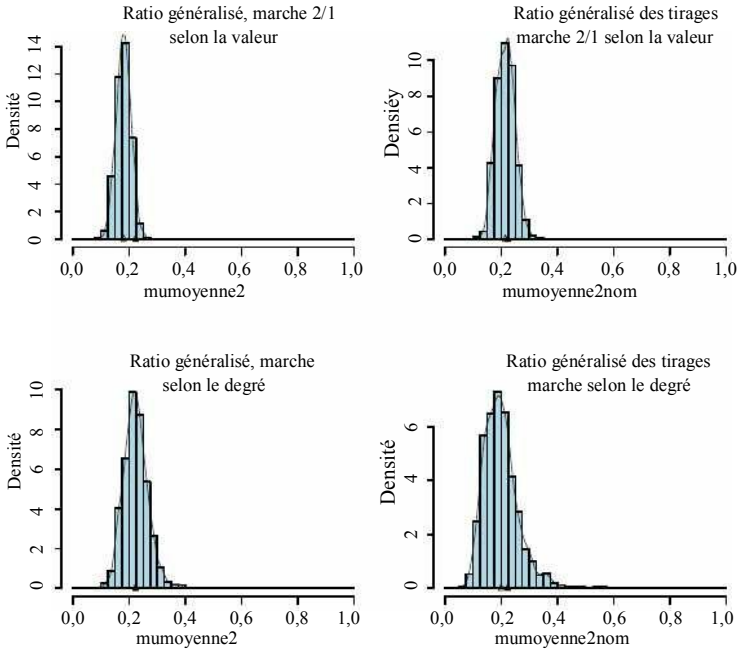


Figure 7. Distributions des estimateurs par le ratio généralisé de la proportion de personnes qui se sont prostituées dans la population empirique de l'étude de Colorado Springs, sous marches ciblées. Le triangle plein représente la proportion réelle dans la population. Le triangle vide représente la moyenne de la distribution de l'estimateur. À noter la surestimation dans le cas des moyennes d'échantillon pour les marches aléatoires ordinaires. Les marches aléatoires sont représentées à la partie supérieure de la figure et les marches uniformes, à la partie inférieure. Le plan comprend 24 marches, de 5 pas chacune, et l'ensemble des 120 observations sont utilisées dans l'estimateur. Le nombre de réalisations de la simulation est égal à 1 000.

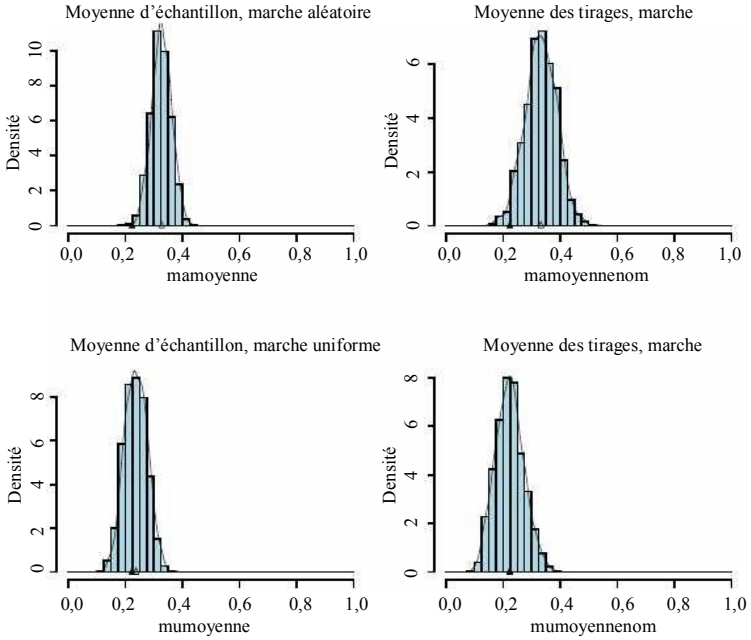


Figure 8. Distributions des moyennes d'échantillon en tant qu'estimateurs de la proportion de personnes qui se sont prostituées dans la population empirique de l'étude de Colorado Springs, sous marches aléatoires et uniformes. Le triangle plein représente la proportion réelle dans la population. Le triangle vide représente la moyenne de la distribution de l'estimateur. À noter la surestimation dans le cas des moyennes d'échantillon pour les marches aléatoires ordinaires. Les marches aléatoires sont représentées à la partie supérieure de la figure et les marches uniformes, à la partie inférieure. Le plan comprend une marche unique de 120 pas. Le nombre de réalisations de la simulation est égal à 1 000.

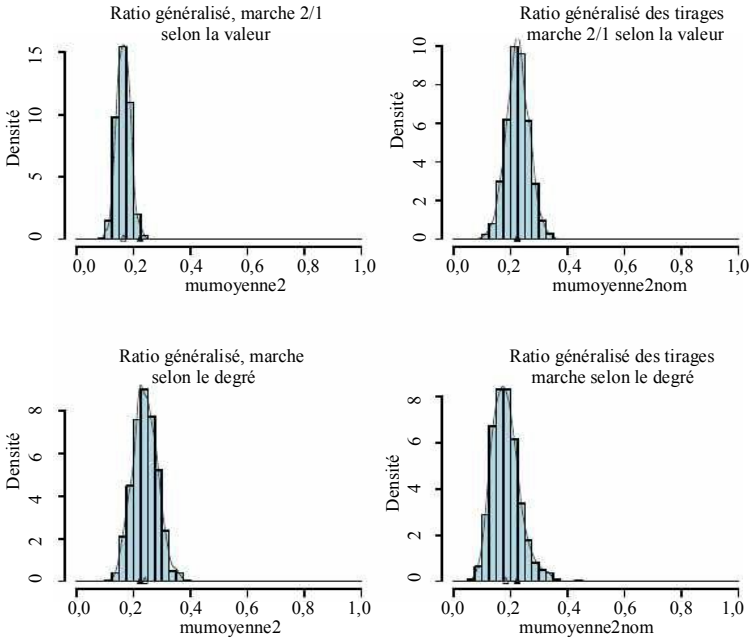


Figure 9. Distributions des estimateurs par le ratio généralisé en tant qu'estimateurs de la proportion de personnes qui se sont prostituées dans la population empirique de l'étude de Colorado Springs, sous marches ciblées. Le triangle plein représente la proportion réelle dans la population. Le triangle vide représente la moyenne de la distribution de l'estimateur. À noter la surestimation dans le cas des moyennes d'échantillon pour les marches aléatoires ordinaires. Les marches aléatoires sont représentées à la partie supérieure de la figure et les marches uniformes, à la partie inférieure. Le plan comprend une marche unique de 120 pas. Le nombre de réalisations de la simulation est égal à 1 000.

Les tableaux 3 à 6 résument les espérances et les erreurs quadratiques moyennes des estimateurs calculées pour les diverses stratégies d'après les 1 000 simulations exécutées en prenant l'ensemble de données de Colorado Springs comme population.

Tableau 3
Moyennes et erreurs quadratiques moyennes des moyennes d'échantillon des unités distinctes et moyennes tirage par tirage pour les marches aléatoires et les marches uniformes. Le plan de sondage comporte 24 marches se poursuivant chacune jusqu'à ce que 5 nœuds distincts soient inclus

Plan :	Marche aléatoire	Marche aléatoire	Marche uniforme	Marche uniforme
Estimateur :	Moyenne d'échantillon	Moyenne du tirage	Moyenne d'échantillon	Moyenne du tirage
moyenne	0,3008000	0,2994872	0,2423000	0,2289125
e.q.m.	0,007617465	0,007608868	0,002016378	0,001974826

Tableau 4
Moyennes et erreurs quadratiques moyennes pour les moyennes pondérées (estimateur par le ratio généralisé), en utilisant les unités distinctes dans chaque marche ou les sélections tirage par tirage pour les marches en fonction de la valeur et les marches en fonction du degré. Le plan comprend 24 marches se poursuivant chacune jusqu'à ce que 5 nœuds distincts soient inclus

Plan :	Marche selon la valeur	Marche selon la valeur	Marche selon le degré	Marche selon le degré
Estimateur :	Unités distinctes	Tirage par tirage	Unités distinctes	Tirage par tirage
moyenne	0,1805114	0,2144555	0,2235257	0,1994530
e.q.m.	0,002546968	0,001195507	0,001807981	0,004382568

Tableau 5
Moyennes et erreurs quadratiques moyennes pour les moyennes d'échantillon des unités distinctes et les moyennes tirage par tirage pour les marches aléatoires et les marches uniformes. Le plan comprend une marche se poursuivant jusqu'à ce que 120 nœuds distincts soient inclus

Plan :	Marche aléatoire	Marche aléatoire	Marche uniforme	Marche uniforme
Estimateur	Moyenne d'échantillon	Moyenne du tirage	Moyenne d'échantillon	Moyenne du tirage
moyenne	0,3274083	0,3325171	0,2379333	0,2232534
e.q.m.	0,012004961	0,014902382	0,001777285	0,002442825

Tableau 6
Moyennes et erreurs quadratiques moyennes pour les moyennes pondérées (estimateur par le ratio généralisé), en utilisant les unités distinctes dans chaque marche ou les sélections tirage par tirage pour les marches selon la valeur et selon le degré. Le plan comprend une marche se poursuivant jusqu'à ce que 120 nœuds distincts soient inclus

Plan :	Marche selon la valeur	Marche selon la valeur	Marche selon le degré	Marche selon le degré
Estimateur :	Unités distinctes	Tirage par tirage	Unités distinctes	Tirage par tirage
moyenne	0,1652275	0,2254267	0,2404622	0,835336
e.q.m.	0,003952703	0,001578039	0,002115518	0,03951540

Les tableau 7 et 8 donnent la variance et l'espérance des variances d'échantillon inter-marches, lorsqu'elles existent, et des variances d'échantillon intramarche pour les plans à marche uniforme.

Tableau 7
Variance des estimateurs et espérance des variances d'échantillon inter-marches et intramarche pour la marche aléatoire uniforme, pour le plan comportant 24 marches de 5 nœuds distincts chacune

Estimateur :	Moyenne d'échantillon	Moyenne tirage par tirage
Variance de l'estimateur :	0,001665709	0,001947796
E (variance inter-marches)	0,001584203	0,001919005
E (variance intramarche moyenne)	0,001515521	0,001231983

Tableau 8
Variance des estimateurs et espérance de la variance d'échantillon intramarche pour la marche aléatoire uniforme, pour le plan comportant une seule marche de 120 nœuds distincts (aucune variance d'échantillon inter-marches n'est disponible pour ce plan)

Estimateur	Moyenne d'échantillon	Moyenne tirage par tirage
Variance de l'estimateur	0,001571384	0,002445194
E (variance intramarche moyenne)	0,001510515	0,001429126

Tableau 9
Taux d'acceptation pour les marches uniforme et ciblée dans la population empirique

Plan :	Marche uniforme	Marche selon la valeur	Marche selon le degré+1	Marche selon le degré
Taux d'acceptation	0,62	0,60	0,85	0,88

8. Taux d'acceptation

Les principaux avantages des plans d'échantillonnage à chaîne de Markov contrôlée, comme les marches uniformes et ciblées, sont les suivants : 1) ils permettent de connaître les probabilités limites de sélection d'après les données, de sorte que celles-ci peuvent être utilisées dans l'estimation, 2) les probabilités limites sont choisies de sorte que certains types de nœuds ou de caractéristiques des graphes puissent être sélectionnés de manière préférentielle, 3) les estimations sont fondées sur le plan d'échantillonnage, de sorte que certaines de leurs propriétés essentielles ne dépendent pas des hypothèses, qui pourraient s'avérer incorrectes, au sujet du graphe de population proprement dit et 4) à mesure que la longueur de la chaîne augmente, l'espérance des estimations a tendance à évoluer vers la quantité correspondante des graphes, même lorsque la loi de sélection initiale diffère de la loi limite. En outre, les plans à marche uniforme produisent un échantillon qui, sans pondération ni analyse, est au pied de la lettre « représentatif » à certains égards de l'ensemble de la population.

Dans le cas des marches uniformes et ciblées, l'une des questions pratiques importantes est celle du taux d'acceptation, c'est-à-dire la probabilité moyenne qu'un nœud sélectionné provisoirement soit accepté. Les nœuds sélectionnés provisoirement qui sont rejetés ne contribuent pas aux estimateurs simples. Dans le cas d'une population telle qu'Internet, pour laquelle les sélections provisoires et les décisions d'acceptation/rejet peuvent être automatisées et exécutées rapidement, le taux d'acceptation n'est pas nécessairement critique. L'échantillonnage se poursuit simplement jusqu'à ce qu'un nombre approprié de nœuds soit accepté. Par contre, lors des études de populations humaines cachées, les tailles d'échantillon sont généralement faibles. Les membres de la population sont difficiles à atteindre et les interviews peuvent prendre beaucoup de temps. Toutefois, dans certaines études, la décision d'accepter ou de rejeter une unité d'après le degré sortant d'une personne sélectionnée provisoirement peut être prise assez rapidement au moyen d'une brève interview de filtrage. Il est malgré tout souhaitable de disposer d'une méthode d'échantillonnage dont le taux d'acceptation est aussi élevé que possible.

Les marches aléatoires sont caractérisées par une probabilité d'acceptation égale à un, mais n'ont généralement pas de probabilités limites connues ou contrôlées. Si l'on se représente la marche aléatoire sous-jacente comme le cheminement naturel, non contrôlé, au sein d'une population, alors on pourrait s'attendre à ce qu'une marche contrôlée ayant une loi limite proche de la marche aléatoire naturelle de la population produise un taux d'acceptation plus élevé qu'une marche contrôlée dont la loi limite diffère considérablement de cette marche aléatoire naturelle. Autrement dit, une marche contrôlée dont la loi stationnaire s'écarte peu de la loi de la marche aléatoire sous-jacente devrait nécessiter moins de modifications par rejet de nœuds sélectionnés provisoirement qu'une marche dont la loi stationnaire s'écarte considérablement de la marche aléatoire représentant la tendance naturelle.

Comme il est mentionné plus haut, les probabilités stationnaires d'une marche aléatoire ordinaire dans un graphe non orienté à une seule composante sont proportionnelles au degré des nœuds. Lorsqu'il existe plus d'une composante connectée, l'ajout du saut aléatoire est nécessaire pour assurer que chaque nœud puisse être atteint, pour produire une loi stationnaire unique ne dépendant pas de la loi initiale et pour faire en sorte que les probabilités limites soient influencées par le degré des nœuds, mais qu'elles n'y soient pas strictement proportionnelles. Même avec l'introduction du saut aléatoire et des probabilités d'acceptation induites, les marches ciblées produisant des probabilités stationnaires proportionnelles au degré de nœud pourraient s'approcher davantage de la loi naturelle de la marche aléatoire que les autres marches contrôlées considérées dans le présent article. En effet, si l'on examine la figure 5, il est évident que, pour la population empirique, la loi d'équilibre de la valeur de nœud espérée obtenue pour la marche selon le degré + 1 est plus proche de la loi d'équilibre de la marche aléatoire avec saut que de celle de tout autre plan contrôlé étudié.

Dans le cas de la population empirique tirée de l'étude sur la transmission hétérosexuelle du VIH/Sida, les taux d'acceptation obtenus pour les divers plans d'échantillonnage sont donnés au tableau 9. Pour le plan à marche uniforme, le taux d'acceptation est de 62 %. Pour la marche selon la valeur de nœud, où la probabilité limite est deux fois plus élevée pour les personnes à haut risque que pour celles à faible risque, le taux d'acceptation est de 60 %. Pour la marche selon le degré de nœud, dans laquelle la probabilité limite est proportionnelle au degré+1, il est de 85 %. Enfin, pour la marche selon le degré, avec une unité ajoutée uniquement au degré des nœuds isolés, il est de 88 %.

9. Discussion

Les plans d'échantillonnage à marche uniforme et à marche ciblée ont pour but de permettre de déterminer les probabilités limites de sélection d'après les données, afin de pouvoir les utiliser dans l'estimation. En outre, les probabilités limites sont choisies de sorte que certains types de nœuds ou de caractéristiques de graphe puissent être sélectionnés de manière préférentielle. La dépendance à l'égard de la sélection initiale, qui peut ne pas être contrôlée, diminue pas à pas.

Les estimateurs utilisés dans le présent article avec les plans d'échantillonnage à marche uniforme et à marche ciblée peuvent être considérés comme des estimateurs fondés sur le plan de sondage. Le plan exact fondé sur les probabilités de sélection pourrait ne pas être connu, si les probabilités de sélection initiales sont inconnues, mais on utilise les probabilités de sélection stationnaires dans les estimateurs. À mesure qu'augmente la longueur de la chaîne, ces probabilités deviennent plus exactes et l'espérance des estimations se rapproche de la quantité de graphe correspondante. L'avantage des méthodes d'estimation fondées sur le plan de sondage est que certaines de leurs propriétés, comme l'absence de biais ou la convergence par rapport au plan, ne dépendent pas d'hypothèses fondées sur un modèle qui pourraient être incorrectes. Les estimations fondées sur le plan de sondage ont la qualité intéressante supplémentaire d'être très simples et faciles à comprendre et à expliquer, et elles peuvent même produire des données qu'il est possible de présenter sans analyse ou interprétation comme étant représentatives de caractéristiques importantes de la population d'intérêt dans son ensemble.

L'utilisation d'algorithmes de Monte Carlo par chaîne de Markov pour l'analyse des données associées à des modèles compliqués est fréquente en statistique. Les approches décrites ici sont inhabituelles en ce sens que les méthodes par chaîne de Markov sont appliquées à des populations réelles pour obtenir effectivement des données, qui peuvent être facilement analysées manuellement. En fait, on pourrait aller une étape plus loin et construire un modèle de graphe stochastique bayésien complexe de la population en utilisant des méthodes de Monte Carlo par chaîne de Markov de la façon classique pour l'analyse des données, ainsi que pour leur collecte.

Les plans d'échantillonnage à marche uniforme ou ciblée sont utiles pour obtenir des échantillons de nœuds acceptés présentant certaines propriétés désirables en ce qui concerne la population, qui fournissent des estimateurs très simples des quantités de population ou qui pourraient fournir un échantillon initial pour un autre plan d'échantillonnage. Il convient de souligner que les nœuds qui ont été observés, puis « rejetés » sous les conditions du plan continuent de faire effectivement partie des données. Leur valeur peut encore être intégrée dans les estimations, au besoin, en appliquant la méthode de Rao-Blackwell, une fois que la chaîne a atteint approximativement l'équilibre, mais dans ces conditions, le calcul des estimations est complexe.

Une autre option consiste à utiliser des méthodes fondées sur un modèle, comme les méthodes d'estimation bayésiennes. En plus de la modélisation appropriée de la population par graphe stochastique, ces méthodes requièrent une procédure de sélection initiale ignorable, condition qui n'est généralement pas satisfaite sous sélections initiales biaisées par les valeurs ou les degrés de nœud, ou bien la modélisation adéquate de la procédure de sélection non ignorable dans les équations de vraisemblance. Les plans d'échantillonnage à marche ciblée produisant une loi asymptotique non corrélée à la procédure de sélection non ignorable et, donc, approximativement non corrélée aux valeurs ou aux degrés de nœud en dehors de l'échantillon pourraient fournir les sélections initiales pour un échantillon auquel les méthodes d'inférence basées sur un modèle pourraient ensuite être appliquées.

Remerciements

La présente étude a été financée par le National Center for Health Statistics, la National Science Foundation (DMS-9626102 et DMS-0406229) et les National Institutes of Health (R01-DA09872). Je tiens à remercier John Potterat et Steve Muth de m'avoir prodigué des conseils et permis d'utiliser les données provenant de l'étude de Colorado Springs.

Bibliographie

Birnbaum, Z.W., et Sirken, M.G. (1965). Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates. *Vital and Health Statistics*, Série 2, No.11. Washington: Government Printing Office.

Brin, S., et Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Proceedings of the 7th International World Wide Web Conference*, Elsevier, 107-117.

Chow, M., et Thompson, S.K. (2003). Estimation avec plans d'échantillonnage par dépistage de liens – Une approche bayésienne. *Techniques d'enquête*, 29, 221-230.

Felix-Medina, M.H., et Thompson, S.K. (2004). Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations. *Journal of Official Statistics*, 20, 19-38.

Frank, O. (1977). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1, 235-264.

Frank, O. (1978). Sampling and estimation in large social networks. *Social Networks*, 1, 91-101.

Frank, O., et Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.

Hastings, W.K. (1970). Monte-Carlo sampling methods using Markov chains and their application. *Biometrika*, 57, 97-109.

Heckathorn, D.D. (1997). Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44, 174-199.

Heckathorn, D.D. (2002). Respondent driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49, 11-34.

Henzinger, M.R., Heydon, A., Mitzenmacher, M. et Najork, M. (2000). On near-uniform URL sampling. *Proceedings of the Ninth International World Wide Web Conference*, Elsevier, 295-308.

Klovdahl, A.S. (1989). Urban social networks: Some methodological problems and possibilities. Dans *The Small World*, (Éd. M. Kochen) Norwood, NJ: Ablex Publishing, 176-210.

Lovász, L. (1993). Random walks on graphs: A survey. Dans *Combinatorics*, *Paul Erdös is Eighty*, (Éds. D. Miklós, D. Sós et T. Szöni), János Bolyai Mathematical Society, Keszthely, Hungary, 2, 1-46.

Potterat, J.J., Woodhouse, D.E., Rothenberg, R.B., Muth, S.Q., Darrow, W.W., Muth, J.B. et Reynolds, J.U. (1993). AIDS in Colorado Springs: Is there an epidemic? *AIDS*, 7, 1517-1521.

Rothenberg, R.B., Woodhouse, D.E., Potterat, J.J., Muth, S.Q., Darrow, W.W. et Klovdahl, A.S. (1995). Social networks in disease transmission: The Colorado Springs study. Dans *Social Networks*, (Éds. R.H. Needle, S.G. Genser et R.T. Trotter) Drug Abuse, and HIV Transmission, NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse, 3-19.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Salganik, M.J., et Heckathorn, D.D. (2004). Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*, 34, 193-239.

Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs: what and why? *Bulletin de Méthodologie Sociologique*, 36, 34-58.

Thompson, S.K., et Collins, L.M. (2002). Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence*, 68, S57-S67.

Thompson, S.K., et Frank, O. (2000). Estimation fondée sur un modèle et comportant des plans d'échantillonnage à dépistage de liens. *Techniques d'enquête*, 26, 99-112.

Techniques d'enquête, juin 2006
Vol. 32, N° 1, pp. 27-39
Statistique Canada, N° 12-001-XIF au catalogue

Utilisation de méthodes de traitement des données manquantes pour corriger l'erreur de mesure dans une fonction de distribution

Gabriele B. Durrant et Chris Skinner [1]

Résumé

Nous examinons le recours à l'imputation et à la pondération pour corriger l'erreur de mesure dans l'estimation d'une fonction de distribution. Le problème qui a motivé l'étude est celui de l'estimation de la distribution de la rémunération horaire au Royaume-Uni au moyen de données provenant de l'Enquête sur la population active. Les erreurs de mesure causent un biais et le but est d'utiliser des données auxiliaires, mesurées avec précision pour un sous-échantillon, en vue de le corriger. Nous envisageons divers estimateurs ponctuels, fondés sur différentes approches d'imputation et de pondération, dont l'imputation fractionnaire, l'imputation par la méthode du plus proche voisin, l'appariement d'après la moyenne prévisionnelle et la pondération par le score de propension à répondre. Nous comparons ensuite ces estimateurs ponctuels d'un point de vue théorique et par simulation. Nous recommandons d'adopter une approche d'imputation fractionnaire par appariement d'après la moyenne prévisionnelle. Elle donne les mêmes résultats que la pondération par le score de propension, mais a l'avantage d'être légèrement plus robuste et efficace.

Mots clés : Imputation par donneur; imputation fractionnaire; imputation hot deck; imputation multiple; imputation par la méthode du plus proche voisin; appariement d'après la moyenne prévisionnelle; pondération par le score de propension.

1. Introduction

L'erreur de mesure peut donner lieu à une estimation biaisée des fonctions de distribution (Fuller 1995). Dans le présent article, nous examinons diverses approches en vue de corriger ce biais quand, en plus des observations sur l'échantillon de la variable mesurée incorrectement, on dispose de valeurs de la variable mesurée correctement pour un sous-échantillon. Si ce dernier est sélectionné aléatoirement, la situation se résume à un cas du problème bien étudié de l'échantillonnage double (par exemple Tenenbein 1970). Dans ces conditions, nous pouvons produire des estimations sans biais à partir du sous-échantillon uniquement, mais utiliser les données concernant la variable substitut corrélée pour l'ensemble de l'échantillon afin d'accroître l'efficacité. Consulter, par exemple, Luo, Stokes et Sager (1998). Dans le présent article, nous supposerons que le sous-échantillon n'est pas sélectionné selon un plan randomisé connu, mais plutôt selon un mécanisme de production de données manquantes inconnu. Nous émettrons simplement l'hypothèse que les données exactes sur la variable manquent au hasard (MAR pour *missing at random*) (Little et Rubin 2002), sachant que les variables sont mesurées sur l'échantillon complet. Nous disposons de certaines méthodes d'inférence pour résoudre ce problème si nous sommes prêts à formuler des hypothèses fortes concernant les paramètres de la distribution réelle (par exemple Buonaccorsi 1990) ou du modèle d'erreur de mesure (par exemple Luo et coll. 1998). Toutefois, nous ne pousserons pas plus loin leur examen, car nous supposons que nous sommes en présence d'une application pour laquelle des hypothèses de ce genre sont irréalistes. La nouveauté du présent article tient plutôt au fait que nous considérons l'inférence sous ces conditions d'erreur de mesure comme étant un problème de données manquantes et que nous étudions l'application de méthodes d'imputation et de pondération décrites dans la littérature sur le traitement des données manquantes. Nous nous concentrerons sur le choix des meilleures méthodes en vue d'améliorer l'estimation ponctuelle de la fonction de distribution, en ce qui a trait au biais, à l'efficacité et à la robustesse aux hypothèses de modélisation. Nous ne nous pencherons que brièvement sur l'estimation de la variance.

La présente étude a été motivée par un effort en vue d'estimer la distribution de la rémunération horaire au Royaume-Uni (R.-U.) à l'aide de données provenant de l'Enquête sur la population active (EPA). L'EPA offre deux moyens de mesurer la rémunération horaire. La méthode classique consiste à recueillir des renseignements sur les gains et le nombre d'heures travaillées, puis à calculer la rémunération horaire d'après cette information. Nous appelons la variable dérivée de cette façon *variable dérivée de rémunération horaire*. Une méthode plus récente de détermination de la rémunération horaire consiste à demander directement aux répondants de déclarer quelle est cette dernière. Nous appelons la mesure résultante de la rémunération horaire la *variable directe*. Skinner et coll. (2002) décrivent, avec preuves empiriques à l'appui, de nombreuses sources d'erreur de mesure dans la variable dérivée et concluent, d'après leur étude, que la variable directe

1. Gabriele B. Durrant et Chris Skinner, Université de Southampton, Royaume-Uni. Courriel : cjs@soton.ac.uk.

mesure la rémunération horaire de façon nettement plus précise que la variable dérivée. Néanmoins, le problème de la variable directe est que les données manquent pour environ 43 % des cas. L'application est décrite dans les grandes lignes à la section 8 et exposée plus en détail dans Skinner et coll. (2002), qui proposent eux aussi de recourir à l'imputation pour régler le problème de l'erreur de mesure. Le présent article prolonge ces travaux en envisageant une plus grande gamme d'approches de traitement des données manquantes et en comparant leurs propriétés du point de vue théorique, ainsi que par simulation. L'approche d'imputation élaborée dans le présent article, qui étend celle considérée par Skinner et coll. (2002), est maintenant appliquée par l'Office for National Statistics du Royaume-Uni comme nouvelle méthode de production d'estimations de la faible rémunération.

La présentation de l'article est la suivante. À la section 2, nous discutons du problème de l'estimation. Aux sections 3 et 4, nous exposons les approches par imputation et par pondération, respectivement. À la section 5, nous étudions et comparons leurs propriétés du point de vue théorique, tandis qu'à la section 7, nous le faisons par étude en simulation. À la section 6, nous nous penchons brièvement sur l'estimation de la variance. À la section 8, nous discutons de l'application des méthodes à l'EPA. Enfin, à la section 9, nous présentons certaines conclusions.

2. Le problème d'estimation

Soit y_i la valeur (vraie) d'une variable d'intérêt associée à l'unité i dans une population finie U. La fonction de distribution de la variable dans U est :

$$F(y) = N^{-1} \sum_{i \in U} I(y_i \le y), \qquad (1)$$

où $I(.)$ est la fonction de vérité ($I(E) = 1$ si E est vrai et $= 0$ autrement) et y peut prendre toute valeur spécifiée. Supposons qu'une enquête soit réalisée auprès d'un échantillon $s \subset U$ et que la variable soit mesurée sous la forme y_i^* pour les unités $i \in s$. La différence entre y_i^* et y_i représente l'erreur de mesure. Supposons que la valeur vraie y_i soit enregistrée pour un sous-ensemble d'unités échantillonnées et que nous écrivions $r_i = 1$ si y_i est enregistré et $r_i = 0$ autrement. Soit x_i un vecteur de variables auxiliaires également enregistrées durant l'enquête. Nos données comprennent les valeurs de y_i^*, x_i et r_i pour les unités $i \in s$ et les valeurs y_i pour les unités $i \in s$ quand $r_i = 1$. Le problème est de savoir comment utiliser ces données pour faire une inférence au sujet de $F(y)$.

Dans l'application de l'EPA, les unités sont les employés, s est l'ensemble d'unités répondantes dans l'échantillon de l'EPA, y_i^* est la valeur de la variable dérivée de rémunération horaire et y_i est la valeur de la variable directe pour l'employé i. Nous supposons que la valeur y_i est égale à la rémunération horaire vraie.

La caractéristique de ce problème d'inférence qui nous intéresse principalement est l'absence de certaines valeurs y_i et nous envisageons deux approches pour traiter ces données manquantes, à savoir :

- l'imputation de y_i pour les unités $i \in s$ où $r_i = 0$, en utilisant les valeurs y_i^* et x_i comme données auxiliaires;
- la pondération d'un estimateur fondé sur le sous-échantillon de répondants $s_1 = \{i \in s;\ r_i = 1\}$, en particulier, l'utilisation de la pondération par le score de propension (Little 1986).

Nous discuterons de ces approches de l'estimation de $F(y)$ aux deux sections qui suivent.

Nous discuterons de l'inférence dans un cadre fondé sur un modèle, dans lequel nous supposons que les valeurs de population (y_i, y_i^*, x_i, r_i), $i \in U$, sont indépendantes et de même loi (IID) et que l'échantillonnage est ignorable, c'est-à-dire que la distribution de (y_i, y_i^*, x_i, r_i) est la même que $i \in s$ ou non. À la section 8, nous expliquerons comment la méthode élaborée sous ces hypothèses peut être adaptée aux conditions du plan d'échantillonnage de l'EPA et au recours à la pondération pour tenir compte de la non-réponse totale durant l'enquête.

3. Approches d'imputation

Supposons au départ qu'il est possible d'observer y_i pour tout $i \in s$. Alors, sous les hypothèses énoncées à la section qui précède,

$$\hat{F}(y) = n^{-1} \sum_{i=1}^{n} I(y_i < y) \qquad (2)$$

sera un estimateur sans biais de $F(y)$, en ce sens que $E[\hat{F}(y) - F(y)] = 0$ pour tout y, où nous écrivons $s = \{1, ..., n\}$ et où l'espérance est obtenue par rapport au modèle, sachant l'échantillon sélectionné s. Pour résoudre le problème dû au fait que y_i manque quand $r_i = 0$, supposons que y_i est remplacé dans (2) par une valeur imputée y_i^I quand $r_i = 0$ (et $i \in s$) et que $\tilde{y}_i = y_i$ si $r_i = 1$ et $\tilde{y}_i = y_i^I$ autrement. L'estimateur résultant de $F(y)$ est

$$\tilde{F}(y) = n^{-1} \sum_{i=1}^{n} I(\tilde{y}_i < y). \qquad (3)$$

Une condition suffisante pour que $\tilde{F}(y)$ soit un estimateur sans biais de $F(y)$ est que la distribution conditionnelle de y_i^I sachant $r_i = 0$, dénotée $f(y_i^I \mid r_i = 0)$, soit la même que la distribution conditionnelle $f(y_i \mid r_i = 0)$.

Cependant, puisque y_i n'est observé que si $r_i = 1$, les données ne fournissent aucun renseignement direct au sujet de $f(y_i \mid r_i = 0)$ si nous n'émettons pas d'hypothèses supplémentaires. Nous considérons deux hypothèses possibles.

Hypothèse (MAR) : r_i et y_i sont conditionnellement indépendants sachant y_i^* et x_i.

Hypothèse (modèle commun d'erreur de mesure) : r_i et y_i^* sont conditionnellement indépendants sachant y_i et x_i.

La première hypothèse est celle faite classiquement lorsque l'on recourt à l'imputation ou à la pondération (Little et Rubin 2002) et celle que nous formulerons ici. La deuxième hypothèse est celle voulant que le modèle d'erreur de mesure, défini comme étant la distribution conditionnelle de y_i^* sachant y_i et x_i, soit le même pour les répondants $(r_i = 1)$ que pour les non-répondants $(r_i = 0)$. Nous utiliserons la deuxième hypothèse pour l'étude en simulation à la section 7 afin d'évaluer la robustesse des méthodes fondées sur le mécanisme MAR. Cependant, sous la deuxième hypothèse, l'inférence est plus difficile et semble nécessiter des hypothèses de modélisation plus fortes au sujet de la distribution de y_i et x_i; nous étudions ce problème dans le cadre d'autres travaux et ne poursuivons pas son examen ici. La vraisemblance de ces deux hypothèses pour l'application de l'EPA est examinée plus en détail dans Skinner et coll. (2002).

Sous l'hypothèse de mécanisme MAR, nous avons $f(y_i \mid y_i^*, x_i, r_i = 0) = f(y_i \mid y_i^*, x_i, r_i = 1)$ et une condition suffisante pour que $\tilde{F}(Y)$ donne une estimation sans biais de $F(Y)$ est que

$$f(y_i^I \mid y_i^*, x_i, r_i = 0) = f(y_i \mid y_i^*, x_i, r_i = 1). \qquad (4)$$

Par conséquent, nous envisageons une approche d'imputation où la distribution conditionnelle de y sachant y^* et x est « ajustée » aux données sur les répondants $(r_i = 1)$, puis les valeurs imputées y_i^I sont « tirées » à partir de cette distribution ajustée, aux valeurs y_i^* et x_i observées pour les non-répondants. Supposons que la distribution conditionnelle $f(y_i \mid y_i^*, x_i, r_i = 1)$ puisse être représentée par un modèle de régression paramétrique tel que :

$$g(y_i) = h(y_i^*, x_i; \beta) + e_i, \; E(e_i \mid y_i^*, x_i) = 0 \qquad (5)$$

où $g(.)$ et $h(.)$ sont des fonctions données et β est un vecteur de paramètres de régression. Un prédicteur ponctuel de y_i, étant donné un estimateur $\hat{\beta}$ de β basé sur les données des répondants, est

$$\hat{y}_i = g^{-1}[h(y_i^*, x_i; \hat{\beta})]. \qquad (6)$$

Toutefois, l'utilisation de $\hat{y}_i$ pour l'imputation peut entraîner une sous-estimation importante de $F(y)$ pour les faibles valeurs de y, puisque une simple imputation par la régression de ce genre devrait, en principe, réduire la variation de $F(y)$ artificiellement (Little et Rubin 2002, page 64). Cet effet pourrait être évité en prenant $y_i^I = g^{-1}[h(y_i^*, x_i; \hat{\beta}) + \hat{e}_i]$, où $\hat{e}_i$ est un résidu empirique sélectionné aléatoirement (Little et Rubin 2002, page 65). Néanmoins, selon notre expérience, cette approche ne permet pas de générer des valeurs imputées qui reproduisent les « pics » des distributions de la rémunération horaire dans notre application et peut donner lieu à un biais autour de ces pics. Nous préférons par conséquent nous limiter aux méthodes d'imputation par donneur, où l'on fixe $y_i^I = y_{d(i)}$ $(r_i = 0)$ pour un répondant donneur $j = d(i)$ pour lequel $r_j = 1$. La valeur imputée d'après le donneur sera toujours une valeur authentique et respectera les pics de la distribution dans notre application. La méthode fondamentale d'imputation par donneur que nous considérons ici est l'imputation par appariement d'après la moyenne prévisionnelle (Little 1988), c'est-à-dire l'imputation par la méthode du plus proche voisin par rapport à $\hat{y}_i$, définie par (6), c'est-à-dire

$$\text{imputer } y_i \text{ par } y_{d(i)} \text{ en satisfaisant } |\hat{y}_i - \hat{y}_{d(i)}| = \min_{j:\, r_j = 1} |\hat{y}_i - \hat{y}_j| \qquad (7)$$

où $r_i = 0$ et $r_{d(i)} = 1$.

Le corollaire 2 du théorème 1 de Chen et Shao (2000) nous donne alors la justification théorique de l'absence approximative de biais dans l'estimateur résultant $\tilde{F}(y)$ de $F(y)$, si les quatre contraintes suivantes sont vérifiées : i) y_i manque au hasard (MAR) sachant $z_i = g^{-1}[h(y_i^*, x_i; \beta)]$, où $\beta = \text{plim}(\hat{\beta})$, ii) l'espérance conditionnelle de y_i sachant z_i est monotone et continue en z_i, iii) les troisièmes moments de z_i et $E(y_i \mid z_i)$ sont finis et iv) la probabilité de réponse sachant z est bornée au-dessus de zéro. Ces contraintes semblent plausibles à condition que l'hypothèse MAR susmentionnée tienne, que la distribution de y_i dépende uniquement de y_i^* et x_i par la voie de z_i et que y_i^* soit une approximation raisonnablement bonne de y_i. En outre, le résultat de Chen et Shao (2000) doit être adapté au fait que le plus proche voisin est défini par rapport à $\hat{\beta}$, tandis que les contraintes susmentionnées sont énoncées par rapport à β. Cette adaptation semble vraisemblable puisque, pour un nombre suffisamment grand de répondants, les voisins proches par rapport à $\hat{y}_i = g^{-1}[h(y_i^*, x_i; \hat{\beta})]$ devraient également être des voisins proches par rapport à $z_i = g^{-1}[h(y_i^*, x_i; \beta)]$.

Des fondements théoriques sous-tendent la notion que l'imputation selon la méthode du plus proche voisin par rapport à $\hat{y}_i$ donnera un estimateur approximativement sans biais de $F(y)$, sous l'hypothèse MAR et certaines autres conditions plausibles. Il est également intéressant de considérer l'efficacité de $\tilde{F}(y)$. La variance de $\tilde{F}(y)$ pour l'imputation par le plus proche voisin pourrait être exagérée si l'on utilise certains donneurs plus fréquemment que

d'autres. Nous considérons plusieurs approches en vue de réduire cet effet d'inflation de la variance.

En premier lieu, nous pouvons restreindre le nombre de fois que des répondants sont utilisés comme donneurs en définissant des classes d'imputation au moyen d'intervalles disjoints de valeurs de $\hat{y}_i$ et en tirant des donneurs pour un receveur donné par échantillonnage aléatoire simple dans la classe dans laquelle est comprise la valeur $\hat{y}_i$ du receveur. Le lissage sera le plus important si nous tirons les donneurs sans remise. Nous dénotons cette méthode hot deck IHDAR ou IHDSR, selon que l'échantillonnage est fait avec ou sans remise. Une deuxième approche consiste à sélectionner les donneurs séquentiellement et à pénaliser la fonction de distance $d(i)$ employée pour déterminer le plus proche voisin comme suit

$$|\hat{y}_i - \hat{y}_{d(i)}|(1+\mu t_{d(i)}) = \min_{j:\, r_j=1}\{|\hat{y}_i - \hat{y}_j|(1+\mu t_j)\}, \quad (8)$$

où $\mu \in \mathbb{R}^+$ est un facteur de pénalité, t_j est le nombre de fois que le répondant j a déjà été utilisé comme donneur, $r_i = 0$ et $r_{d(i)} = 1$ (Kalton 1983). Une troisième approche consiste à employer des valeurs imputées répétées $y_i^{I(m)}$, $m = 1, ..., M$, pour chaque receveur $i \in s$, telles que $r_i = 0$. L'estimateur résultant de $F(y)$ est $M^{-1}\sum_m \tilde{F}^{(m)}(y)$, la moyenne des estimateurs résultants $\tilde{F}^{(m)}(y)$. Nous dénommons cette troisième approche imputation fractionnaire (Kalton et Kish 1984; Fay 1996) plutôt qu'imputation multiple (Rubin 1996), parce qu'il n'est pas nécessaire que notre méthode d'imputation soit « appropriée », c'est-à-dire qu'elle remplisse les conditions assurant que l'estimateur de la variance par imputation multiple soit convergent. Nous ne stipulons pas cette exigence ici, parce que notre objectif premier est l'estimation ponctuelle. Lorsque nous utilisons l'imputation fractionnaire, nous visons à sélectionner des donneurs $d(i, m)$, $m = 1, ..., M$ qui sont chacun un voisin proche de i, de sorte que $\tilde{F}^{(m)}(y)$ demeure approximativement sans biais pour $F(y)$. Nous examinons les variantes suivantes de cette approche.

i) Les $M/2$ voisins les plus proches au-dessus et au-dessous de la valeur $\hat{y}_i$ sont tirés, pour $M = 2$ ou 10, et dénotés PPV2 et PPV10, respectivement.

ii) $M/2$ donneurs sont sélectionnés par échantillonnage aléatoire simple avec remise parmi les M répondants pour lesquels la valeur est supérieure à $\hat{y}_i$ et parmi les M répondants pour lesquels elle est inférieure, pour $M = 2$ ou 10, et nous les dénotons PPV2(4) et PPV10(20), respectivement.

iii) $M = 10$ donneurs sont sélectionnés par échantillonnage aléatoire simple avec ou sans remise dans les classes d'imputation auxquelles nous avons fait allusion dans les méthodes IHDAR et IHDSR décrites plus haut. Nous nommons ces méthodes IHDAR10 et IHDSR10.

Aux fins de comparaison, nous envisageons aussi la méthode bootstrap bayésienne approximative d'imputation multiple (Rubin et Schenker 1986), dénotée BBA10, définie par rapport aux classes d'imputation mentionnées au sujet des méthodes IHDAR et IHDSR.

4. Estimation pondérée

L'estimateur $\tilde{F}(y)$ implicite dans les diverses approches d'imputation envisagées à la section précédente peut être exprimé sous la forme pondérée :

$$\tilde{F}(y) = \sum_{i\in s_1} w_i\, I(y_i < y) \Big/ \sum_{i\in s_1} w_i, \quad (9)$$

où $s_1 = \{i \in s;\ r_i = 1\}$ est l'ensemble de répondants et $w_i = 1 + d_i/M$, où d_i est le nombre total de fois que le répondant i est utilisé comme donneur sur les M imputations répétées. Notons que $\sum_{s_1} w_i = n$. Un autre choix de pondération consisterait à fixer que w_i est égal à l'inverse de la valeur estimée du score de propension, $\Pr(r_i = 1 | y_i^*, x_i)$ (Little 1986). Cette approche a été proposée pour l'estimation de la rémunération horaire par Dickens et Manning (2004). Le score de propension pourrait être estimé, par exemple, sous un modèle de régression logistique reliant r_i à y_i^* et x_i. Sous l'hypothèse MAR, l'estimateur résultant $\tilde{F}(y)$ sera approximativement sans biais si l'on suppose que le modèle est valide pour la distribution conditionnelle $f(r_i | y_i^*, x_i)$ et certaines conditions de régularité, telles que celles décrites à la section 3 pour l'estimateur imputé. Notons que la nécessité de modéliser $f(r_i | y_i^*, x_i)$ remplace la nécessité de modéliser $f(y_i | y_i^*, x_i)$ dans l'approche d'imputation.

5. Propriétés des approches d'imputation et de pondération

À la présente section, nous investiguons et comparons les propriétés théoriques de l'approche d'imputation et de l'approche de pondération par le score de propension présentées aux deux sections précédentes sous diverses hypothèses simplificatrices. Nous fixons y et établissons $u_i = I(y_i < y)$. Posant que $N \to \infty$, nous supposons que le paramètre d'intérêt est $\theta = E(u_i)$. Nous considérons l'approche d'imputation pour commencer et supposons que y_i dépend de y_i^* et x_i uniquement par la voie de $z_i = g^{-1}[h(y_i^*, x_i; \beta]$ et que y_i manque au hasard, sachant z_i. En ignorant la différence entre β et $\hat{\beta}$, et en supposant que s_1 est grand, nous considérons l'imputation par le plus

proche voisin par rapport à z_i. Comme dans (9), l'estimateur imputé de θ peut être exprimé sous la forme

$$\hat{\theta}_{\text{IMP}} = \sum_{i \in s_1} w_i u_i \Big/ \sum_{i \in s_1} w_i \qquad (10)$$

où $w_i = 1 + d_i / M$ (et $\sum_{s_1} w_i = n$). Nous écrivons l'expression correspondante pour la pondération par le score de propension sous la forme $\hat{\theta}_{\text{PS}}$ avec w_i remplacé par $w_{\text{PS}i}$. Représentons par $z_{\text{PS}i}$ la fonction scalaire de y_i^*, x_i dont dépend r_i et écrivons :

$$\Pr(r_i = 1 \mid y_i^*, x_i) = \pi(z_{\text{PS}i}). \qquad (11)$$

Tout comme nous avons ignoré l'écart entre β et $\hat{\beta}$, nous ignorons au départ l'erreur lors de l'estimation de $\pi(z_{\text{PS}i})$ et écrivons $w_{\text{PS}i} = \pi(z_{\text{PS}i})^{-1}$.

Nous pouvons nous attendre à ce que les approches d'imputation et de pondération par le score de propension produisent des estimateurs semblables si z_i et $z_{\text{PS}i}$ sont semblables, c'est-à-dire s'ils sont proches de fonctions déterministes l'un de l'autre, et que M est grand. Pour le montrer, considérons un exemple simple de l'approche d'imputation, où le donneur est tiré aléatoirement dans une classe d'imputation c de voisins proches par rapport à z_i, contenant m_c répondants et $n_c - m_c$ non-répondants, comme il est décrit à la section 3. Dans ce cas, w_i s'approchera de $1 + (n_c - m_c)/m_c = n_c / m_c$ lorsque $M \to \infty$ et il s'agit de l'inverse du taux de réponse dans la classe (David, Little, Samuhel et Triest 1983). De façon plus générale, si nous suivons l'approche d'imputation fractionnaire par le plus proche voisin envisagée à la section 3, le poids $w_i = 1 + d_i / M$ peut être interprété comme une estimation locale (par rapport à z_i) non paramétrique de $\Pr(r_i = 1 \mid z_i)^{-1}$, malgré le fait que l'imputation est basée sur un modèle pour y_i sachant z_i plutôt que pour r_i sachant z_i. Donc, nous pouvons nous attendre à ce que la méthode d'imputation mène à des résultats d'estimation semblables à la méthode de pondération par le score de propension si z_i et $z_{\text{PS}i}$ sont des fonctions déterministes l'un de l'autre. Cependant, en général, il n'en est pas ainsi. Puisque $\Pr(r_i = 1 \mid z_i)$ peut être exprimé comme une moyenne de $\Pr(r_i = 1 \mid y^*, x)$ sur les valeurs de y^* et x pour lesquels $z = z_i$, , nous pouvons interpréter w_i comme étant une version lissée de $w_{\text{PS}i}$ et pouvons nous attendre à ce que sa dispersion soit plus faible. Cela donne à penser qu'il pourrait être possible d'utiliser l'imputation pour améliorer l'efficacité des estimations fondées sur la pondération par le score de propension, comme en ont déjà discuté David et coll. (1983) et Rubin (1996, section 4.6). Pour étudier davantage cette possibilité, émettons l'hypothèse MAR, ainsi que les autres hypothèses formulées aux sections 3 et 4 sur lesquelles les approches sont fondées, de sorte que les approches d'imputation et de pondération mènent toutes deux à une estimation approximativement sans biais de $F(y)$ et que nous puissions donc nous concentrer sur la comparaison des efficacités relatives.

Il découle de l'équation (3.3) de Chen et Shao (2000) que la variance de $\hat{\theta}_{\text{IMP}}$ peut être approximée, pour une grande valeur de n, par

$$\text{var}(\hat{\theta}_{\text{IMP}}) \approx n^{-2} E\Big[\sum_{s_1} w_i^2 V(u_i \mid z_i)\Big] + n^{-1} V[\psi(z_i)\Big], \quad (12)$$

où $\psi(z_i) = E(u_i \mid z_i)$ et tout effet de l'estimation de β est ignoré. Notons que Chen et Shao (2000) considèrent le cas de l'imputation unique avec $M = 1$, mais que leur preuve de ce résultat est transposable si $M > 1$. Il est commode de réexprimer ce résultat sous la forme

$$\text{var}(\hat{\theta}_{\text{IMP}}) \approx n^{-1}\sigma^2 + n^{-2} E\Big[\sum_{s_1} (w_i^2 - w_i) V(u_i \mid z_i)\Big], \quad (13)$$

en utilisant l'identité

$$V[\psi(z_i)] = \sigma^2 - E[V(u_i \mid z_i)], \qquad (14)$$

où $\sigma^2 = V(u_i)$ et un corollaire du théorème 1 de Chen et Shao (2000) selon lequel

$$E\Big[n^{-1}\sum_{s_1} w_i V(u_i \mid z_i)\Big] = E[V(u_i \mid z_i)] + o_p(n^{-1/2}). \quad (15)$$

Notons que $w_i^2 - w_i = (d_i / M)(1 + d_i / M) \geq 0$. L'expression (13) peut être interprétée sous l'angle des « données manquantes » ainsi que sous celui de l'« erreur de mesure ». Du point de vue des données manquantes, le premier terme de (13) est simplement la variance de $\hat{\theta}$ en l'absence de données manquantes et le deuxième terme représente l'inflation de cette variance due à l'erreur d'imputation. Du point de vue de l'erreur de mesure, nous pouvons considérer les propriétés limites sous les « conditions asymptotiques de faible erreur de mesure » (Chesher 1991), c'est-à-dire quand $y_i^* \to y_i$ et que $V(u_i \mid z_i)$ s'approche de zéro. Dans ce cas, le deuxième terme tend aussi vers zéro et $\hat{\theta}_{\text{IMP}}$ devient « entièrement efficace », c'est-à-dire que sa variance s'approche de σ^2 / n.

Considérons maintenant la pondération par le score de propension. Nous formulons l'hypothèse correspondante que y_i manque au hasard sachant $z_{\text{PS}i}$. En linéarisant le ratio de l'expression (9), avec $w_{\text{PS}i}$ à la place de w_i, en utilisant le fait que $E(\sum_{s_1} w_{\text{PS}i}) = n$ et en ignorant au départ l'effet de l'estimation du score de propension, nous pouvons écrire

$$\text{var}(\hat{\theta}_{\text{PS}}) \approx n^{-2}\,\text{var}\Big[\sum_{s_1} w_{\text{PS}i}(u_i - \theta)\Big] = n^{-1}\, E[w_{\text{PS}i}(u_i - \theta)^2], \qquad (16)$$

que nous pouvons aussi exprimer sous la forme

$$\text{var}(\hat{\theta}_{\text{PS}}) \approx n^{-2} E\left[\sum_{s_1} w_{\text{PS}i}^2 V(u_i \mid \text{z}_{\text{PS}i})\right] + n^{-1} E\{w_{\text{PS}i}[\psi(\text{z}_{\text{PS}i}) - \theta]^2\}. \quad (17)$$

Pour comparer l'efficacité de la pondération et de l'imputation, il est commode d'utiliser les expressions (14) et (15) (qui sont également vraies si $w_{\text{PS}i}$ remplace w_i) pour obtenir

$$\text{var}(\hat{\theta}_{\text{PS}}) \approx n^{-1}\sigma^2 + n^{-2} E\left[\sum_{s_1} (w_{\text{PS}i}^2 - w_{\text{PS}i}) V(u_i \mid \text{z}_{\text{PS}i})\right] + n^{-1} E\left\{\sum_{s_1} [w_{\text{PS}i} - 1]\,[\psi(\text{z}_{\text{PS}i}) - \theta]^2\right\}. \quad (18)$$

Notons que, comparativement à (13), cette expression contient un troisième terme, qui ne converge pas nécessairement vers zéro quand y_i^* s'approche de y_i et que $V(u_i \mid \text{z}_{\text{PS}i}) \to 0$. Donc, la pondération par le score de propension ne devient pas entièrement efficace quand l'erreur de mesure disparaît. Nous pouvons aussi nous attendre à ce que le deuxième terme de (18) domine le deuxième terme de (13) quand $V(u_i \mid \text{z}_i)$ et $V(u_i \mid \text{z}_{\text{PS}i})$ sont constantes et égales, puisque, en se souvenant que $\sum_{s_1} w_i = E(\sum_{s_1} w_{\text{PS}i}) = n$, ces deuxièmes termes sont principalement déterminés par les variances des coefficients de pondération w_i et $w_{\text{PS}i}$, et que, à condition que M soit suffisamment grand, nous pouvons nous attendre à ce que w_i soit moins variable que $w_{\text{PS}i}$, comme nous l'avons soutenu plus haut.

La discussion qui précède ne tient pas compte de l'effet éventuel de l'estimation de β ou de l'estimation d'un vecteur de paramètre α dont on peut supposer que dépend le score de propension $\Pr(r_i = 1 \mid y_i^*, x_i)$. Kim (2004) montre, en fait, que l'estimation de α par son estimateur du maximum de vraisemblance $\hat{\alpha}$ réduit la variance de $\hat{\theta}_{\text{PS}}$ comme suit :

$$\text{var}(\hat{\theta}_{\text{PS}}) \approx \text{var}(\tilde{\theta}_{\text{PS}}) - \text{cov}(\tilde{\theta}_{\text{PS}}, \hat{\alpha})\,\text{var}(\hat{\alpha})^{-1}\,\text{cov}(\hat{\alpha}, \tilde{\theta}_{\text{PS}}), \quad (19)$$

où $\tilde{\theta}_{\text{PS}}$ est l'estimateur $\hat{\theta}_{\text{PS}}$ dans lequel les scores de propension estimés sont remplacés par leur valeur réelle et où le premier membre de (16), (17) et (18) devrait maintenant être $\text{var}(\tilde{\theta}_{\text{PS}})$. Nous concluons de ce fait et de la discussion qui précède qu'en général, $\hat{\theta}_{\text{IMP}}$ n'est pas forcément plus efficace que $\hat{\theta}_{\text{PS}}$ ou inversement, et nous nous appuyons sur l'étude en simulation présentée à la section 7 pour des preuves numériques. Cependant, la conclusion que $\hat{\theta}_{\text{IMP}}$ devient plus efficace à mesure que l'erreur de mesure disparaît et que $y_i^* \to y_i$ demeure valide, même en présence d'une erreur d'estimation dans α et dans β, puisque l'effet de l'erreur d'estimation dans β disparaîtra dans ce cas lorsque $\text{z}_i \to y_i^*$, tandis que le deuxième terme de (19), lorsqu'il est ajouté à l'expression (18), ne réduira généralement pas $\text{var}(\hat{\theta}_{\text{PS}})$ à σ^2 / n dans ce cas.

Enfin, considérons l'effet des écarts par rapport à l'hypothèse de mécanisme MAR. Sous des conditions asymptotiques de faible erreur de mesure, où $y_i^* \to y_i$ et $V(u_i \mid \text{z}_i) \to 0$, de sorte que $y_i^I \to y_i$, l'approche d'imputation donnera une inférence convergente au sujet de θ, même si l'hypothèse de mécanisme MAR ne tient pas. Par contre, ce n'est pas le cas pour l'approche de pondération par le score de propension. Cela donne à penser que l'approche d'imputation pourrait être plus robuste aux écarts par rapport à l'hypothèse MAR si l'erreur de mesure est relativement faible.

6. Estimation de la variance

Bien que l'estimation ponctuelle soit le sujet principal du présent article, nous allons maintenant considérer brièvement l'estimation de la variance par linéarisation. Dans le cas de la pondération par le score de propension, nous nous référons aux travaux de Kim (2004). Pour les méthodes d'imputation simple ou fractionnaire fondées sur l'imputation par le plus proche voisin décrites à la section 3, nous pouvons considérer une approche simplifiée basée sur l'hypothèse IID formulées à la section 2 et l'expression de la variance de $\hat{\theta}_{\text{IMP}}$ dans (13).

L'estimateur simple du premier terme σ^2 / n :

$$n^{-1}\hat{\sigma}^2 = n^{-2}\sum_{s_1} w_i (u_i - \hat{\theta}_{\text{IMP}})^2 \quad (20)$$

est approximativement sans biais d'après le corollaire 1 de Chen et Shao (2000). Il s'ensuit qu'un estimateur approximativement sans biais de $\text{var}(\hat{\theta}_{\text{IMP}})$ est

$$\hat{V}(\hat{\theta}_{\text{IMP}}) = n^{-1}\hat{\sigma}^2 + n^{-2}\sum_{s_1} (w_i^2 - w_i)\,\hat{V}(u_i \mid \text{z}_i) \quad (21)$$

si nous pouvons construire un estimateur approximativement sans biais $\hat{V}(u_i \mid \text{z}_i)$ de $V(u_i \mid \text{z}_i)$. Diverses approches d'estimation de $V(u_i \mid \text{z}_i)$ semblent possibles. À l'instar de Fay (1999), nous pourrions prendre en considération la variance d'échantillon de u_j valeurs pour les voisins répondants proches de i par rapport à z. Une autre approche serait d'adopter une méthode fondée sur un modèle selon laquelle un modèle est ajusté à $\psi(\text{z}_i) = E(u_i \mid \text{z}_i)$ pour $i \in s$ sachant $\hat{\psi}(\text{z}_i)$ et de fixer $\hat{V}(u_i \mid \text{z}_i) = \hat{\psi}(\text{z}_i)[1 - \hat{\psi}(\text{z}_i)]$. Nous avons envisagé des méthodes non paramétriques d'ajustement de $\psi(\text{z}_i)$, mais avons constaté qu'avec les données de l'EPA, elles mènent à des valeurs de $\hat{V}(\hat{\theta}_{\text{IMP}})$ fort semblables à celles produites par un modèle de régression logistique pour $\psi(\text{z}_i)$.

Il pourrait être possible d'appliquer les idées de Chen et Shao (2001) ou de Kim et Fuller (2002) en vue d'étendre l'approche susmentionnée de façon à pouvoir tenir compte des poids de sondage et du plan de sondage complexe. Consulter Rancourt (1999) et Fay (1999) pour d'autres

approches d'estimation de la variance sous imputation par le plus proche voisin, ainsi que Little et Rubin (2002) pour les approches d'imputation multiple.

7. Étude par simulation

Le but de l'étude est de générer des échantillons répétés indépendants $s^{(h)}$, $h = 1, ..., H$, avec des valeurs y_i, y_i^*, x_i, r_i, $i \in s^{(h)}$ réalistes dans le contexte de l'application de l'EPA, examinée plus loin à la section 8, afin de calculer les estimations correspondantes $\tilde{F}^{(h)}(y)$ pour diverses approches de traitement des valeurs manquantes de y et d'évaluer empiriquement les propriétés des estimateurs $\tilde{F}(y)$. Afin d'utiliser des valeurs réalistes, nous avons tiré les échantillons $s^{(h)}$ de taille n avec remise (c'est-à-dire par la méthode du bootstrap) à partir d'un échantillon réel d'environ 16 000 employés pour le trimestre de mars à mai 2000 utilisé pour l'EPA (seuls les emplois principaux des employés de 18 ans et plus ont été pris en considération et le très petit nombre de cas pour lesquels des valeurs de y_i^* ou x_i manquaient ont été omis). Les valeurs de x_i pour chaque échantillon $s^{(h)}$ ont été tirées directement des valeurs dans l'échantillon de l'EPA. Les critères utilisés pour choisir les variables incluses dans x_i étaient qu'elles soient corrélées à la rémunération horaire, à l'erreur de mesure dans y_i^* ou à la réponse r_i (voir Skinner et coll. 2002). Ces variables comprennent, par exemple l'âge, le sexe, la position dans le ménage, les qualifications, la profession, la durée de l'emploi, le travail à temps plein/temps partiel, l'industrie et la région (plusieurs de ces variables ont été représentées par des variables muettes). Nous avons fixé $n = 15\,000$, de sorte que chaque $s^{(h)}$ soit de taille similaire à celle de l'échantillon original de l'EPA, et $H = 1\,000$. Les valeurs de y_i, y_i^* et r_i pour chaque échantillon $s^{(h)}$ ont été générées d'après des modèles, plutôt que directement d'après les données de l'EPA, pour les raisons qui suivent.

y_i : ces valeurs ont été générées d'après un modèle, parce qu'elles manquaient fréquemment dans l'EPA. Nous avons utilisé un modèle de régression linéaire reliant $\ln(y_i)$ à $\ln(y_i^*)$ et x_i avec une erreur normale et avec 20 covariables, y compris des termes quadratiques pour $\ln(y_i^*)$ et l'âge, et des termes d'interaction entre $\ln(y_i^*)$ et cinq composantes de x_i. Le modèle a été ajusté aux 7 000 cas environ pour lesquels y_i avait été observé.

y_i^* : ces valeurs ont été générées d'après un modèle afin d'éviter les valeurs en double de (y_i^*, x_i) dans chaque échantillon $s^{(h)}$, ce qui, à notre avis, aurait pu donner lieu à une distribution irréaliste des distances entre les unités pour la méthode du plus proche voisin. Nous avons utilisé un modèle de régression linéaire reliant $\ln(y_i^*)$ à x_i avec une erreur normale et avec 12 covariables, y compris un terme quadratique pour l'âge et un terme d'interaction, que nous avons ajusté aux données de l'EPA.

r_i : ces valeurs ont été générées d'après un modèle afin d'assurer que le mécanisme de production des données manquantes soit connu. Nous avons ajusté plusieurs modèles. Le seul présenté ici est une régression logistique reliant r_i à $\ln(y_i^*)$ et à x_i avec 17 covariables, y compris le carré de $\ln(y_i^*)$ et des termes d'interactions entre $\ln(y_i^*)$ et deux covariables. Nous avons ajusté le modèle aux données de l'EPA. Le mécanisme de production des données manquantes est MAR sachant les y_i^* et x_i pour tous les résultats présentés, sauf ceux du tableau 5.

Nous avons obtenu les estimations $\hat{\theta}_t^{(h)}$ de deux paramètres $(t = 1, 2)$ pour chaque échantillon $s^{(h)}$,

θ_1 = proportion de travailleurs dont la rémunération est inférieure au salaire minimum national (= 3,00 £ de l'heure pour les personnes de 18 à 21 ans, 3,60 £ de l'heure pour les personnes de 22 ans et plus)

θ_2 = proportion de travailleurs dont la rémunération est comprise entre le salaire minimum et 5 £ de l'heure.

Les valeurs réelles sont $\theta_1 = 0,056$ et $\theta_2 = 0,185$. Nous avons estimé le biais et l'erreur-type sous les formes $\hat{\text{biais}}(\hat{\theta}_t) = \bar{\theta}_t - \theta_t$ et $\hat{e}.-t.(\hat{\theta}_t) = [H^{-1}\sum_{h=1}^{H}(\hat{\theta}_t^{(h)} - \bar{\theta}_t)^2]^{1/2}$, où $\bar{\theta}_t = H^{-1}\sum_h \theta_t^{(h)}$.

Dans le cas des méthodes d'imputation fractionnaire, nous avons examiné plusieurs valeurs de M et avons choisi $M = 10$ ou 20 afin d'obtenir un accroissement de l'efficacité, tout en restant capable de définir raisonnablement une imputation par le plus proche voisin.

Nous commençons par comparer les résultats pour les diverses approches d'imputation. Le tableau 1 donne les estimations des biais des estimateurs de θ_1 et θ_2 pour diverses méthodes d'imputation, sous un mécanisme de données manquantes MAR. Nous ne dégageons aucune preuve d'un biais important pour aucune des méthodes par le plus proche voisin (PPV). Les ratios biais/erreur-type sont faibles et nous pouvons nous attendre à ce qu'ils soient encore plus petits pour les estimations pour des domaines tels que les régions ou les groupes d'âge. Nous concluons qu'il n'existe aucune preuve d'un biais important pour ces méthodes, à condition que l'hypothèse d'un mécanisme MAR tienne et que le modèle soit spécifié correctement.

Nous dégageons certaines preuves de l'existence d'un biais statistiquement significatif pour chacune des trois méthodes fondées sur les classes d'imputation (IHDAR10, IHDSR10, BBA10) peut-être dû à la largeur des classes, quoique le biais semble faible comparativement à l'erreur-type. Étant donné l'inconvénient supplémentaire que

représente la spécification arbitraire des bornes des classes, ces méthodes semblent moins intéressantes que les méthodes d'imputation par le plus proche voisin. Ce résultat va à l'encontre de la préférence parfois exprimée (par exemple Brick et Kalton 1996, page 227) pour les méthodes stochastiques d'imputation, telles que les méthodes IHD, comparativement aux méthodes déterministes, comme l'imputation par le plus proche voisin, pour estimer les paramètres de la distribution.

Les estimations correspondantes des erreurs-types sont présentées au tableau 2. Nous constatons, comme prévu, que la méthode d'imputation simple PPV1 est celle qui produit les erreurs-types les plus importantes. L'utilisation de la méthode avec fonction de pénalité (PPV1P) réduit la variance d'environ 10 %. L'utilisation de deux imputations (PPV2 ou PPV2(4)) donne une réduction de 10 % à 20 %, et l'utilisation de dix imputations (PPV10, PPV10(20)), IHDAR10, IHDSR10, BBA10), une réduction d'environ 20 %. Pour un nombre donné d'imputations (2 ou 10), il ne semble pas exister d'effet systématique manifeste de l'utilisation d'une méthode stochastique (PPV2(4) ou PPV10(20)) plutôt qu'une méthode déterministe (PPV2 ou PPV10). Nous nous attendrions à ce que les erreurs-types ne soient pas plus faibles pour IHDAR10 que pour IHDSR10, ce qui est le cas pour $\hat{\theta}_1$ au tableau 2. La légère réduction de l'erreur-type de l'estimateur $\hat{\theta}_2$ est vraisemblablement causée par un nombre comparativement faible d'itérations dans la simulation ($H = 1\,000$), nombre qui pourrait ne pas suffire entièrement pour l'estimation de l'erreur-type. Nous concluons que PPV10 est l'approche la plus prometteuse, car elle permet d'éviter le biais des méthodes basées sur des classes d'imputation et de réaliser des gains appréciables d'efficacité par rapport aux méthodes générant une ou deux imputations.

Nous allons maintenant comparer l'approche d'imputation PPV10 à la pondération par le score de propension (PSP). Nous considérons non seulement le cas où la spécification du modèle utilisé pour l'imputation ou pour la pondération correspond au modèle utilisé pour la simulation, comme au tableau 1, mais aussi certains cas de spécification incorrecte. Pour assurer que la comparaison de la pondération et de l'imputation soit équitable, nous utilisons les mêmes covariables lors de l'ajustement des deux modèles générant y_i et r_i. Nous considérons pour commencer les biais estimés présentés au tableau 3. Lorsque le modèle pour l'imputation (PPV10) ou pour la pondération par le score de propension est spécifié correctement, ni l'une ni l'autre méthode ne donne lieu à un biais significatif dans l'estimation de θ_1 ou θ_2. Toutefois, nous observons un biais significatif dans les deux cas, si le modèle est mal spécifié en oubliant d'inclure les covariables utilisées dans la simulation. Néanmoins, le biais est sensiblement plus important pour l'approche de pondération. Par exemple, pour l'estimateur $\hat{\theta}_1$, le biais est de 3 à 7 fois plus élevé sous la méthode PSP que sous la méthode PPV10, selon l'erreur de spécification. L'effet de l'erreur de spécification semble plus important pour l'estimateur $\hat{\theta}_2$, en particulier sous la méthode PSP. Dans le cas de cet estimateur, nous observons un biais de 6 à 15 fois plus grand pour cette dernière que pour la méthode PPV10.

Tableau 1
Estimations par simulation des biais des estimateurs de θ_1 et θ_2 pour diverses méthodes d'imputation, sous l'hypothèse de mécanisme MAR et de covariables correctes ($H = 1\,000$)

Méthode d'imputation	Biais de $\hat{\theta}_1$	Biais rel. de $\hat{\theta}_1$	Biais de $\hat{\theta}_2$	Biais rel. de $\hat{\theta}_2$
PPV1	$1{,}2*10^{-4}$	0,2 %	$0{,}9*10^{-4}$	0,0 %
	$(0{,}9*10^{-4})$		$(1{,}7*10^{-4})$	
PPV1P[1]	$4{,}4*10^{-4}$	0,8 %	$0{,}3*10^{-4}$	0,0 %
	$(2{,}6*10^{-4})$		$(5{,}1*10^{-4})$	
PPV2	$0{,}6*10^{-4}$	0,1 %	$1{,}6*10^{-4}$	0,0 %
	$(0{,}8*10^{-4})$		$(1{,}5*10^{-4})$	
PPV2(4)	$1{,}4*10^{-4}$	0,2 %	$-2{,}5*10^{-4}$	−0,1 %
	$(0{,}9*10^{-4})$		$(1{,}5*10^{-4})$	
PPV10	$0{,}2*10^{-4}$	0,0 %	$-1{,}2*10^{-4}$	−0,1 %
	$(0{,}8*10^{-4})$		$(1{,}5*10^{-4})$	
PPV10(20)	$0{,}2*10^{-4}$	0,0 %	$0{,}7*10^{-4}$	0,0 %
	$(0{,}8*10^{-4})$		$(1{,}5*10^{-4})$	
IHDAR10	$2{,}8*10^{-4}$	0,5 %	$26{,}2*10^{-4}$	1,4 %
	$(0{,}8*10^{-4})$		$(1{,}5*10^{-4})$	
IHDSR10	$2{,}5*10^{-4}$	0,4 %	$28{,}0*10^{-4}$	1,5 %
	$(0{,}8*10^{-4})$		$(1{,}5*10^{-4})$	
BBA10	$4{,}6*10^{-4}$	0,8 %	$29{,}8*10^{-4}$	1,6 %
	$(0{,}8*10^{-4})$		$(1{,}5*10^{-4})$	

Les erreurs-types des estimations du biais figurent entre parenthèses sous les estimations.
[1] Nota : $H = 100$ itérations ont été utilisées à cause du temps de calcul.

Les estimations correspondantes des erreurs-types de $\hat{\theta}_1$ et $\hat{\theta}_2$ sont données au tableau 4. Celles-ci ont aussi tendance à être plus importantes pour la méthode de pondération, l'accroissement étant de 5 % à 15 % par rapport à la méthode d'imputation. L'augmentation de l'erreur-type est plus importante pour le deuxième estimateur $\hat{\theta}_2$, variant de 12 % à 15 %, que pour l'estimateur $\hat{\theta}_1$, pour lequel l'accroissement est de 5 % à 12 %, selon l'erreur de spécification. Par conséquent, l'erreur quadratique moyenne est également plus grande pour la méthode de pondération, l'accroissement variant de 20 % à 28 % pour les six valeurs au tableau 4. Du moins sous l'hypothèse de mécanisme MAR, la méthode d'imputation PPV10 semble être préférable à la pondération par le score de propension en ce qui concerne le biais et la variance.

Enfin, nous comparons les propriétés des méthodes d'imputation (PPV10) et de pondération par le score de propension lorsque l'hypothèse de mécanisme MAR ne tient pas. Nous simulons maintenant l'absence de données conformément à l'hypothèse du modèle commun d'erreur de mesure décrit à la section 3. Nous utilisons le même modèle logistique avec les mêmes coefficients que pour la simulation précédente, excepté que y_i^* est remplacé par y_i à titre de covariable. Les estimations en simulation des biais et des erreurs-types sont présentées au tableau 5. Nous observons un biais relatif significatif non négligeable d'environ 5 % pour l'approche d'imputation et un peu plus élevé pour l'approche de pondération par le score de propension. La direction positive du biais de $\hat{\theta}_1$ est conforme aux attentes fondées sur les arguments de Dickens et Manning (2004) et de Skinner et coll. (2002). Les méthodes basées sur l'hypothèse MAR auront tendance à surestimer le nombre de travailleurs faiblement rémunérés, si l'hypothèse d'erreur de mesure commune tient. Il en est ainsi parce que les employés pour lesquels les valeurs y_i sont observées ont tendance à être moins bien rémunérés que ceux pour lesquels les valeurs y_i manquent et qu'une méthode d'imputation fondée sur l'hypothèse MAR, même connaissant les autres variables, aurait tendance à imputer des valeurs plus faibles de rémunération horaire que cela ne serait le cas sous l'hypothèse d'erreur de mesure commune qui permet la dépendance par rapport à la rémunération horaire réelle. Bien que l'on puisse prévoir la direction de l'effet, la grandeur de celui-ci a une certaine importance en ce qui concerne la robustesse des méthodes fondées sur l'hypothèse MAR. Le biais relatif de 5 % de la méthode PPV10 ne semble toutefois pas rendre les estimations résultantes inutilisables.

Tableau 2
Estimations par simulation des erreurs-types des estimateurs de θ_1 et θ_2 pour diverses méthodes d'imputation, sous l'hypothèse de mécanisme MAR et de covariables correctes ($H = 1\,000$)

Méthode d'imputation	e.-t.($\hat{\theta}_1$)	e.-t.($\hat{\theta}_2$)	$\frac{V(\hat{\theta}_1)}{V_{\text{PPV1}}(\hat{\theta}_1)}$	$\frac{V(\hat{\theta}_2)}{V_{\text{PPV1}}(\hat{\theta}_2)}$
PPV1	$2{,}79*10^{-3}$	$5{,}43*10^{-3}$	1	1
PPV1P[2]	$2{,}60*10^{-3}$	$5{,}15*10^{-3}$	0,87	0,91
PPV2	$2{,}68*10^{-3}$	$5{,}05*10^{-3}$	0,91	0,86
PPV2(4)	$2{,}73*10^{-3}$	$4{,}88*10^{-3}$	0,94	0,80
PPV10	$2{,}56*10^{-3}$	$4{,}88*10^{-3}$	0,83	0,81
PPV10(20)	$2{,}57*10^{-3}$	$4{,}79*10^{-3}$	0,84	0,77
IHDAR10	$2{,}52*10^{-3}$	$4{,}66*10^{-3}$	0,82	0,74
IHDSR10	$2{,}48*10^{-3}$	$4{,}72*10^{-3}$	0,78	0,76
BBA10	$2{,}63*10^{-3}$	$4{,}87*10^{-3}$	0,88	0,80

[2] Nota : $H = 100$ itérations ont été utilisées à cause du temps de calcul.

Tableau 3
Estimations par simulation des biais des estimateurs de θ_1 et θ_2 pour l'imputation par le plus proche voisin (PPV10) et la pondération par le score de propension, sous l'hypothèse de mécanisme MAR et de covariables correctes et spécifiées incorrectement ($H = 1\,000$)

Méthode	Covariables hypothétiques	Biais de $\hat{\theta}_1$	Biais rel. de $\hat{\theta}_1$	Biais de $\hat{\theta}_2$	Biais rel. de $\hat{\theta}_2$
PPV10	M1 (correct)	$-0{,}18*10^{-4}$ ($0{,}64*10^{-4}$)	– 0,03 %	$-5{,}8*10^{-4}$ ($1{,}20*10^{-4}$)	– 0,31 %
	M2	$-1{,}31*10^{-4}$ ($0{,}65*10^{-4}$)	– 0,24 %	$-4{,}74*10^{-4}$ ($1{,}23*10^{-4}$)	– 0,25 %
	M3	$-1{,}66*10^{-4}$ ($0{,}63*10^{-4}$)	– 0,30 %	$-10{,}6*10^{-4}$ ($1{,}23*10^{-4}$)	– 0,57 %
Pondération par le score de propension	M1 (correct)	$0{,}15*10^{-4}$ ($0{,}72*10^{-4}$)	0,03 %	$-2{,}62*10^{-4}$ ($1{,}35*10^{-4}$)	– 0,14 %
	M2	$-8{,}96*10^{-4}$ ($0{,}68*10^{-4}$)	– 1,64 %	$70{,}2*10^{-4}$ ($1{,}40*10^{-4}$)	3,80 %
	M3	$-5{,}02*10^{-4}$ ($0{,}68*10^{-4}$)	– 0,92 %	$67{,}8*10^{-4}$ ($1{,}41*10^{-4}$)	3,66 %

Nota : M1 est le modèle correct.
M2 correspond à l'exclusion des termes d'interactions et des termes quadratiques du modèle correct.
M3 correspond à l'élimination de covariables supplémentaires par rapport au modèle M2.

Tableau 4
Estimations par simulation des erreurs-types des estimateurs de θ_1 et θ_2 pour l'imputation par le plus proche voisin (PPV10) et la pondération par le score de propension, sous l'hypothèse de mécanisme MAR et de covariables correctes et spécifiées incorrectement ($H = 1\,000$)

Méthode	Covariables hypothétiques	e.-t.($\hat{\theta}_1$)	e.-t.($\hat{\theta}_2$)	EQM($\hat{\theta}_1$)	EQM($\hat{\theta}_2$)
PPV10	M1 (correct)	$2{,}02*10^{-3}$	$3{,}80*10^{-3}$	$4{,}10*10^{-6}$	$1{,}49*10^{-5}$
	M2	$2{,}06*10^{-3}$	$3{,}88*10^{-3}$	$4{,}29*10^{-6}$	$1{,}54*10^{-5}$
	M3	$2{,}01*10^{-3}$	$3{,}89*10^{-3}$	$4{,}10*10^{-6}$	$1{,}63*10^{-5}$
Pondération par le score de propension	M1 (correct)	$2{,}27*10^{-3}$	$4{,}27*10^{-3}$	$5{,}16*10^{-6}$	$1{,}83*10^{-5}$
	M2	$2{,}17*10^{-3}$	$4{,}42*10^{-3}$	$5{,}51*10^{-6}$	$6{,}90*10^{-5}$
	M3	$2{,}16*10^{-3}$	$4{,}46*10^{-3}$	$4{,}94*10^{-6}$	$6{,}59*10^{-5}$

Tableau 5
Estimations par simulation des biais et des erreurs-types des estimateurs de θ_1 et θ_2 pour l'imputation par le plus proche voisin (PPV10) et la pondération par le score de propension, sous le modèle (non-MAR) commun d'erreur de mesure ($H = 1\,000$)

Méthode	Biais de $\hat{\theta}_1$	Biais rel. de $\hat{\theta}_1$	Biais de $\hat{\theta}_2$	Biais rel. de $\hat{\theta}_2$	e.-t.($\hat{\theta}_1$)	e.-t.($\hat{\theta}_2$)
PPV10	$29{,}0*10^{-4}$ ($0{,}8*10^{-4}$)	5,1 %	$92{,}0*10^{-4}$ ($1{,}48*10^{-4}$)	5,0 %	$2{,}53*10^{-3}$	$4{,}70*10^{-3}$
Pondération par le score de propension	$32{,}3*10^{-4}$ ($0{,}73*10^{-4}$)	5,7 %	$100*10^{-4}$ ($1{,}40*10^{-4}$)	5,7 %	$2{,}31*10^{-3}$	$4{,}42*10^{-3}$

8. Application à l'Enquête sur la population active

À la présente section, nous examinons l'application des méthodes élaborées aux sections 2 à 4 aux données de l'EPA. Cette dernière représente une source importante de données pour l'estimation de la distribution de la rémunération horaire au Royaume-Uni (Stuttard et Jenkins 2001). Il s'agit d'une enquête trimestrielle réalisée auprès de ménages sélectionnés à partir d'un fichier national d'adresses postales avec probabilités égales par échantillonnage systématique stratifié. Tous les adultes compris dans les ménages sélectionnés sont inclus dans l'échantillon. L'échantillon résultant est mis en grappes selon l'appartenance au ménage, mais non selon les caractéristiques géographiques. Chaque ménage sélectionné demeure dans l'échantillon en vue d'être interviewé pendant cinq trimestres consécutifs, puis est éliminé de l'échantillon et remplacé. Les questions concernant la rémunération horaire sont posées durant les première et cinquième interviews seulement, ce qui produit des données à ce sujet pour environ 16 000 employés par trimestre.

Deux mesures de la rémunération horaire sont construites, comme il est décrit à la section 1. La variable dérivée de la rémunération horaire de l'EPA est définie comme suit : a) des questions sont posées aux employés au sujet de leur emploi principal afin de déterminer les gains au cours d'une période de référence, b) des questions sont posées afin de déterminer le nombre d'heures travaillées au cours de la période de référence et c) le résultat de a) est divisé par le résultat de b). La variable directe est obtenue en commençant par demander si le répondant est payé à un taux horaire fixe, puis, si la réponse est affirmative, en demandant quel est ce taux (de base). Skinner et coll. (2002) discutent des nombreuses sources d'erreur de mesure qui affectent la variable dérivée, comme dans le cas d'enquêtes semblables réalisées par d'autres pays (Rodgers, Brown et Duncan 1993; Moore, Stinson et Welniak 2000). Ils concluent que la variable directe mesure la rémunération horaire de façon nettement plus précise. Dans la présente application, l'une des hypothèses de travail est que la variable directe mesure la rémunération horaire sans erreur. Toutefois, le problème que pose cette variable directe est dû au fait que sa valeur manque pour les répondants qui ne sont pas rémunérés à un taux horaire fixe (et pour les non-répondants à la question) et que cette absence de données est associée positivement à la rémunération horaire. La proportion de répondants de l'EPA ayant un emploi (principal) qui fournissent une réponse à la question directe est d'environ 43 %. Cette proportion a tendance à être plus élevée pour les employés faiblement rémunérés; par exemple, le taux est de 72 % pour ceux appartenant au décile inférieur de la variable dérivée. L'information sur la variable directe n'est pas recueillie pour le deuxième emploi (ni les emplois suivants) et nous limitons par conséquent notre examen aux emplois principaux uniquement. Le but est d'utiliser les méthodes de traitement des données manquantes élaborées dans le présent article pour corriger l'erreur de mesure qui entache l'estimation de la distribution de la rémunération horaire. Skinner et coll. (2002) discutent de la vraisemblance de deux hypothèses relatives aux données manquantes énoncées à la section 3 pour cette application.

Les méthodes décrites aux sections 2 à 4 ont été élaborées sous l'hypothèse d'un modèle IID et d'un échantillonnage ignorable. Dans le cas de l'EPA, les employés sont sélectionnés avec probabilités égales, de sorte que l'échantillonnage peut être considéré comme ignorable en ce qui concerne le biais de l'estimation ponctuelle, mais la non-réponse totale est vraisemblablement différentielle, de sorte que les poids de sondage sont calculés de façon à en tenir compte (ONS 1999). Nous proposons d'intégrer ces poids de sondage dans l'estimateur (3) ou, de façon équivalente, de multiplier les coefficients de pondération w_i dans (9) par les poids de sondage. Cette approche est analogue à la façon dont les estimateurs sont pondérés en se fondant sur une hypothèse IID dans l'approche de la pseudovraisemblance (Skinner 1989). Le but est d'utiliser les méthodes décrites aux sections 2 à 4 pour corriger le biais dû à l'erreur de mesure et à la non-réponse partielle, ainsi que les poids de sondage pour corriger le biais dû à l'échantillonnage et à la non-réponse totale. Nous n'avons pas essayé de tenir compte des pondérations dans les méthodes d'imputation et cette question pourrait être étudiée dans le cadre de futurs travaux.

Nous appliquons maintenant l'imputation par le plus proche voisin, l'imputation hot deck dans les classes et la pondération par le score de propension aux données de l'EPA. Nous appliquons la pondération par les poids de sondage à toutes les méthodes. À la figure 1, nous comparons une distribution estimée, qui ne tient pas compte de l'erreur de mesure (courbe en trait plein) aux estimations obtenues par les trois méthodes de traitement des données manquantes (les trois courbes en trait interrompu). Nous soutenons que l'absence approximative de biais est plus importante pour ces dernières estimations que pour la première. Les trois ajustements pour les données manquantes produisent, comme prévu, une forte « cassure » dans la distribution au niveau de la rémunération horaire minimale nationale contrairement à la variable dérivée. Les estimations correspondantes de deux proportions de faible rémunération horaire d'intérêt sont présentées au tableau 6. Les « ajustements pour les données manquantes » ont un effet appréciable comparativement aux estimations fondées sur la variable dérivée. Les résultats laissent entendre que la proportion d'emplois rémunérés au taux horaire minimal national ou à un taux inférieur pourrait être surestimée d'un facteur quatre ou cinq si l'on ne tient pas compte de l'erreur de mesure. Les écarts entre les méthodes de traitement des données manquantes sont nettement plus faibles. Nous voyons que les estimations sous la pondération par le score de propension diffèrent des estimations calculées par les méthodes d'imputation, du moins pour le trimestre de juin à août 1999. Il convient de souligner que le taux de réponse durant ce trimestre de l'EPA a été plus faible que pour les trimestres suivants à cause de modifications apportées au questionnaire de l'enquête. Nous avons constaté que pour les trimestres suivants, pour lesquels le taux de réponse était de l'ordre de 43 %, la pondération et l'imputation ont donné des estimations fort semblables des proportions de travailleurs faiblement rémunérés, comme l'illustre le tableau 7 pour le trimestre de mars à mai 2000. La diminution de la proportion d'employés faiblement rémunérés au cours du temps est due à l'effet de la loi sur le salaire minimum national. En outre, nous utilisons différents modèles d'imputation et de pondération par le score de propension en vue d'analyser les effets de diverses spécifications du modèle sur les estimations de la proportion d'employés faiblement rémunérés. Le tableau 6 montre que l'utilisation de différents modèles pourrait avoir une incidence sur les estimations. À mesure que le modèle devient plus complexe, nous observons une réduction des estimations dans le cas des deux estimateurs ponctuels. Cela pourrait refléter un écart par rapport à l'hypothèse de mécanisme MAR pour les modèles d'imputation plus simples. Du moins pour le trimestre de 1999, les différences d'estimations semblent être plus importantes entre les méthodes de pondération et d'imputation qu'entre les modèles. Soulignons que les estimations présentées ici pourraient différer légèrement des estimations officielles pour le Royaume-Uni, puisque, par exemple, les estimations officielles sont fondées sur des modèles d'imputation différents, qui traitent différemment les valeurs extrêmes ou qui imputent différemment les valeurs pour certaines professions.

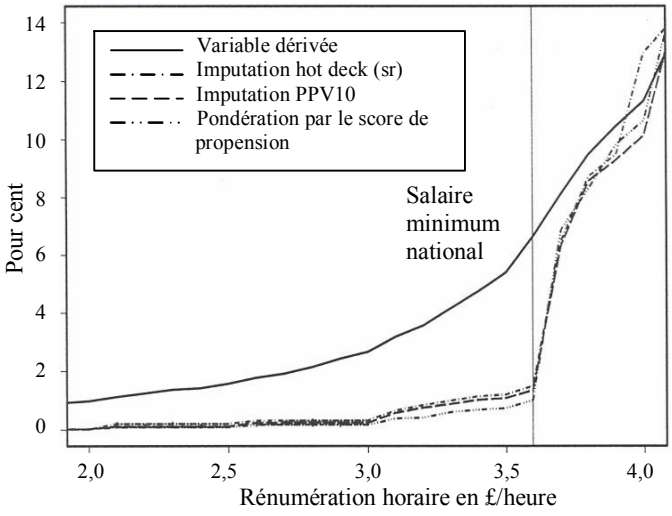


Figure 1. Diverses estimations de la distribution de la rémunération horaire de 2 £ à 4 £ pour le groupe des 22 ans et plus, juin à août 1999.

Tableau 6
Estimations de θ_1 et θ_2 (pondérées) pour le groupe des 18 ans et plus par application de divers modèles de pondération par le score de propension et d'imputation aux données de l'EPA, juin à août 1999

Méthode	Modèle de pondération par le score de propension ou modèle d'imputation	(Pondérée) $\hat{\theta}_1$ (%)	(Pondérée) $\hat{\theta}_2$ (%)
Variable dérivée	–	7,13	20,5
Pondération par le score de propension	M1	0,96	34,5
	M2	1,08	38,4
	M3	1,08	38,4
IHDSR10	M1	1,44	32,1
	M2	1,41	32,9
	M3	1,50	33,2
PPV10	M1	1,32	32,6
	M2	1,44	32,8
	M3	1,50	33,0

Nota : M1 est le modèle le plus complexe comprenant des termes quadratiques et des termes d'interactions.
M2 exclut les termes d'interactions et les termes quadratiques inclus dans le modèle M1.
M3 correspond à l'élimination de covariables supplémentaires par rapport au modèle M2.

Tableau 7
Estimations de θ_1 et θ_2 (pondérées) pour le groupe des 18 ans et plus par application de la pondération par le score de propension et de l'imputation aux données de l'EPA, mars à mai 2000

Méthode	Modèle de pondération par le score de propension ou modèle d'imputation	(Pondérée) $\hat{\theta}_1$ (%)	(Pondérée) $\hat{\theta}_2$ (%)
Pondération par le score de propension	M1	0,54	27,10
IHDSR10	M1	0,57	26,01
PPV10	M1	0,55	26,61

9. Conclusion

Dans le présent article, nous avons examiné l'application de diverses méthodes de traitement des données manquantes en vue de corriger le biais causé par l'erreur de mesure dans l'estimation d'une fonction de distribution. Parmi les méthodes d'imputation, celles par le plus proche voisin donnent les résultats les plus prometteurs en ce qui concerne le biais. Il n'existe aucun signe que ces méthodes déterministes produisent un biais plus important que les méthodes d'imputation stochastiques. L'imputation fractionnaire donne lieu à des gains d'efficacité appréciables comparativement à l'imputation simple et semble être plus efficace que la pénalisation de la fonction de distance ou que l'échantillonnage sans remise avec imputation simple. Comparativement à la méthode de pondération par le score de propension, l'imputation fractionnaire par le plus proche voisin donne des résultats comparables, mais présente de légers avantages en ce qui a trait à la robustesse et à l'efficacité. L'étude en simulation laisse entendre que l'effet sur le biais sous un modèle mal spécifié est plus important dans le cas de la pondération par le score de propension et que les erreurs-types sous l'approche de pondération sont supérieures de 5 % à 15 % à celles observées pour la méthode d'imputation.

Nous avons entrepris d'autres travaux en vue d'élaborer et d'évaluer des méthodes d'estimation de la variance associées, ainsi que d'autres méthodes d'estimation ponctuelle fondées sur le modèle commun d'erreur de mesure décrit à la section 2.

Remerciements

Nous remercions Danny Pfeffermann pour ses commentaires concernant une version antérieure du présent article.

Bibliographie

Brick, J.M., et Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.

Buonaccorsi, J.P. (1990). Double sampling for exact values in some multivariate measurement error problems. *Journal of the American Statistical Association*, 85, 1075-1082.

Chen, J., et Shao, J. (2000). Nearest neighbour imputation for survey data. *Journal of Official Statistics*, 16, 113-131.

Chen, J., et Shao, J. (2001). Jackknife variance estimation for nearest neighbour imputation. *Journal of the American Statistical Association*, 96, 453, 260-269.

Chesher, A. (1991). The effect of measurement error. *Biometrika*, 78, 451-462.

David, M.H., Little, R., Samuhel, M. et Triest, R. (1983). Imputation models based on the propensity to respond. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 168-173.

Dickens, R., et Manning, A. (2004). Has the national minimum wage reduced UK wage inequality? *Journal of the Royal Statistical Society*, Séries A, 4, 613-626.

Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.

Fay, R.E. (1999). Theory and application of nearest neighbour imputation in census 2000. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 112-121.

Fuller, W.A. (1995). Estimation in the presence of measurement error. *Revue Internationale de Statistique*, 63, 121-141.

Kalton, G. (1983). *Compensating for missing survey data*. Michigan, Institute for Social Research.

Kalton, G., et Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics*, *Part A*, *Theory and Methods*, 13, 1919-1939.

Kim, J.K. (2004). Efficient nonresponse weighting adjustment using estimated response probability. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Kim, J.-K., et Fuller, W.A. (2002). Variance estimation for nearest neighbour imputation. Manuscript non-publié.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54, 139-157.

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287-301.

Little, R.J.A., et Rubin, D.B. (2002). *Statistical analysis with missing data*. New York: John Wiley & Sons, Inc.

Luo, M., Stokes, L. et Sager, T. (1998). Estimation of the CDF of a finite population in the presence of a calibration sample. *Environmental and Ecological Statistics*, 5, 277-289.

Moore, J.C., Stinson, L.L. et Welniak, E.J. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics*, 16, 331-361.

ONS (1999). *Labour Force Survey*. User Guide, Volume 1, Background and Methodology, London.

Rancourt, E. (1999). Estimation with nearest neighbour imputation at Statistics Canada. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 131-138.

Rodgers, W.L., Brown, C. et Duncan, G.J. (1993). Errors in survey reports of earnings, hours worked and hourly wages. *Journal of the American Statistical Association*, 88, 1208-1218.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

Rubin, D.B., et Schenker N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable non-response. *Journal of the American Statistical Association*, 81, 366-374.

Skinner, C.J. (1989). Domain means, regression and multivariate analysis. Dans *Analysis of Complex Surveys*, (Éds. C.J. Skinner, D. Holt et T.M.F. Smith), Chichester, Wiley.

Skinner, C., Stuttard, N., Beissel-Durrant, G. et Jenkins, J. (2002). The measurement of low pay in the UK Labour Force Survey. *Oxford Bulletin of Economics and Statistics*, 64, 653-676.

Stuttard, N., et Jenkins, J. (2001). Measuring low pay using the new earnings survey and the Labour Force Survey. *Labour Market Trends*, janvier 2001, 55-66.

Tenenbein, A. (1970). A double sampling scheme for estimating from binary data with misclassifications. *Journal of the American Statistical Association*, 65, 1350-1361.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, juin 2006
Vol. 32, N° 1, pp. 41-57
Statistique Canada, N° 12-001-XIF au catalogue

De l'estimation des quantiles par calage

Torsten Harms et Pierre Duchesne [1]

Résumé

Le présent article traite de l'application du paradigme de calage à l'estimation des quantiles. La méthodologie proposée suit une approche semblable à celle qui donne lieu aux estimateurs par calage originaux de Deville et Särndal (1992). Une propriété intéressante de cette nouvelle méthodologie est qu'elle ne nécessite pas la connaissance des valeurs des variables auxiliaires pour toutes les unités de la population. Il suffit de connaître les quantiles correspondants de ces variables auxiliaires. L'adoption d'une métrique quadratique permet d'obtenir une représentation analytique des poids de calage, qui sont alors similaires à ceux menant à l'estimateur par la régression généralisée (GREG). Nous discutons de l'estimation de la variance et de la construction des intervalles de confiance. Au moyen d'une petite étude par simulation, nous comparons l'estimateur par calage à d'autres estimateurs fréquemment utilisés des quantiles qui s'appuient également sur des données auxiliaires.

Mots clés : Estimateurs par calage; estimateurs par la différence; estimateurs par le ratio; quantiles.

1. Introduction

Ces dernières années, dans le contexte des sondages, beaucoup d'attention a été accordée à l'estimation des fonctions de répartition des populations. À cet égard, la médiane, souvent considérée comme une mesure d'emplacement plus satisfaisante que la moyenne, particulièrement si la variable d'intérêt suit une loi asymétrique, a suscité un intérêt particulier. Habituellement, les estimateurs traditionnels des moyennes et des totaux de population peuvent être améliorés sensiblement si l'on dispose d'information auxiliaire pertinente. Par conséquent, il paraît fort souhaitable d'utiliser ce genre d'information dans les estimateurs des quantiles d'échantillon.

Adoptant une approche basée sur un modèle, Chambers et Dunstan (1986) ont examiné des estimateurs des quantiles basés sur un estimateur de la fonction de répartition auquel sont intégrés des renseignements auxiliaires. Rao, Kovar et Mantel (1990) ont proposé des variantes fondées sur le plan de sondage de l'approche basée sur un modèle. Ils ont comparé, dans des expériences par simulation, deux estimateurs des quantiles, basés sur des estimateurs par le ratio et par différence, à l'estimateur fondé sur le plan de sondage simple dans lequel n'est utilisée aucune information auxiliaire. Il convient de souligner que ni l'un ni l'autre des estimateurs proposés fondés sur le plan de sondage ne nécessite la connaissance de l'information auxiliaire pour chaque unité de la population; il suffit de connaître les quantiles correspondants. Bien que l'estimateur basé sur un modèle proposé par Chambers et Dunstan (1986) puisse être plus efficace que l'option fondée sur le plan de sondage si le modèle est spécifié correctement, Rao et coll. (1990) soulignent l'avantage des estimateurs fondés sur le plan de sondage en cas de spécification incorrecte du modèle. Chambers, Dorfman et Hall (1992) ont comparé théoriquement la convergence, le biais asymptotique et la variance de ces deux estimateurs sous un modèle de population. Leur conclusion principale est qu'aucune des deux méthodes n'est sensiblement meilleure que l'autre. Dorfman (1993) a réévalué les résultats des simulations de Rao et coll. (1990) et proposé une version modifiée de leur méthode reposant sur des arguments fondés sur un modèle. Les estimateurs de la variance dans le contexte de l'approche basée sur un modèle de Chambers et Dunstan (1986) et des estimateurs fondés sur le plan de sondage de Rao et coll. (1990) sont examinés dans Wu et Sitter (2001).

Parmi les autres travaux relatifs aux estimateurs des quantiles et de la médiane, mentionnons ceux de Kuk (1988) qui propose des estimateurs des quantiles sous échantillonnage PPT (*probabilité proportionnelle à la taille*) et ceux de Kuk et Mak (1989) qui utilisent une méthode basée sur la classification croisée des individus compris dans l'échantillon en fonction de la variable d'intérêt et d'une variable auxiliaire unique. Meeden (1995) adopte une approche différente pour construire un estimateur de la médiane basé sur des données auxiliaires univariées, en utilisant le concept bayésien de l'échantillonnage de Pólya pour imputer toutes les valeurs de population inconnues de la variable cible selon une approche fondée sur le ratio. Récemment, Rueda, Arcos et Martínez (2003) ont construit des estimateurs des quantiles qui étendent les estimateurs par le ratio, par la différence et par la régression de façon similaire à ceux élaborés pour la moyenne de population.

Dans le présent article, nous appliquons le concept de calage que Deville (1988) a été le premier à proposer afin de dériver un estimateur des quantiles. L'approche par calage

1. Torsten Harms et Pierre Duchesne, Université de Montréal, Département de mathématiques et de statistique, CP 6128 Succursale Centre-Ville, Montréal, (Québec), H3C 3J7, Canada. Courriel : duchesne@dms.umontreal.ca.

est devenue populaire dans les applications pratiques, parce que les estimateurs résultants sont faciles à interpréter et à justifier, étant donné qu'ils s'appuient sur les poids d'échantillonnage et des contraintes de calage naturelles. Cette approche a été élaborée dans le cadre des travaux fondamentaux de Deville et Särndal (1992) à titre de nouveau moyen d'intégrer l'information auxiliaire dans l'estimation des totaux de population. Les poids dits calés sont obtenus en minimisant une mesure de distance entre les poids d'échantillonnage et les nouveaux poids sous certaines contraintes de calage. Pour l'estimation des totaux, les poids calés remplacent les poids de sondage originaux utilisés dans les estimateurs de type Horvitz-Thompson. Lorsqu'on les applique aux variables auxiliaires disponibles dans l'échantillon, les nouveaux poids reproduisent exactement les totaux connus de population de ces variables, d'où le nom d'estimateurs par calage donné aux estimateurs de cette classe. Voir aussi Singh et Mohl (1996) qui fournissent des justifications simples des estimateurs par calage. Ils présentent en outre un traitement très général et harmonisé des méthodes par calage produisant des poids qui satisfont à certaines restrictions concernant les fourchettes de valeurs et certaines contraintes d'étalonnage.

Essentiellement, notre but est de proposer pour les quantiles des estimateurs par calage aussi faciles à appliquer et à interpréter que les estimateurs par calage des totaux mis au point par Deville et Särndal (1992). Comparativement aux estimateurs des quantiles décrits dans la littérature, les nouveaux estimateurs par calage devraient aussi donner des résultats avantageux en ce qui concerne le biais, la variance et les taux de couverture des intervalles de confiance. Les premiers estimateurs par calage proposés pour les fonctions de répartition et les quantiles comprennent ceux de Kovačević (1997), qui a étudié des estimateurs de la fonction de répartition calés sur les moments des variables auxiliaires. Harms (2003) a suivi une approche analogue comportant des applications à la version finlandaise du Panel européen des ménages. Ren (2002) semble avoir été le premier à élaborer un traitement unifié de l'application des estimateurs par calage aux fonctions de répartition et aux quantiles. Les estimateurs par calage applicables aux quantiles présentés ici constituent une prolongation des travaux de Ren (2002). Nous adhérons aussi étroitement que possible au paradigme de calage original appliqué aux totaux : lorsque le paramètre d'intérêt est un total, il semble logique de faire le calage sur les variables auxiliaires. Ici, puisque le paramètre d'intérêt est un quantile, les contraintes de calage imposent l'utilisation de poids tels que les estimateurs des quantiles d'échantillon des variables auxiliaires et de leurs quantiles de population correspondants soient égaux. Autrement dit, les estimateurs pondérés des quantiles des variables auxiliaires devraient produire exactement les quantiles de population, que l'on suppose connus. Nous présentons des arguments qui justifient le calage sur les quantiles, quand le paramètre d'intérêt est lui-même un quantile. Fait intéressant, notre méthode ne nécessite pas que l'on connaisse les valeurs des variables auxiliaires pour toutes les unités de la population. Puisque les estimateurs résultants présentent une forme structurelle fort semblable à celle des estimateurs par calage originaux des totaux, nous nous attendons à ce que, sous des conditions générales, les estimateurs proposés des quantiles soient asymptotiquement sans biais par rapport au plan de sondage. En outre, ces similarités nous permettent de dériver des estimateurs de la variance qui admettent une forme familière. Contrairement à certains autres estimateurs, l'approche proposée est également applicable aux variables auxiliaires vectorielles (c'est-à-dire les situations où plusieurs variables auxiliaires sont disponibles), tout en ne requérant que des renseignements auxiliaires minimes. Cependant, certaines restrictions pourraient s'appliquer lorsque l'échantillon est fortement non représentatif de la population échantillonnée ou que les quantiles que l'on estime sont très proches du minimum ou du maximum de population. Notons que des échantillons fortement non représentatifs peuvent aussi causer des problèmes dans le cas des estimateurs par calage des totaux utilisés couramment; le cas échéant, l'algorithme pour le calcul de ces estimateurs peut ne pas converger pour de nombreuses mesures de distance présentant un intérêt pratique.

La présentation de l'article est la suivante. À la section 2, nous donnons certains préliminaires, dont un bref examen des estimateurs par calage des totaux. À la section 3.1, nous décrivons l'élaboration des nouveaux estimateurs par calage des quantiles. La fonction de répartition standard peut être interprétée comme étant un estimateur d'Horvitz-Thompson, qui offre une approche possible de la construction d'un estimateur calé de la fonction de répartition. Les estimateurs des quantiles sont alors dérivés naturellement par inversion de l'estimateur de la fonction de répartition (voir, par exemple, Ren (2002)). Comme dans le cas des estimateurs par calage des totaux, les poids de sondage peuvent être remplacés par des poids d'échantillonnage plus généraux, afin de tenir compte de l'information auxiliaire. Toutefois, dans de nombreuses situations présentant un intérêt pratique, il se peut qu'aucune solution n'existe pour les contraintes de calage lorsqu'on adopte ce genre d'estimateur des fonctions de répartition, parce que celui-ci correspond à une fonction échelon. Afin d'éviter les problèmes d'existence de solutions pour les contraintes de calage, nous présentons un nouvel estimateur de la fonction de répartition fondé sur le concept naturel d'interpolation. À la section 3.2, nous présentons, sous les conditions de la métrique quadratique ordinaire, une représentation analytique des poids de calage; à la section 3.3, nous discutons des estimateurs de la

variance et des intervalles de confiance. L'un des aspects pratiques consiste à évaluer la méthodologie proposée en l'appliquant à des populations réelles et à plusieurs plans d'échantillonnage. Par conséquent, à la section 4, nous présentons une petite étude par simulation, où nous comparons notre approche, en ce qui concerne la variance, le biais et le taux de couverture des intervalles de confiance, à celle de Chambers et Dunstan (1986), ainsi qu'à certains estimateurs proposés par Rao et coll. (1990). Enfin, à la section 5, nous présentons nos conclusions.

2. Certains préliminaires sur les estimateurs par calage

À la présente section, nous présentons les concepts fondamentaux et les notations qui seront utiles dans la suite. En outre, nous passons brièvement en revue les estimateurs par calage des totaux.

Soit $U = \{1, ..., k, ..., N\}$ une population finie de taille N. Soit $T_y = \sum_U y_k$ le total de population de la variable d'intérêt y (notons que pour un ensemble A, $A \subseteq U$, nous utiliserons $\sum_A$ comme abréviation de $\sum_{k \in A}$). Nous tirons un échantillon $s \subset U$ de taille n conformément à un plan d'échantillonnage. Soit $\pi_k = \Pr(s \ni k)$ et $\pi_{kl} = \Pr(s \ni k, l)$ les probabilités d'inclusion de premier et de deuxième ordre, respectivement. Nous dénotons les poids de sondage par $d_k = \pi_k^{-1}$ et $\hat{T}_{y,\mathrm{HT}} = \sum_s d_k y_k$ représente l'estimateur d'Horvitz-Thompson (HT) de T_y.

Soit $\mathbf{x}_k = (x_{1k}, ..., x_{Jk})'$ un vecteur de variables auxiliaires associé à l'unité k, $k \in U$. Les estimateurs par calage incluent naturellement l'information auxiliaire dans l'estimation. Soit $s = \{k_1, ..., k_n\}$, $s \subset U$. En partant du vecteur de poids originaux $\mathbf{d} = (d_{k_1}, ..., d_{k_n})'$, nous trouvons de nouveaux poids qui, lorsqu'ils sont appliqués aux variables auxiliaires disponibles dans s, permettent d'extraire les totaux de population connus pour les J variables auxiliaires $\mathbf{T}_\mathbf{x} = \sum_U \mathbf{x}_k = (T_{x_1}, ..., T_{x_J})'$. Les estimateurs par calage des totaux seront définis de façon plus précise dans la définition 1.

Définition 1 (estimateur par calage des totaux). *Soit* $\mathbf{d} = (d_{k_1}, ..., d_{k_n})'$ *les poids de sondage. L'estimateur par calage des totaux prend la forme* $\hat{T}_{y,\mathrm{cal}} = \sum_s w_{ks} y_k$, *où les poids* w_{ks}, $k \in s$ *sont obtenus en résolvant le problème de minimisation qui suit par rapport à la variable* $\mathbf{v} = (v_{k_1}, ..., v_{k_n})'$:

$$\mathbf{w} = \mathrm{argmin}_\mathbf{v}\, \mathrm{D}(\mathbf{v}, \mathbf{d}), \qquad (1)$$

sous les contraintes de calage $\sum_s v_k \mathbf{x}_k = \mathbf{T}_\mathbf{x}$, *où* $\mathrm{D}(\cdot,\cdot)$ *représente la mesure de distance et* $\mathbf{w} = (w_{k_1}, ..., w_{k_n})'$ *correspond au vecteur des poids calés.*

Pour simplifier la notation, nous écrivons $w_k \equiv w_{ks}$ dans la définition 1 quand aucune confusion n'est possible. Il est courant, en pratique, de poser $x_{1k} \equiv 1$, $\forall k \in U$, et par conséquent $T_{x_1} = N$. Cela signifie que les poids calés satisfont la contrainte naturelle $\sum_s w_k = N$. De nombreuses fonctions de distance D sont proposées dans la littérature (voir, par exemple, Deville et Särndal (1992); Chen et Qin (1993); Thompson (1997)). Considérons la fonction de distance quadratique.

$$\mathrm{D}(\mathbf{v}, \mathbf{d}) = \sum_s \frac{(v_k - d_k)^2}{d_k q_k}, \qquad (2)$$

où q_k détermine l'importance de l'unité $k \in s$ dans le problème de calage. Les problèmes d'hétéroscédasticité peuvent être réglés en choisissant les valeurs de q_k de façon appropriée. En résolvant le problème d'optimisation (1) par la technique du multiplicateur de Lagrange (voir Deville et Särndal 1992, entre autres), nous obtenons les poids $w_k = d_k(1 + q_k \mathbf{x}_k' \boldsymbol{\lambda}_s)$, où $\boldsymbol{\lambda}_s = (\sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k')^{-1}(\mathbf{T}_\mathbf{x} - \hat{\mathbf{T}}_{\mathbf{x},\mathrm{HT}})$ et $\hat{\mathbf{T}}_{\mathbf{x},\mathrm{HT}}$ représente l'estimateur HT de $\mathbf{T}_\mathbf{x}$. Ce choix de la fonction de distance aboutit aux poids de l'estimateur par la régression généralisée (GREG) bien connu de Cassel, Särndal et Wretman (1976), qui est étudiée en détail dans Särndal, Swensson et Wretman (1992). Sous des exigences minimales concernant la mesure de distance D, Deville et Särndal (1992) ont montré que tous les estimateurs par calage de cette classe sont asymptotiquement équivalents à l'estimateur GREG. Pour faciliter l'interprétation et pour d'autres raisons esthétiques, certains utilisateurs pourraient souhaiter obtenir des poids positifs ou les contraindre à un intervalle particulier (voir aussi Singh et Mohl 1996). Dans les applications pratiques, ces caractéristiques numériques des poids semblent être le motif principal de divers choix de D.

3. Nouveaux estimateurs par calage

À la présente section, nous élaborons des estimateurs par calage pour les quantiles, selon des idées similaires à celles donnant lieu aux estimateurs par calage des totaux de population décrits à la section 2. Nous présentons les nouveaux estimateurs par calage pour les quantiles à la sous-section suivante, en utilisant des estimateurs interpolés de la fonction de répartition. Puis, nous accordons une attention spéciale à la fonction de distance quadratique. À la dernière sous-section, nous présentons l'estimation de la variance et la construction des intervalles de confiance.

3.1 Définition des estimateurs par calage des quantiles

Soit $\mathbf{Q}_{\mathbf{x},\alpha} = (Q_{x_1,\alpha}, ..., Q_{x_J,\alpha})'$ le vecteur connu des quantiles de population pour le vecteur de variables auxiliaires $\mathbf{x}_k = (x_{1k}, ..., x_{Jk})'$, $k \in U$. La fonction de Heaviside $H(z)$ est donnée par :

$$H(z) = \begin{cases} 1, & z \geq 0, \\ 0, & z < 0. \end{cases}$$

Nous définissons la fonction de répartition d'une variable auxiliaire scalaire x dans la population de la manière habituelle par $F_x(t) = N^{-1} \sum_U H(t - x_k)$, et nous obtenons le quantile de population $Q_{x,\alpha}$ en posant $Q_{x,\alpha} = \inf\{t \mid F_x(t) \geq \alpha\}$.

Le vecteur $\mathbf{Q}_{\mathbf{x},\alpha}$ contient les quantiles des variables auxiliaires, obtenus d'après l'information tirée d'enquêtes antérieures ou de sources administratives disponibles. Par exemple, pour les lois asymétriques qui sont assez fréquentes dans le cas des enquêtes auprès des entreprises et des enquêtes économiques, il paraît plus logique de garder dans les fichiers d'enregistrements les médianes plutôt que les moyennes de population; le cas échéant, il semble naturel de supposer que $\mathbf{Q}_{\mathbf{x},0,5}$ est connu. Cela donne à penser qu'en suivant la même approche que celle menant au calage des totaux décrits à la section 2, l'estimateur proposé pour les quantiles de population $Q_{y,\alpha}$ de la variable d'intérêt y, noté $\hat{Q}_{y,\text{cal},\alpha}$, pourrait être obtenu par inversion d'un certain estimateur de la fonction de répartition (dont nous discutons plus loin), sous des contraintes de calage telles que $\hat{Q}_{x_j,\text{cal},\alpha} = Q_{x_j,\alpha}$, $j = 1, ..., J$. Suivant l'interprétation habituelle, si les poids calés nous permettent d'extraire les quantiles de population connus des variables auxiliaires, alors, sous certaines conditions, ils devraient produire des estimateurs raisonnables des quantiles de la variable d'intérêt y.

Plus précisément, nous obtenons les poids calés en résolvant le problème d'optimisation suivant :

$$\mathbf{w} = \text{argmin}_{\mathbf{v}}\, \mathrm{D}(\mathbf{v}, \mathbf{d}), \tag{3}$$

sous les contraintes de calage $\sum_s v_k = N$ et $\hat{\mathbf{Q}}_{\mathbf{x},\text{cal},\alpha} = (\hat{Q}_{x_1,\text{cal},\alpha}, ..., \hat{Q}_{x_J,\text{cal},\alpha})' = \mathbf{Q}_{\mathbf{x},\alpha}$.

Les estimateurs $\hat{\mathbf{Q}}_{\mathbf{x},\text{cal},\alpha}$ et $\hat{Q}_{y,\text{cal},\alpha}$ s'appuient sur le vecteur de poids $\mathbf{w}$, issus de la résolution du problème de calage (3). Pour calculer ces estimateurs pour les quantiles, nous devons construire les estimateurs pondérés par les poids w de la fonction de répartition des variables $\mathbf{x}$ et y. En se basant sur les poids d'échantillonnage $\mathbf{d}$, un estimateur logique de la fonction de répartition d'échantillonnage est donné par

$$\tilde{F}_y(t) = \sum_s d_k H(t - y_k) \Big/ \sum_s d_k, \tag{4}$$

qui fournit un estimateur convergent de $F_y(t)$. De même, $F_{x_j}(t)$ peut être estimée de façon convergente par $\tilde{F}_{x_j}(t) = \sum_s d_k H(t - x_{jk}) / \sum_s d_k$, $j = 1, ..., J$. Un estimateur de la fonction de répartition pondéré par les poids w de $F_{x_j}(t)$ est donné par

$$\tilde{F}_{x_j,\text{cal}}(t) = \sum_s w_k H(t - x_{jk}) \Big/ \sum_s w_k. \tag{5}$$

Une formule similaire est vérifiée pour $\tilde{F}_{y,\text{cal}}(t)$. Ces estimateurs pondérés par les poids w sont considérés dans Ren (2002). Cependant, si l'on estime $Q_{x_j,\alpha}$ par $\hat{Q}_{x_j,\alpha} = \inf\{t \mid \tilde{F}_{x_j}(t) \geq \alpha\}$, ou que l'on fait une estimation similaire en utilisant une version pondérée par les poids w, alors il est généralement impossible d'atteindre une solution exacte du problème de calage (3). En effet, si l'on utilise la définition qui précède pour estimer les quantiles par inversion de la fonction de distribution en utilisant les définitions antérieures, les contraintes du problème d'optimisation (3) ne seront généralement pas remplies, à moins que l'échantillon s contienne précisément une unité k telle que $x_{jk} = Q_{x_j,\alpha}$. Quand J est grand, ce problème peut être plus prononcé. En outre, même si l'échantillon contient une telle valeur, il est parfois impossible d'obtenir les poids nécessaires pour minimiser la fonction de distance, parce que, dans certaines circonstances, les poids qui satisfont les contraintes de calage forment un ensemble ouvert, tandis que les poids optimaux se situent précisément sur la limite de cet ensemble. L'exemple qui suit illustre cette situation.

Exemple 1

Considérons une population U de taille $N = 30$, telle que la médiane de population de x soit $Q_{x,0,5} = 2$. Nous tirons un échantillon s de taille $n = 3$ et supposons que $x_k = k, \forall k \in s = \{1, 2, 3\}$. Pour simplifier, nous adoptons la mesure de distance $\mathrm{D}(\mathbf{v}, \mathbf{d}) = \sum_s (v_k - d_k)^2$; nous supposons que les poids d'échantillonnage sont $(d_1, d_2, d_3) = (15, 9, 6)$. En nous basant sur (5), la contrainte de calage est $\tilde{Q}_{x,\text{cal},0,5} = \inf\{t \mid \tilde{F}_{x,\text{cal}}(t) \geq 0,5\} = 2$, qui implique $\sum_s w_k H(2 - x_k) \geq 15$ et $\sum_s w_k H(1 - x_k) < 15$. De façon équivalente, $w_1 + w_2 \geq 15$ et $w_1 < 15$. Donc, nous devons choisir w_1 de la forme $w_1 = 15 - \epsilon$, pour $\epsilon > 0$. Dans ce cas, puisque $w_1 + w_2 + w_3 = 30$, nous avons que $\mathrm{D}(\mathbf{v}, \mathbf{d}) = \epsilon^2 + (w_2 - 9)^2 + (w_2 - 9 - \epsilon)^2$, ce qui nous mène à la solution optimale $(w_1, w_2, w_3) = (15 - \epsilon, 9 + \epsilon/2, 6 + \epsilon/2)$. Par conséquent, pour ces poids, $\mathrm{D}(\mathbf{v}, \mathbf{d}) = 3\epsilon^2/2$, qui est de toute évidence minimisé quand $\epsilon \to 0$. Toutefois, la limite se réduit à $\mathbf{w} = (w_1, w_2, w_3) = (15, 9, 6)$ avec $\mathrm{D}(\mathbf{w}, \mathbf{d}) = 0$, mais, d'après ces poids, $\tilde{Q}_{x,\text{cal},0,5} = 1 \neq Q_{x,0,5} = 2$.

Néanmoins, il est possible d'éviter ces difficultés en envisageant un estimateur lisse de la fonction de répartition. Pour simplifier, considérons ici un estimateur de la fonction de répartition calculé par interpolation linéaire (nous discutons d'une autre possibilité à la section 5), qui est défini précisément dans la définition 2.

Définition 2 (Estimateur interpolé de la fonction de répartition). *Définissons*

$$\hat{F}_{y,\text{cal}}(t) = \frac{\sum_s w_k H_{y,s}(t, y_k)}{\sum_s w_k}, \tag{6}$$

$$\hat{F}_{x_j,\text{cal}}(t) = \frac{\sum_s w_k H_{x_j,s}(t, x_{jk})}{\sum_s w_k}, \qquad (7)$$

où, dans (4) *et* (5)*, la fonction de Heaviside* H *est remplacée par la fonction légèrement modifiée*

$$H_{y,s}(t, y_k) = \begin{cases} 1, & y_k \le L_{y,s}(t), \\ \beta_{y,s}(t) & y_k = U_{y,s}(t), \\ 0, & y_k > U_{y,s}(t), \end{cases} \qquad (8)$$

où $L_{y,s}(t) = \max\{\{y_k,\ k \in s \mid y_k \le t\} \cup \{-\infty\}\}$, $U_{y,s}(t) = \min\{\{y_k, k \in s \mid y_k > t\} \cup \{\infty\}\}$ *et* $\beta_{y,s}(t) = \{t - L_{y,s}(t)\}/\{U_{y,s}(t) - L_{y,s}(t)\}$. *La fonction* $H_{x_j,s}(t, x_k)$ *est définie de la même façon. Les estimateurs* (6) *et* (7)*, basés sur les fonctions* $H_{y,s}(t, y_k)$ *et* $H_{x_j,s}(t, x_k)$*, sont appelés estimateurs interpolés pour les fonctions de distribution* $F_y(t)$ *et* $F_{x_j}(t)$*, respectivement.*

Dans (8), les diverses grandeurs sont faciles à interpréter : $L_{y,s}$ et $U_{y,s}$ représentent les voisins inférieur et supérieur de t dans les valeurs échantillonnées $y_k, k \in s$, et $\beta_{y,s}(t)$ représente le coefficient d'interpolation linéaire entre ces deux grandeurs. En particulier, pour tout $t \in \{y_k, k \in s\}$, nous avons $H_{y,s}(t, y_k) = H(t - y_k)$. Par conséquent, les relations $\hat{F}_{y,\text{cal}}(t) = \tilde{F}_{y,\text{cal}}(t)$ sont satisfaites pour tout $t \in \{y_k, k \in s\}$. Pour toutes les autres valeurs de t, $\hat{F}_{y,\text{cal}}(t)$ consiste en une interpolation linéaire entre ces grandeurs. Dans l'exemple qui suit, nous réexaminons l'exemple 1 en utilisant l'estimateur interpolé de la fonction de répartition (7).

Exemple 2

Dans l'exemple 1, si nous utilisons la version interpolée (7), les contraintes deviennent $w_1 + w_2 + w_3 = 30$ et $(w_1 + w_2)/(w_1 + w_2 + w_3) = 0{,}5$. Par conséquent, $w_3 = 15$, $w_1 + w_2 = 15$. Des opérations algébriques simples montrent que la solution optimale est $(w_1, w_2, w_3) = (10{,}5, 4{,}5, 15)$, qui est maintenant bien définie.

Si nous utilisons les estimateurs interpolés de la fonction de répartition, $\hat{F}^{-1}_{y,\text{cal}}(\alpha)$ et $\hat{F}^{-1}_{x_j,\text{cal}}(\alpha)$ sont maintenant des estimateurs bien définis du quantile α pour tout $\alpha \in (0, 1)$, à condition de pouvoir s'assurer que les poids w_k sont tous strictement positifs. En posant $\hat{Q}_{x_j,\text{cal},\alpha} = \hat{F}^{-1}_{x_j,\text{cal}}(\alpha)$, nous définissons l'estimateur par calage proposé $\hat{Q}_{y,\text{cal},\alpha}$ pour le quantile $Q_{y,\alpha}$, en utilisant l'estimateur interpolé de la fonction de répartition donné dans la définition 2.

Définition 3 (Estimateur par calage des quantiles). *Considérons le problème d'optimisation* (3)*, sous les contraintes de calage* $\sum_s v_k = N$ *et* $\hat{\mathbf{Q}}_{\mathbf{x},\text{cal},\alpha} = (\hat{Q}_{x_1,\text{cal},\alpha}, \ldots, \hat{Q}_{x_J,\text{cal},\alpha})' = \mathbf{Q}_{\mathbf{x},\alpha}$. *Si nous résolvons ce problème d'optimisation et dénotons les poids résultant par* $\mathbf{w}$*, l'estimateur par calage proposé des quantiles* $Q_{y,\alpha}$ *est défini par*

$$\hat{Q}_{y,\text{cal},\alpha} = \hat{F}^{-1}_{y,\text{cal}}(\alpha), \qquad (9)$$

où $\hat{F}_{y,\text{cal}}(t)$ *est donné par* (6).

L'une des propriétés séduisantes de l'estimateur proposé (9) est qu'il donne des quantiles de population exacts quand la relation entre y et une variable auxiliaire scalaire x est exactement linéaire. Émettons l'hypothèse que $y_k = a + bx_k$ tient parfaitement pour toutes les unités $k \in U$ et supposons que les unités de l'échantillon s sont telles que $x_k < Q_{x,\alpha} < x_l$ pour certaines unités x_k et x_l, $k, l \in s$. Pour l'estimateur calé (9), nous avons que $\hat{F}_{x,\text{cal}}(Q_{x,\alpha}) = \alpha$. Nous devons faire la distinction entre les deux cas $b > 0$ et $b < 0$ (le cas $b = 0$ est trivial puisque y_k est alors identiquement égal à une constante). En premier lieu, considérons la situation $b > 0$. Comme la relation linéaire $y_k = a + bx_k$ est satisfaite pour toutes les unités k et que $b > 0$, les relations qui suivent sont vérifiées : $L_{y,s}(a + bt) = a + bL_{x,s}(t)$; $U_{y,s}(a + bt) = a + bU_{x,s}(t)$ et $\beta_{y,s}(a + bt) = \beta_{x,s}(t)$. Ces relations mènent à $H_{y,s}(a + bt, y_k) = H_{x,s}(t, x_k)$. Il s'ensuit que $\hat{F}_{y,\text{cal}}(a + bt) = \hat{F}_{x,\text{cal}}(t)$. En outre, $\hat{F}_{y,\text{cal}}(a + bQ_{x,\alpha}) = \alpha$ et, en utilisant la relation $a + bQ_{x,\alpha} = Q_{y,\alpha}$, nous déduisons que $\hat{F}_{y,\text{cal}}(Q_{y,\alpha}) = \alpha$. Par conséquent, lorsqu'une relation exactement linéaire est vérifiée et que $b > 0$, $\hat{Q}_{y,\text{cal},\alpha} = \hat{F}^{-1}_{y,\text{cal}}(\alpha) = Q_{y,\alpha}$. En deuxième lieu, considérons le cas $b < 0$. Nous déduisons, dans ce cas, les relations qui suivent : $L_{y,s}(a + bt) = a + bU_{x,s}(t)$; $U_{y,s}(a + bt) = a + bL_{x,s}(t)$; $\beta_{y,s}(a + bt) = 1 - \beta_{x,s}(t)$ et $H_{y,s}(a + bt, y_k) = 1 - H_{x,s}(t, x_k)$. Puisque $b < 0$, la relation entre les quantiles de x et de y est donnée par $a + bQ_{x,\alpha} = Q_{y,1-\alpha}$. Alors, nous déduisons que $\hat{F}_{y,\text{cal}}(Q_{y,1-\alpha}) = \hat{F}_{y,\text{cal}}(a + bQ_{x,\alpha}) = 1 - \hat{F}_{x,\text{cal}}(Q_{x,\alpha}) = 1 - \alpha$. Donc, dans cette situation, $Q_{y,1-\alpha}$ est estimé exactement par $\hat{Q}_{y,\text{cal},1-\alpha}$. Autrement dit, lorsqu'une relation exacte est vérifiée, si $b > 0$, l'estimateur par calage proposé $\hat{Q}_{y,\text{cal},\alpha}$ produit des estimateurs parfaits, de biais nul et de variance $Q_{y,\alpha}$. Par ailleurs, si $b < 0$ et que le calage est fait sur $Q_{x,\alpha}$, $Q_{y,1-\alpha}$ est estimé exactement par $\hat{Q}_{y,\text{cal},1-\alpha}$ (ce qui est sensé, parce que la relation parfaitement linéaire entre x et y est telle que le paramètre de pente est négatif).

Notons que, si $\hat{F}_{y,\text{cal}}$ et $\hat{F}_{x_j,\text{cal}}$ sont intervertibles aux points $Q_{y,\alpha}$ et $Q_{x_j,\alpha}$, les contraintes de calage exprimées en (3) peuvent être réécrites par rapport aux fonctions de répartition, ce qui signifie que les contraintes de calage fondées sur les quantiles sont équivalentes à $\hat{F}_{x_j,\text{cal}}(Q_{x_j,\alpha}) = \alpha$, $j = 1, \ldots, J$. Autrement dit, le problème de calage original peut être réexprimé par rapport aux fonctions de répartition avec les contraintes susmentionnées.

Une question naturelle est celle de savoir s'il existe une solution au problème d'optimisation (3). Même si celui-ci est formulé au moyen des fonctions de répartition interpolées, il n'est pas toujours possible de trouver une solution. Par exemple, si $Q_{x_j,\alpha}$ est plus petit ou plus grand

que toutes les valeurs x_{jk} figurant dans l'échantillon s, alors $\hat{F}_{x_j,\text{cal}}(Q_{x_j,\alpha})$ sera égal à zéro ou à un, quels que soient les poids **w** que l'on choisis. Donc, dans ces cas, il peut arriver que les contraintes de calage ne puissent être satisfaites. Cependant, quand le comportement de l'échantillon diffère considérablement de celui de la population cible, il convient d'examiner tout ajustement d'un œil très critique et de considérer la situation comme quelque peu extrême. En pratique, elle se produit rarement, à moins que l'on choisisse une valeur de α très proche de zéro ou de un. Notons qu'il pourrait être impossible d'obtenir une solution si la taille n de l'échantillon est petite. Le cas échéant, nous pourrions considérer le minimum ou le maximum d'échantillon comme un estimateur possible ou recourir au simple estimateur de la fonction de répartition fondé sur le plan de sondage.

Le deuxième problème éventuel est que certains poids w_k pourraient être négatifs. Dans ces conditions, $\hat{F}_{y,\text{cal}}$ n'est plus bijectif, ce qui ne cause pas d'ennui à condition que $\hat{F}^{-1}_{y,\text{cal}}(\alpha)$ demeure déterminé de façon unique. Nous pouvons éviter ce problème en contraignant tous les poids à des valeurs strictement positives, grâce à l'utilisation d'une métrique appropriée $\mathrm{D}(\cdot,\cdot)$. Cette approche a été adoptée par Kovačević (1997) (pour plus de précisions sur les fonctions de distance produisant des poids positifs, voir également Deville et Särndal (1992), ainsi que Singh et Mohl (1996)).

Remarque 1

Les estimateurs proposés des fonctions de répartition (6) et (7) s'appuient sur une interpolation linéaire. Par souci d'uniformisation, la fonction de répartition de la population, qui est aussi une fonction échelon, pourrait également être définie en se fondant sur une interpolation linéaire. En pratique, les deux définitions correspondent à des comportements qui ne diffèrent que légèrement si la population N est suffisamment grande. Cependant, il convient de souligner que, si la taille de la population N est assez faible, l'utilisation d'une interpolation pour définir les fonctions de répartition pourrait valoir la peine.

Remarque 2

Dans le problème d'optimisation (3), nous avons réalisé le calage sur un quantile particulier. Cette approche pourrait être étendue en permettant le calage sur un ensemble fini de quantiles, si ce genre d'information est disponible. Plus précisément, supposons que, pour une variable auxiliaire x, les quantiles α_m dénotés Q_{x,α_m}, $m = 1, \ldots, M$ sont connus, où $M < n-1$. Dans ce cas, nous pourrions envisager les contraintes de calage $\hat{F}_{x,\text{cal}}(Q_{x,\alpha_m}) = \alpha_m$, $m = 1, \ldots, M$ et résoudre le problème d'optimisation (3) avec ces contraintes supplémentaires. Naturellement, cette information produit une description plus complète de la distribution des variables auxiliaires; donc, l'efficacité des estimateurs par calage ainsi obtenus devrait, en principe, être plus grande.

Remarque 3

L'estimateur par calage proposé (9) est obtenu par calage sur les quantiles de population. Une autre possibilité a été examinée par Ren (2002) qui a calé les estimateurs sur les moments de population, jusqu'à l'ordre m, d'une même loi. Plus précisément, Ren (2002) a proposé des estimateurs par calage des quantiles satisfaisant des contraintes de la forme $\sum_s w_k x_k^m = \sum_U x_k^m$, $m = 0, 1, \ldots, M$. Le calage sur divers moments de la même loi est étroitement associé au calage sur divers quantiles de la même variable, et toutes ces contraintes offrent une description plus complète de la distribution de la variable auxiliaire. Pour d'autres généralisations du paradigme de calage sur les moments, consulter aussi Ren et Deville (2000), ainsi que Harms (2003).

3.2 Solution analytique des poids calés quand D est la métrique quadratique

Lorsque l'on adopte la fonction de distance quadratique (2), il est possible de dériver une solution explicite du problème d'optimisation (3). Cette situation est semblable à celle des estimateurs par calage des totaux, où les poids de l'estimateur GREG sont obtenus explicitement sous la métrique (2). Une analyse minutieuse du problème d'estimation dans le cas des quantiles révèle d'importantes similarités, dues au fait que les estimateurs donnés par (7) sont des sommes pondérées des variables $\{H_{x_j,s}(t, x_{jk}),\ k \in s\}$, $j = 1, \ldots, J$. Cela est énoncé dans la proposition 1.

Proposition 1 (poids calés pour la métrique quadratique). *Considérons la fonction de distance quadratique* (2). *Le vecteur de poids* **w** *qui résout le problème d'optimisation* (3) *satisfait la relation :*

$$w_k = d_k(1 + q_k \mathbf{a}_k' \boldsymbol{\lambda}_s),\ k \in s, \tag{10}$$

où le vecteur $\boldsymbol{\lambda}_s = (\lambda_0, \ldots, \lambda_J)'$ *est déterminé par la voie des* $J+1$ *contraintes sous la forme :*

$$\boldsymbol{\lambda}_s = \left(\sum_s d_k\, q_k\, \mathbf{a}_k\, \mathbf{a}_k'\right)^{-1}\left(\mathbf{T}_\mathbf{a} - \sum_s d_k\, \mathbf{a}_k\right), \tag{11}$$

avec $\mathbf{T}_\mathbf{a} = (N, \alpha, \ldots, \alpha)'$ *et les composantes de* $\mathbf{a}_k = (1, a_{1k}, \ldots, a_{Jk})'$ *sont données par*

$$a_{jk} = \begin{cases} N^{-1}, & x_{jk} \le L_{x_j,s}(Q_{x_j,\alpha}), \\ N^{-1}\beta_{x_j,s}(Q_{x_j,\alpha}), & x_{jk} = U_{x_j,s}(Q_{x_j,\alpha}), \\ 0, & x_{jk} > U_{x_j,s}(Q_{x_j,\alpha}), \end{cases}$$

avec $j = 1, \ldots, J$.

Preuve. Afin de prouver la proposition 1, soulignons d'abord que, puisque que la première contrainte $\sum_s w_k = N$ doit être satisfaite, il s'en suit que $\hat{F}_{x_j,\,\mathrm{cal}}(t) = N^{-1}\sum_s w_k H_{x_j,s}(t,\, x_{jk})$. En nous inspirant de Deville et Särndal (1992), nous pouvons montrer que le vecteur $\mathbf{a}_k = (1, a_{1k}, \ldots, a_{Jk})'$ satisfait

$\mathbf{a}_k =$

$$\left(1, \frac{\partial \hat{F}_{x_1,\,\mathrm{cal}}}{\partial w_k}, \ldots, \frac{\partial \hat{F}_{x_J,\,\mathrm{cal}}}{\partial w_k}\right)'\Bigg|_{\sum_s w_k = N;\, \hat{F}_{x_j,\mathrm{cal}}(Q_{x_j,\alpha}) = \alpha,\ j=1,\ldots,J}, \quad (12)$$

que nous évaluons maintenant explicitement. L'évaluation des dérivées nous donne $a_{jk} = N^{-1}H_{x_j,\,s}(t,\, x_{jk}),\quad j = 1, \ldots, J,$ évalué à $t = Q_{x_j,\,\alpha}$. Ceci mène à

$$a_{jk} = \begin{cases} N^{-1}, & x_{jk} \le L_{x_j,s}(Q_{x_j,\alpha}), \\ N^{-1}\beta_{x_j,s}(Q_{x_j,\alpha}), & x_{jk} = U_{x_j,s}(Q_{x_j,\alpha}), \\ 0, & x_{jk} > U_{x_j,s}(Q_{x_j,\alpha}), \end{cases}$$

$j = 1, \ldots, J,$ tel qu'annoncé.

Dans (11), $\mathbf{T}_\mathbf{a}$ peut être interprété comme étant la valeur espérée de $\sum_s d_k \mathbf{a}_k$. Dans l'estimateur de la fonction de répartition (6), les poids dérivés (10) dépendent des variables $\mathbf{a}_k,\ k \in s$ définies par (12). Notons qu'elles correspondent à une certaine transformation de la variable auxiliaire $\mathbf{x}_k$. La différence entre les poids utilisés pour les totaux et pour les quantiles tient à cette variable $\mathbf{a}_k$; lorsque $\mathbf{a}_k$ est remplacée par $\mathbf{x}_k$, nous obtenons les poids originaux pour les totaux. Par conséquent, il est utile d'interpréter cette nouvelle variable. Lors de l'estimation d'un total, l'effet sur la j^{e} contrainte de calage est mesuré par x_{jk}, pour chaque unité $k \in s$. Dans notre cadre, l'effet de l'unité k est maintenant donné par N^{-1} si $x_{jk} \le L_{x_j,s}(Q_{x_j,\alpha})$; il correspond au facteur $N^{-1}\beta_{x_j,s}(Q_{x_j,\alpha})$ quand $x_{jk} = U_{x_j,s}(Q_{x_j,\alpha})$ et est nul ailleurs. À la section 5, nous discuterons d'autres problèmes d'estimation menant à des variables $\mathbf{a}_k$ différentes.

Compte tenu des similarités entre les estimations des totaux et des quantiles, nous pouvons également envisager l'estimation de la variance. Nous abordons cette question à la sous-section suivante.

3.3 Estimation de la variance et intervalles de confiance

Comme nous l'avons décrit à la section précédente, l'estimateur $\hat{Q}_{y,\,\mathrm{cal},\,\alpha}$ présente plusieurs similarités avec l'estimateur GREG habituel pour les totaux de population. Les variables transformées données par (12) constituent la principale différence entre les estimateurs par calage des quantiles et ceux des totaux. Il se trouve que, étant donné la similarité structurelle avec les estimateurs par calage originaux, il est facile de déterminer un intervalle de confiance pour l'estimateur proposé $\hat{Q}_{y,\,\mathrm{cal},\,\alpha}$. Nous considérons la construction d'intervalles de confiance suivant l'approche de Woodruff (1952). L'intervalle de confiance est donné au résultat 1.

Résultat 1 (intervalle de confiance de Woodruff pour l'estimateur par calage des quantiles). *L'intervalle de confiance basé sur l'approche de Woodruff* (1952)*, lorsqu'on utilise l'estimateur par calage* (9) *pour le quantile* $Q_{y,\,\alpha}$*, est donné par*

$$[\hat{F}^{-1}_{y,\,\mathrm{cal}}(\hat{c}_{1y}),\, \hat{F}^{-1}_{y,\,\mathrm{cal}}(\hat{c}_{2y})], \quad (13)$$

où $\hat{c}_{1y} = \alpha - z_{1-\gamma/2}[\hat{V}\{\hat{F}_{y,\,\mathrm{cal}}(Q_{y,\,\alpha})\}]^{1/2}$ *et* $\hat{c}_{2y} = \alpha + z_{1-\gamma/2}[\hat{V}\{\hat{F}_{y,\,\mathrm{cal}}(Q_{y,\,\alpha})\}]^{1/2}$. *La procédure résultante donne un intervalle de confiance approximatif pour* $Q_{y,\,\alpha}$ *à un niveau de confiance* $1-\gamma$ *précisé.*

Preuve. En supposant que $\hat{F}_{y,\,\mathrm{cal},\,\alpha}(Q_{y,\,\alpha})$ suit approximativement une loi normale, il s'ensuit que $\Pr(c_{1y} \le \hat{F}_{y,\,\mathrm{cal},\,\alpha}(Q_{y,\,\alpha}) \le c_{2y})$ devrait être approximativement égal à $1-\gamma$, si l'on choisit

$$c_{1y} = \alpha - z_{1-\gamma/2}[V\{\hat{F}_{y,\,\mathrm{cal}}(Q_{y,\,\alpha})\}]^{1/2}, \quad (14)$$

$$c_{2y} = \alpha - z_{1-\gamma/2}[V\{\hat{F}_{y,\,\mathrm{cal}}(Q_{y,\,\alpha})\}]^{1/2}, \quad (15)$$

où z_γ représente le γ^{e} quantile de la loi normale standard $N(0,1)$. Puisque $\hat{F}_{y,\,\mathrm{cal},\,\alpha}(Q_{y,\,\alpha})$ représente essentiellement une moyenne d'échantillon, un estimateur possible de la variance justifié par la linéarisation de Taylor classique est donnée par

$$\hat{V}\{\hat{F}_{y,\,\mathrm{cal}}(Q_{y,\,\alpha})\} = N^{-2}\sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}}(w_k\, e_k)\,(w_l\, e_l), \quad (16)$$

où $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$; les poids $w_k, k \in s,$ correspondent aux poids calés (3) qui se réduisent à (10) lorsque D est la fonction de distance quadratique (2); les résidus sont donnés par $e_k = H_{y,\,s}(\hat{Q}_{y,\,\mathrm{cal},\,\alpha},\, y_k) - \mathbf{a}_k'\,\hat{\mathbf{B}}_s$ où

$$\hat{\mathbf{B}}_s = \left(\sum_s w_k\, q_k\, \mathbf{a}_k\, \mathbf{a}_k'\right)^{-1}\sum_s w_k\, q_k\, \mathbf{a}_k\, H_{y,\,s}(Q_{y,\,\mathrm{cal},\,\alpha},\, y_k)$$

représente l'estimateur des coefficients de régression. Puisque les constantes c_{1y} et c_{2y} données par (14) et (15) dépendent de $V\{\hat{F}_{y,\,\mathrm{cal}}(Q_{y,\,\alpha})\}$, nous pouvons les estimer en utilisant l'estimateur de la variance (16).

Dans le résultat 1, soulignons que Deville et Särndal (1992) ont préconisé l'utilisation d'un estimateur de la variance pondéré par les poids *w* semblable à (16) pour estimer la variance des estimateurs par calage des totaux de population. À la section 4, nous étudions empiriquement les propriétés de l'estimateur par calage proposé (9) et l'intervalle de confiance donné par l'expression (13).

4. Résultats des simulations

D'un point de vue pratique, il est logique d'étudier les propriétés des nouveaux estimateurs par calage pour des

échantillons finis et de les comparer à celles des estimateurs des quantiles généralement décrits dans la littérature. À la présente section, nous entreprenons des expériences par simulation afin d'illustrer empiriquement les nouveaux estimateurs. Nous nous intéressons surtout à leur biais et à leur variance empirique en population réelle. Nous étudions également les propriétés de couverture des intervalles de confiance, qui représentent aussi une question d'intérêt pratique.

Afin de répondre partiellement à ces questions, nous avons exécuté trois petites études par simulation, dans le cadre desquelles, pour plusieurs plans d'échantillonnage et pour des populations réelles, nous comparons l'estimateur par calage proposé des quantiles aux estimateurs généralement utilisés à l'heure actuelle. À la sous-section 4.1, nous décrivons en détail les populations étudiées et discutons des plans d'échantillonnage choisis. À la sous-section 4.2, nous présentons les estimateurs inclus dans l'étude empirique et, à la sous-section 4.3, nous décrivons les mesures fréquentistes (biais, variance et erreur quadratique moyenne empirique, taux de couverture des intervalles de confiance). Enfin, à la sous-section 4.4, nous analysons nos résultats empiriques.

4.1 Description des populations réelles et des plans d'échantillonnage

Les populations réelles sont représentées aux figures 1 à 6. La première, notée MU284, est tirée de Särndal et coll. (1992, annexe B). Elle est constituée de $N = 284$ municipalités de la Suède. Nous retenons comme variable d'intérêt la population en 1985 (variable P85) et nous supposons que l'information auxiliaire disponible est la population en 1975 (variable P75). Les deux variables sont mesurées en milliers. À la figure 1, la variable P85 est exprimée en fonction de la variable P75; comme prévu, la relation entre P85 et P75 est fortement linéaire. La variable P95 suit une loi hautement asymétrique, comme l'illustre la figure 2. Dans cette population, nous avons tiré 500 échantillons selon un plan d'échantillonnage aléatoire simple sans remise (EAS). En outre, nous avons exécuté la même étude sous un plan d'échantillonnage avec probabilités de sélection inégales, c'est-à-dire le plan d'échantillonnage de Poisson (PO). Les propriétés du plan d'échantillonnage PO sont décrites dans Särndal et coll. (1992). Étant donné la grande fourchette de valeurs de y, nous n'avons pas pu construire des probabilités de sélection d'échantillon π_k de la forme $\pi_k \propto y_k$, car certaines π_k auraient dû être supérieures à l'unité. Aux fins de notre illustration, nous avons déterminé les probabilités de sélection en utilisant la relation $\pi_k \propto 0,2y_k + 0,05$ (nous reconnaissons que les π_k sont idéalisées, puisque y_k n'est pas disponible en pratique). Sous le plan d'échantillonnage EAS (plan d'échantillonnage PO), nous considérons les tailles d'échantillon (tailles espérées d'échantillon) $n = 25$ et $n = 50$.

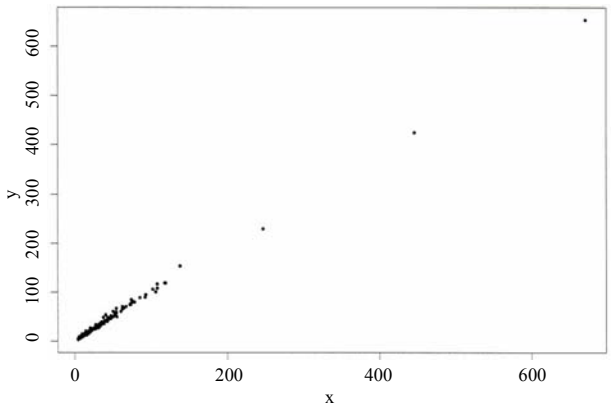


Figure 1. Population MU284, où y = P85 et x = P75 .

Pour la deuxième étude, nous avons choisi la population MU284, mais retenu comme variable d'intérêt y = RMT85, qui représente les recettes de l'impôt municipal de 1985 (en millions de couronnes). Ici, la variable auxiliaire choisie est x = REV84, qui représente les valeurs immobilières selon les évaluations de 1984 de chaque municipalité (en millions de couronnes). Comme le montre la figure 3, la relation entre x et y est quelque peu étalée pour les grandes valeurs de x. L'histogramme de la variable RMT85 révèle que celle-ci suit une loi asymétrique (figure 4). Pour cette étude, nous avons tiré, selon le plan d'échantillonnage EAS, 500 échantillons de taille $n = 25$ et $n = 50$.

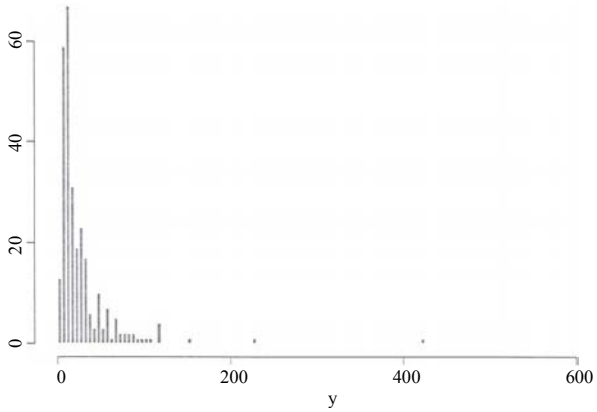


Figure 2. Histogramme de la variable P85 dans la population MU284.

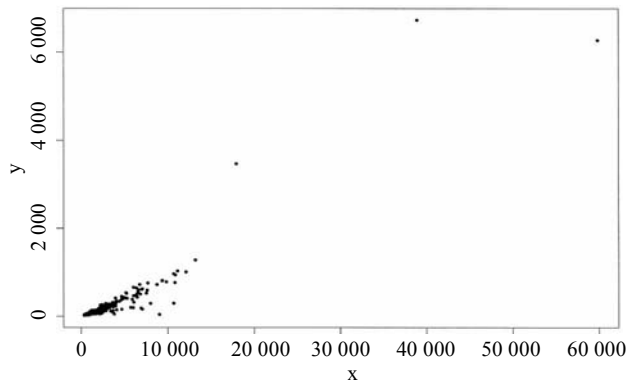


Figure 3. Population MU284, où y = RMT85 et x = REV84.

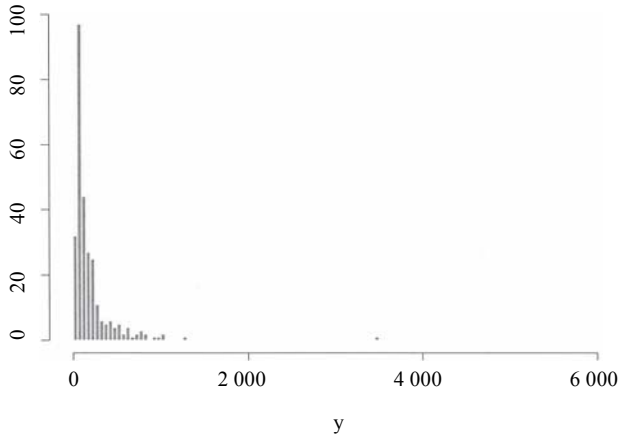


Figure 4. Histogramme de la variable RMT85 dans la population MU284.

La troisième population est basée sur un sous-échantillon aléatoire de l'*Enquête sur la dynamique du travail et du revenu*, noté SLID982 (SLID, *Survey of Labour and Income Dynamics*). L'enquête a été réalisée par Statistique Canada en 1998. Pour simplifier, nous n'avons sélectionné que les entrées pour lesquelles aucune valeur ne manquait. La taille du sous-échantillon est $N = 2\,000$ et, aux fins de notre étude, nous supposons qu'il s'agit d'une population (la taille originale de l'échantillon de cette enquête est d'environ 60 000). Le revenu imposable (en milliers de dollars) est la variable cible, tandis que la durée en mois de l'emploi courant est la variable auxiliaire. Comme l'illustre la figure 5, la relation linéaire entre le revenu imposable et la durée de l'emploi est moins prononcée. Cependant, les deux variables ne semblent pas être indépendantes. À la figure 6, nous voyons que la variable d'intérêt présente un coefficient d'asymétrie élevé. Nous avons tiré 500 échantillons à partir de la population SLID982, selon les plans d'échantillonnage EAS et PO. Nous avons considéré comme taille d'échantillon (taille espérée d'échantillon) $n = 100$ et $n = 200$. Pour l'échantillonnage PO, nous avons défini les probabilités de sélection de premier ordre, $\pi_k, k \in U$, en fonction de deux règles. Sous la première, nous avons créé les π_k de telle sorte qu'elles soient approximativement proportionnelles à la variable d'intérêt, c'est-à-dire le revenu imposable (aux fins de notre étude, nous supposons qu'il est possible de créer de telles π_k). Puisque certaines valeurs de y_k sont négatives dans la population, nous choisissons $p_{1k} = y_k - \min\{y_k, k \in U\} + 1$ et nous définissons $\pi_k = E(n_s)p_{1k} / \sum_U p_{1k}$, où $E(n_s)$ représente la taille espérée de l'échantillon, dans notre cas $E(n_s) = 100$ et 200. En vertu de la deuxième règle, les π_k ont été créées proportionnelles aux entrées du tableau 1. Autrement dit, pour chaque $k \in U$, il existe un facteur p_{2k}, qui est déterminé par le groupe âge-sexe de l'individu k. Alors, $\pi_k = E(n_s)p_{2k} / \sum_U p_{2k}$, où les facteurs p_{2k} sont donnés au tableau 1. Les facteurs p_{2k} du tableau 1 sont fondés sur un plan d'échantillonnage hypothétique, dans lequel nous supposons que ces facteurs fournissent des mesures de taille appropriée pour les unités dans les diverses classes âge-sexe (voir, par exemple, Särndal et coll. (1992, page 87)); pour ces unités, plus d'hommes que de femmes sont susceptibles d'être sélectionnés et pour les deux sexes, les adultes de 27 à 37 ans et ceux de 38 à 46 ans sont plus susceptibles que les autres d'être inclus dans l'échantillon.

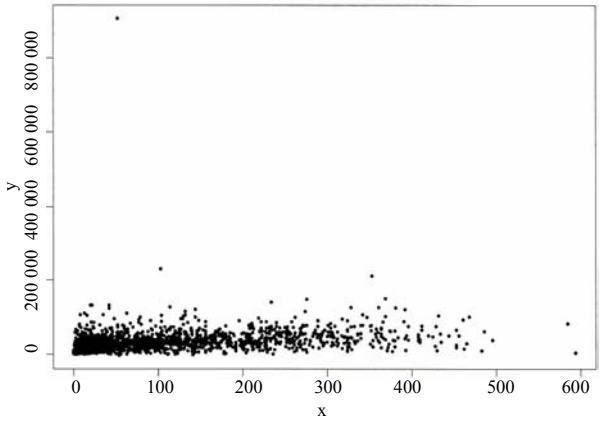


Figure 5. Population SLID982, où la variable dépendante est le revenu imposable et la variable indépendante, la durée de l'emploi courant (en mois).

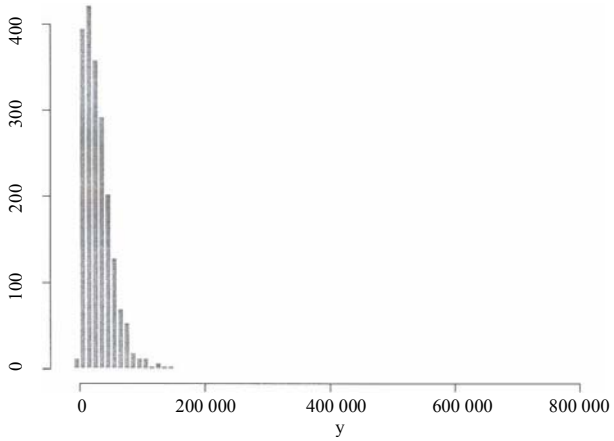


Figure 6. Histogramme du revenu imposable dans la population SLID982.

Tableau 1
Facteur p_{2k} selon l'âge et le sexe de l'individu k, dans la population SLID982

		Âge			
		16 à 25 ans	27 à 37 ans	38 à 46 ans	47 à 69 ans
Sexe	Hommes	3	6	5	4
	Femmes	1	2	3	2

Dans ces trois études, nous estimons les quartiles, c'est-à-dire les paramètres de population $Q_{y,\alpha}$ pour lesquels $\alpha = 0{,}25$, $0{,}5$ et $0{,}75$. Puisque les variables d'intérêt présentent une distribution fortement asymétrique, il semble particulièrement intéressant d'étudier le quartile correspondant à

$\alpha = 0{,}75$, en plus de la médiane et du premier quartile. À la section suivante, nous décrivons les estimateurs utilisés dans l'étude.

4.2 Estimateurs inclus dans l'étude empirique

Puisque l'un de nos objectifs est de proposer des estimateurs ayant des propriétés raisonnables en ce qui concerne le biais, la variance et les taux de couverture des intervalles de confiance, nous comparons le nouvel estimateur défini par (9) fondé sur la métrique (2) à certains estimateurs des quantiles populaires proposés dans la littérature.

Pour commencer, nous incluons l'estimateur fondé sur le plan de sondage simple basé sur l'inversion de l'estimateur $\hat{F}_y(t) = \sum_s d_k\, H_{y,s}(t, y_k) / \sum_s d_k$:

$$\hat{Q}_{y,\mathrm{HT},\alpha} = \hat{F}_y^{-1}(\alpha). \tag{17}$$

L'estimateur (17) n'utilise pas d'information auxiliaire. Un estimateur possible de la variance est

$$\hat{V}\{\hat{F}_y(Q_{y,\alpha})\} = \hat{N}^{-2} \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}} \left\{ \frac{H_{y,s}(\hat{Q}_{y,\mathrm{HT},\alpha}, y_k) - \alpha)}{\pi_k} \right\} \left\{ \frac{H_{y,s}(\hat{Q}_{y,\mathrm{HT},\alpha}, y_l) - \alpha}{\pi_l} \right\},$$

où $\hat{N} = \sum_s d_k$, et les intervalles de confiance peuvent être calculés au moyen de

$$[\hat{F}_y^{-1}(\tilde{c}_{1y}),\ \hat{F}_y^{-1}(\tilde{c}_{2y})],$$

où

$$\tilde{c}_{1y} = \alpha - z_{1-\gamma/2}[\hat{V}\{\hat{F}_y(Q_{y,\alpha})\}]^{1/2}, \tag{18}$$

$$\tilde{c}_{2y} = \alpha + z_{1-\gamma/2}[\hat{V}\{\hat{F}_y(Q_{y,\alpha})\}]^{1/2}. \tag{19}$$

Pour plus de détails, consulter Särndal et coll. (1992, page 202).

Nous incluons également dans notre étude empirique l'estimateur fondé sur un modèle de Chambers et Dunstan (1986), qui est motivé par un modèle de superpopulation linéaire $y_k = \beta_0 + \boldsymbol{\beta}'\mathbf{x}_k + \epsilon_k$, $k \in U$, où ϵ_k forme une suite de variables aléatoires indépendantes et de même loi de moyenne nulle et de variance finie. Leur estimateur est défini par

$$\hat{Q}_{y,\mathrm{CD},\alpha} = \inf\{t \mid \hat{F}_{y,\mathrm{CD}}(t) \geq \alpha\}, \tag{20}$$

où $\hat{F}_{y,\mathrm{CD}}(t) = N^{-1}\{\sum_s H(t - y_k) + \sum_{U/s} \hat{G}(t - \hat{y}_k)\}$ représente un estimateur fondé sur un modèle de la fonction de répartition,

$$\hat{G}(u) = n^{-1} \sum_s H(u - \hat{\epsilon}_k) \tag{21}$$

représente la fonction de répartition empirique des résidus $\hat{\epsilon}_k = y_k - \hat{y}_k$, $k \in s$, et $\hat{y}_k = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\mathbf{x}_k$, $k \in U/s$ correspond aux prédictions selon les moindres carrés. Puisque l'estimateur (20) impute essentiellement la valeur inconnue y_k pour $k \in U/s$, soulignons qu'il nécessite la connaissance complète de $\mathbf{x}_k$ pour $k \in U$.

La construction d'un intervalle de confiance pour $\hat{Q}_{y,\mathrm{CD},\alpha}$ repose sur l'estimation de la variance $V\{\hat{F}_{y,\mathrm{CD}}(t)\}$. Toutefois, ce problème d'estimation pose des difficultés, puisque toute formule analytique de la variance dépend du modèle hypothétique. En outre, les expressions analytiques de ce genre font intervenir des estimateurs à noyau de la densité, qui demandent beaucoup de calculs et dépendent du choix d'une fonction noyau et d'une fenêtre (bandwidth). Pour toutes ces raisons, nous avons décidé d'appliquer les estimateurs de la variance par le jackknife avec suppression d'une unité étudiés par Wu et Sitter (2001), qui ont montré la convergence des estimateurs proposés de la variance. Dans le contexte des sondages, diverses méthodes de rééchantillonnage, y compris le jackknife, sont présentées dans Kovar, Rao et Wu (1988). La technique du jackknife consiste à supprimer une unité et à recalculer l'estimateur. Soit $s_i = s/\{i\}$ l'échantillon sans l'unité i. Considérons $\hat{\beta}_{0i}$ et $\hat{\boldsymbol{\beta}}_i$, les estimateurs par la régression de β_0 et $\boldsymbol{\beta}$ calculés sur s_i. Sous un modèle de régression simple, définissons

$$F_i^* = (n-1)^{-1} \sum_{k \in s_i} \left[N^{-1} \sum_{l \in U/s} H\{\hat{Q}_{y,\mathrm{CD},\alpha} - \hat{\beta}_i(x_l - x_k) - y_k\} \right].$$

Un estimateur de la variance convergent de $V\{\hat{F}_{y,\mathrm{CD}}(Q_{y,\mathrm{CD},\alpha})\}$ est donné par

$$\hat{V}_{y,\mathrm{CD}}\{\hat{F}_{y,\mathrm{CD}}(Q_{y,\alpha})\} = \frac{n-1}{n} \sum_{i \in s} (F_i^* - \bar{F}^*)^2 + \frac{f(1-f)}{N-n} \sum_{k \in U/s} \hat{G}(\hat{Q}_{y,\mathrm{CD},\alpha} - \hat{y}_k)\{1 - \hat{G}(\hat{Q}_{y,\mathrm{CD},\alpha} - \hat{y}_k)\},$$

où $f = n/N$ est la fraction d'échantillonnage, $\bar{F}^* = n^{-1}\sum_s F_i^*$, et $\hat{G}$ est donné par (21). En partant de $\hat{V}\{\hat{F}_{y,\mathrm{CD}}(Q_{y,\alpha})\}$, nous pouvons maintenant calculer les intervalles de confiance pour $Q_{y,\alpha}$ en suivant l'approche de l'inversion.

Puisque notre méthode ne nécessite que la connaissance du vecteur des quantiles $\mathbf{Q}_{\mathbf{x},\alpha}$, nous incluons dans notre étude les estimateurs par le ratio et par la différence des quantiles étudiés dans Rao et coll. (1990) :

$$\hat{Q}_{y,\mathrm{ra},\alpha} = Q_{x,\alpha}(\hat{Q}_{y,\mathrm{HT},\alpha} / \hat{Q}_{x,\mathrm{HT},\alpha}), \tag{22}$$

$$\hat{Q}_{y,\mathrm{diff},\alpha} = \hat{Q}_{y,\mathrm{HT},\alpha} + \hat{R}(Q_{x,\alpha} - \hat{Q}_{x,\mathrm{HT},\alpha}), \tag{23}$$

où $\hat{Q}_{y,\mathrm{HT},\alpha}$ est donné par (17) et $\hat{Q}_{x,\mathrm{HT},\alpha}$ est calculé de la même façon; l'estimateur par le ratio donné par $\hat{R} = \sum_s d_k y_k / \sum_s d_k x_k$ fournit un estimateur convergent de $R = \sum_U y_k / \sum_U x_k$. Notons que les estimateurs (22) et (23) sont élaborés en se fondant sur une variable auxiliaire

scalaire, c'est-à-dire $J = 1$. Des estimateurs de la variance valides de (22) et (23) sont donnés par :

$$\hat{V}(\hat{Q}_{y,\mathrm{ra},\alpha}) = \hat{V}(\hat{Q}_{y,\mathrm{HT},\alpha}) + \left(\frac{\hat{Q}_{y,\mathrm{HT},\alpha}}{\hat{Q}_{x,\mathrm{HT},\alpha}}\right)^2 \hat{V}(\hat{Q}_{x,\mathrm{HT},\alpha}) - 2\,\frac{\hat{Q}_{y,\mathrm{HT},\alpha}}{\hat{Q}_{x,\mathrm{HT},\alpha}}\,\hat{C}(\hat{Q}_{y,\mathrm{HT},\alpha}, \hat{Q}_{x,\mathrm{HT},\alpha}),$$

$$\hat{V}(\hat{Q}_{y,\mathrm{diff},\alpha}) = \hat{V}(\hat{Q}_{y,\mathrm{HT},\alpha}) + \hat{R}^2\,\hat{V}(\hat{Q}_{y,\mathrm{HT},\alpha}) - 2\hat{R}\,\hat{C}(\hat{Q}_{y,\mathrm{HT},\alpha}, \hat{Q}_{x,\mathrm{HT},\alpha}).$$

Ces estimateurs s'appuient sur la variance de $\hat{Q}_{y,\mathrm{HT},\alpha}$, ainsi que sur la covariance de $\hat{Q}_{y,\mathrm{HT},\alpha}$ et de $\hat{Q}_{x,\mathrm{HT},\alpha}$, qui sont estimées selon l'approche de Woodruff (1952) :

$$\hat{V}(\hat{Q}_{y,\mathrm{HT},\alpha}) = \frac{W_y^2}{4z_{1-\gamma/2}^2},$$

$$\hat{C}(\hat{Q}_{y,\mathrm{HT},\alpha}, \hat{Q}_{x,\mathrm{HT},\alpha}) = \frac{W_y W_x \hat{C}\{\hat{F}_x(Q_{x,\alpha}), \hat{F}_y(Q_{y,\alpha})\}}{4z_{1-\gamma/2}^2[\hat{V}\{\hat{F}_x(Q_{x,\alpha})\}]^{1/2}[\hat{V}\{\hat{F}_y(Q_{y,\alpha})\}]^{1/2}},$$

où $W_y = \hat{F}_y^{-1}(\tilde{c}_{2y}) - \hat{F}_y^{-1}(\tilde{c}_{1y})$ et $W_x = \hat{F}_x^{-1}(\tilde{c}_{2x}) - \hat{F}_x^{-1}(\tilde{c}_{1x})$ représentent les intervalles de Woodruff associés à y et à x, avec $\tilde{c}_{1y}$ et $\tilde{c}_{2y}$ définis par (18) et (19), $\tilde{c}_{1x} = \alpha - z_{1-\gamma/2}[\hat{V}\{\hat{F}_x(Q_{x,\alpha})\}]^{1/2}$, $\tilde{c}_{2x} = \alpha + z_{1-\gamma/2}[\hat{V}\{\hat{F}_x(Q_{x,\alpha})\}]^{1/2}$ et

$$\hat{C}\{\hat{F}_y(Q_{y,\alpha}),\ \hat{F}_x(Q_{x,\alpha})\} = \hat{N}^{-2}\sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\left\{\frac{H_{y,s}(\hat{Q}_{y,\mathrm{HT},\alpha},\ y_k) - \alpha}{\pi_k}\right\}\left\{\frac{H_{x,s}(\hat{Q}_{x,\mathrm{HT},\alpha},\ x_l) - \alpha}{\pi_l}\right\}.$$

Brièvement, nous nous attendons à ce que $\hat{Q}_{y,\mathrm{CD},\alpha}$ donne de bons résultats quand le modèle linéaire décrit la population adéquatement, ce qui motive la comparaison de la nouvelle méthode à un estimateur fondé sur un modèle. En outre, il semble intéressant d'évaluer $\hat{Q}_{y,\mathrm{cal},\alpha}$ et les principales propositions fondées sur le plan de sondage, telles que $\hat{Q}_{y,\mathrm{diff},\alpha}$ et $\hat{Q}_{y,\mathrm{ra},\alpha}$. Les estimateurs $\hat{Q}_{y,\mathrm{cal},\alpha}$, $\hat{Q}_{y,\mathrm{diff},\alpha}$ et $\hat{Q}_{y,\mathrm{ra},\alpha}$ utilisent $Q_{x,\alpha}$ uniquement pour améliorer les estimations et tiennent compte du plan d'échantillonnage; il s'agit donc de concurrents naturels. Notons que les divers estimateurs inclus dans notre étude sont élaborés sous diverses hypothèses quant à la dimension du vecteur de variables auxiliaires $\mathbf{x}$ et à la disponibilité de $\mathbf{x}_k$. Le tableau 2 donne une comparaison des divers estimateurs décrits à la présente section.

Tableau 2
Comparaison des estimateurs par calage proposés et de certains estimateurs importants des quantiles proposés dans la littérature, en ce qui concerne la dimension J de $\mathbf{x}$ et l'information requise au sujet de $\mathbf{x}$

Estimateur	Dimension de $\mathbf{x}$	Information requise sur $\mathbf{x}$
$\hat{Q}_{y,\mathrm{HT},\alpha}$	s.o.	aucune
$\hat{Q}_{y,\mathrm{CD},\alpha}$	$J \geq 1$	$\mathbf{x}_k, k \in U/s$
$\hat{Q}_{y,\mathrm{ra},\alpha}$	$J = 1$	$Q_{x,\alpha}$
$\hat{Q}_{y,\mathrm{diff},\alpha}$	$J = 1$	$Q_{x,\alpha}$
$\hat{Q}_{y,\mathrm{cal},\alpha}$	$J \geq 1$	$Q_{\mathbf{x},\alpha}$

4.3 Mesures fréquentistes

Notre objectif est d'évaluer les estimateurs en nous basant sur le biais et la variance. D'autres considérations importantes sont l'erreur quadratique moyenne (EQM) et les taux de couverture des intervalles de confiance.

Soit $\hat{Q}_{y,\alpha}$ un estimateur du quantile de population $Q_{y,\alpha}$. Supposons que $\hat{Q}_{y,\alpha}^{(v)}$ est l'estimateur du quantile calculé en utilisant l'échantillon v, $v = 1, \ldots, K$. La moyenne de Monte Carlo E_{MC}, le biais de Monte Carlo B_{MC} et la variance de Monte Carlo V_{MC} sont donnés par les formules usuelles, c'est-à-dire

$$E_{\mathrm{MC}}(\hat{Q}_{y,\alpha}) = K^{-1}\sum_{v=1}^{K}\hat{Q}_{y,\alpha}^{(v)},$$

$$B_{\mathrm{MC}} = E_{\mathrm{MC}}(\hat{Q}_{y,\alpha}) - Q_{y,\alpha},$$

$$V_{\mathrm{MC}}(\hat{Q}_{y,\alpha}) = K^{-1}\sum_{v=1}^{K}\{\hat{Q}_{y,\alpha}^{(v)} - E_{\mathrm{MC}}(\hat{Q}_{y,\alpha})\}^2.$$

Notre critère principal de détermination de l'efficacité est l'EQM de Monte Carlo, définie par $\mathrm{EQM}_{\mathrm{MC}} = K^{-1}\sum_{v=1}^{K}(\hat{Q}_{y,\alpha}^{(v)} - Q_{y,\alpha})^2$. Nous calculons les intervalles de confiance au niveau de confiance de 95 %, selon les procédures décrites aux sections qui précèdent. Pour un estimateur $\hat{Q}_{y,\alpha}^{(v)}$ et l'estimateur de sa variance $\hat{V}^{(v)}$, $v = 1, \ldots, K$, les taux de couverture (TC) au niveau de confiance de 95 % sont calculés selon l'expression

$$\mathrm{TC}(\hat{Q}_{y,\alpha}) = K^{-1}\sum_{v=1}^{K} I\left(\left\{Q_{y,\alpha} \in \left[\hat{Q}_{y,\alpha}^{(v)} - 1.96\sqrt{\hat{V}^{(v)}},\ \hat{Q}_{y,\alpha}^{(v)} + 1.96\sqrt{\hat{V}^{(v)}}\right]\right\}\right),$$

où $I(A)$ est la fonction indicateur de l'ensemble A. Les taux de couverture sont donnés dans la colonne intitulée TC. Rappelons que nous adoptons $K = 500$ pour toutes les études.

4.4 Discussion des résultats empiriques

Les résultats sont présentés aux tableaux 3 à 8. Nous commençons par discuter des résultats exposés aux tableaux 3 et 4, obtenus en échantillonnant la population MU284 selon un plan aléatoire simple EAS ou un plan de Poisson PO. Comme on peut le voir, tous les estimateurs ont le même comportement dans les deux études. L'estimateur fondé sur un modèle $\hat{Q}_{y,\mathrm{CD},\alpha}$ semble être le plus efficace parmi ceux analysés lors de l'examen du cas $\alpha = 0{,}75$ et est, en général, très efficace. Nous nous attendions à ce résultat, puisque la relation entre $x = \mathrm{P75}$ et $y = \mathrm{P85}$ est fortement linéaire et que l'estimateur fondé sur le modèle est basé sur un modèle hypothétique de régression simple. Par contre, pour $\alpha = 0{,}25$, les différences d'efficacité sont prononcées par rapport aux autres estimateurs fondés sur des données auxiliaires. Ceux n'utilisant que $Q_{x,\alpha}$ comme information sur la variable auxiliaire produisent des résultats assez semblables. Lorsque la taille d'échantillon est petite, les taux de couverture s'écartent habituellement du niveau nominal de 95 %, surtout ceux de $\hat{Q}_{y,\mathrm{cal},\alpha}$, qui sont quelque peu sous-estimés. Cependant, nous observons une certaine amélioration pour $n = 50$, ce qui témoigne de la cohérence des procédures étudiées. Par ailleurs, les taux de couverture de $\hat{Q}_{y,\mathrm{ra},\alpha}$ et de $\hat{Q}_{y,\mathrm{diff},\alpha}$ sont toujours égaux à un, ce qui donne à penser que les variances de ces estimateurs sont surestimées. À cause d'une composante de biais importante dans l'EQM, les taux de couverture de l'estimateur fondé sur un modèle se détériorent parfois à mesure qu'augmente la taille de l'échantillon. Nous obtenons les meilleurs taux de couverture en utilisant l'estimateur HT simple, $\hat{Q}_{y,\mathrm{HT},\alpha}$, qui est cependant moins efficace que les autres.

Le tableau 5 donne les résultats pour la deuxième population, qui est la population MU284, avec $y = \mathrm{RMT85}$ et $x = \mathrm{REV84}$. La figure 3 semble révéler un phénomène d'hétéroscédasticité dans cette population. Par conséquent, puisque l'utilisation de l'estimateur par le ratio est justifiée quand la population sous-jacente manifeste ce genre de comportement, il n'est pas étonnant que l'estimateur par le ratio $\hat{Q}_{y,\mathrm{ra},\alpha}$ donne de bons résultats dans cette situation particulière; il surpasse $\hat{Q}_{y,\mathrm{diff},\alpha}$ dans plusieurs cas. Pour toutes les tailles d'échantillon, l'estimateur par le ratio se comporte généralement mieux que $\hat{Q}_{y,\mathrm{cal},\alpha}$. Cependant, pour $n = 50$, l'estimateur par calage semble donner des résultats aussi bons ou légèrement meilleurs que l'estimateur par le ratio. Dans cette expérience, le biais et la variance de l'estimateur fondé sur un modèle $\hat{Q}_{y,\mathrm{CD},\alpha}$ font augmenter sensiblement l'EQM. En outre, dans certains cas, nous n'avons pas pu obtenir les intervalles de confiance pour cet estimateur, car la méthode de Woodruff ne convient pas lorsque la variance est extrêmement grande (l'intervalle de Woodruff devient trop grand et l'hypothèse de linéarité de la fonction de répartition dans cet intervalle n'est plus vérifiée). Nous pensons qu'un modèle tenant compte de l'hétéroscédasticité pourrait améliorer les propriétés de l'estimateur fondé sur un modèle. Cela met en relief le fait que, pour que l'efficacité des estimateurs fondés sur un modèle soit grande, le modèle doit être spécifié correctement.

Les résultats des tableaux 6 à 8 ont trait à la population SLID982, sous des plans d'échantillonnage EAS et PO avec deux règles pour la détermination des probabilités π_k. Tous les estimateurs présentés au tableau 6 donnent des estimations raisonnablement bonnes du premier quartile et de la médiane, sauf l'estimateur par le ratio $\hat{Q}_{y,\mathrm{ra},\alpha}$, qui est le moins efficace. Le fait que la relation entre les variables dépendante et indépendante ne corresponde pas exactement à un modèle linéaire pourrait expliquer partiellement la performance médiocre de l'estimateur par le ratio dans ce cas. La relation entre x et y n'est pas proportionnelle, si bien que l'estimateur par la différence $\hat{Q}_{y,\mathrm{diff},\alpha}$ semble être préférable à $\hat{Q}_{y,\mathrm{ra},\alpha}$. Cependant, pour $\alpha = 0{,}75$, ces estimateurs affichent l'EQM la plus élevée, étant tous deux les moins efficaces. Curieusement, dans cette partie de l'expérience, $\hat{Q}_{y,\mathrm{cal},\alpha}$ est supérieur aux estimateurs fondés sur le plan de sondage en ce qui concerne l'EQM. Toutefois, si α est petit, $\hat{Q}_{y,\mathrm{diff},\alpha}$ et $\hat{Q}_{y,\mathrm{cal},\alpha}$ donnent des résultats similaires. Notons que, pour un échantillon de plus grande taille, $\hat{Q}_{y,\mathrm{cal},\alpha}$ et $\hat{Q}_{y,\mathrm{CD},\alpha}$ sont les plus efficaces pour la médiane et le troisième quartile. En fait, l'estimateur fondé sur un modèle $\hat{Q}_{y,\mathrm{CD},\alpha}$ surpasse légèrement $\hat{Q}_{y,\mathrm{cal},\alpha}$, mais il faut souligner qu'il utilise plus d'information auxiliaire.

Les tableaux 7 et 8 présentent les résultats sous les plans d'échantillonnage de Poisson. En général, les estimateurs fondés sur le plan de sondage donnent des résultats fort comparables à ceux obtenus sous un plan d'échantillonnage aléatoire simple (EAS). Par contre, il n'en n'est pas ainsi de l'estimateur fondé sur un modèle, qui est moins efficace, vraisemblablement parce qu'il n'intègre pas d'information au sujet du plan d'échantillonnage. Plus précisément, le tableau 7 présente les résultats des simulations sous échantillonnage PO, en utilisant la première règle pour les π_k, $k \in U$. Les taux de couverture de l'estimateur fondé sur un modèle sont particulièrement décevants dans cette expérience; les composantes de biais sont trop importantes dans l'EQM. Les estimateurs fondés sur le plan de sondage fournissent des taux de couverture empiriques nettement plus proche du niveau de confiance nominal de 95 %. Pour les valeurs moyennes et grandes de α, $\hat{Q}_{y,\mathrm{cal},\alpha}$ est l'estimateur le plus efficace. En fait, l'estimateur par calage $\hat{Q}_{y,\mathrm{cal},\alpha}$ donne de bons résultats dans cette expérience. Enfin, le tableau 8 contient les résultats obtenus sous échantillonnage PO avec la deuxième règle pour les π_k. Dans ce cas, $\hat{Q}_{y,\mathrm{ra},\alpha}$ est l'estimateur le moins efficace pour le premier quartile et la médiane, et $\hat{Q}_{y,\mathrm{diff},\alpha}$ est le moins efficace pour $\alpha = 0{,}75$. En général, $\hat{Q}_{y,\mathrm{cal},\alpha}$ est supérieur aux autres estimateurs dans cette situation, offrant la plus grande efficacité.

Tableau 3
Résultats des simulations de Monte Carlo pour l'échantillonnage de la population MU284, $y = \text{P85}$, $x = \text{P75}$, sous un plan d'échantillonnage EAS. Nombre de répétitions fixé à $K = 500$

		$n = 25$				$n = 50$			
α	Estimateur	B_{MC}	V_{MC}	EQM_{MC}	TC	B_{MC}	V_{MC}	EQM_{MC}	TC
0,25	$\hat{Q}_{y,\text{cal},\alpha}$	−0,0343	0,5075	0,5077	0,886	−0,0499	0,2437	0,2457	0,828
	$\hat{Q}_{y,\text{HT},\alpha}$	−0,0266	2,3196	2,3157	0,952	0,0035	1,1087	1,1065	0,936
	$\hat{Q}_{y,\text{ra},\alpha}$	−0,1444	0,3869	0,4070	1,000	−0,0774	0,1684	0,1741	1,000
	$\hat{Q}_{y,\text{diff},\alpha}$	−0,1486	0,3901	0,4114	1,000	−0,0734	0,1723	0,1774	1,000
	$\hat{Q}_{y,\text{CD},\alpha}$	0,4855	0,2791	0,5143	0,906	0,5485	0,1981	0,4985	0,824
0,5	$\hat{Q}_{y,\text{cal},\alpha}$	−0,2762	1,6499	1,7229	0,918	−0,2835	0,9585	1,0370	0,944
	$\hat{Q}_{y,\text{HT},\alpha}$	0,2605	12,5161	12,5589	0,922	−0,0064	5,8466	5,8349	0,916
	$\hat{Q}_{y,\text{ra},\alpha}$	−0,2586	0,8828	0,9479	1,000	−0,4296	0,6701	0,8533	1,000
	$\hat{Q}_{y,\text{diff},\alpha}$	−0,2775	0,9898	1,0648	1,000	−0,4331	0,7492	0,9352	1,000
	$\hat{Q}_{y,\text{CD},\alpha}$	0,9431	0,4054	1,2940	0,866	0,9884	0,2410	1,2175	0,714
0,75	$\hat{Q}_{y,\text{cal},\alpha}$	−0,6229	3,3241	3,7055	0,614	−0,3661	1,8107	1,9411	0,710
	$\hat{Q}_{y,\text{HT},\alpha}$	−0,1414	53,1951	53,1088	0,948	−0,3692	18,8586	18,9572	0,964
	$\hat{Q}_{y,\text{ra},\alpha}$	−0,7925	3,0021	3,6242	1,000	−1,0004	1,4594	2,4573	1,000
	$\hat{Q}_{y,\text{diff},\alpha}$	−0,8230	3,4379	4,1083	1,000	−1,0396	1,5267	2,6044	1,000
	$\hat{Q}_{y,\text{CD},\alpha}$	0,4343	0,5108	0,6984	0,954	0,4485	0,2618	0,4624	0,974

Tableau 4
Résultats des simulations de Monte Carlo pour l'échantillonnage de la population MU284, $y = \text{P85}$, $x = \text{P75}$, sous un plan d'échantillonnage PO. Nombre de répétitions fixé à $K = 500$

		$n = 25$				$n = 50$			
α	Estimateur	B_{MC}	V_{MC}	EQM_{MC}	TC	B_{MC}	V_{MC}	EQM_{MC}	TC
0,25	$\hat{Q}_{y,\text{cal},\alpha}$	−0,0441	0,4886	0,4896	0,888	−0,0169	0,2601	0,2599	0,828
	$\hat{Q}_{y,\text{HT},\alpha}$	−0,1698	2,2825	2,3068	0,936	−0,0384	1,1828	1,1819	0,928
	$\hat{Q}_{y,\text{ra},\alpha}$	−0,1509	0,3857	0,4076	1,000	−0,0913	0,2100	0,2179	1,000
	$\hat{Q}_{y,\text{diff},\alpha}$	−0,1634	0,3821	0,4080	1,000	−0,0877	0,2149	0,2221	1,000
	$\hat{Q}_{y,\text{CD},\alpha}$	0,6709	0,3310	0,7805	0,896	0,8792	0,1339	0,9066	0,554
0,5	$\hat{Q}_{y,\text{cal},\alpha}$	−0,3610	1,4881	1,6155	0,920	−0,3236	0,8833	0,9863	0,936
	$\hat{Q}_{y,\text{HT},\alpha}$	−0,0612	11,3969	11,3778	0,926	−0,2712	5,2672	5,3302	0,906
	$\hat{Q}_{y,\text{ra},\alpha}$	−0,3735	1,0009	1,1385	1,000	−0,4130	0,5486	0,7181	1,000
	$\hat{Q}_{y,\text{diff},\alpha}$	−0,3962	1,1271	1,2818	1,000	−0,4217	0,5962	0,7729	1,000
	$\hat{Q}_{y,\text{CD},\alpha}$	1,1740	0,4947	1,8719	0,820	1,3297	0,2146	1,9822	0,474
0,75	$\hat{Q}_{y,\text{cal},\alpha}$	−0,6420	2,6605	3,0674	0,608	−0,4476	1,6212	1,8183	0,708
	$\hat{Q}_{y,\text{HT},\alpha}$	−0,6200	51,2934	51,5752	0,956	−0,6632	17,3625	17,7677	0,966
	$\hat{Q}_{y,\text{ra},\alpha}$	−0,8686	2,8841	3,6329	1,000	−0,9683	1,6494	2,5837	1,000
	$\hat{Q}_{y,\text{diff},\alpha}$	−0,9025	2,9826	3,7911	1,000	−1,0177	1,6340	2,6665	1,000
	$\hat{Q}_{y,\text{CD},\alpha}$	0,4620	0,4501	0,6627	0,982	0,5388	0,2329	0,5228	0,980

Tableau 5
Résultats des simulations de Monte Carlo pour l'échantillonnage de la population MU284, y = RMT85, x = REV84, sous un plan d'échantillonnage EAS. Nombre de répétitions fixé à $K = 500$

		$n = 25$				$n = 50$			
α	Estimateur	B_{MC}	V_{MC}	EQM_{MC}	TC	B_{MC}	V_{MC}	EQM_{MC}	TC
0,25	$\hat{Q}_{y,\text{cal},\alpha}$	1,0161	51,5421	52,4714	0,892	0,6499	24,0662	24,4404	0,954
	$\hat{Q}_{y,\text{HT},\alpha}$	0,3733	110,2572	110,1760	0,960	0,3383	47,2921	47,3120	0,962
	$\hat{Q}_{y,\text{ra},\alpha}$	3,0025	65,4135	74,2979	0,998	2,3856	30,7284	36,3580	0,992
	$\hat{Q}_{y,\text{diff},\alpha}$	2,5952	107,7891	114,3084	0,994	2,4083	55,6977	61,3862	0,986
	$\hat{Q}_{y,\text{CD},\alpha}$	−16,5165	1661,0257	1930,4983	0,990	−17,3217	820,7447	1119,1443	0,960
0,5	$\hat{Q}_{y,\text{cal},\alpha}$	−1,6219	215,0326	217,2330	0,870	−0,3419	118,2125	118,0930	0,922
	$\hat{Q}_{y,\text{HT},\alpha}$	0,0075	763,6236	762,0964	0,910	−0,3977	331,2357	330,7314	0,914
	$\hat{Q}_{y,\text{ra},\alpha}$	0,7712	212,8298	212,9988	0,996	−0,2810	136,4382	136,2443	0,996
	$\hat{Q}_{y,\text{diff},\alpha}$	0,3415	283,6718	283,2210	0,998	−1,0104	201,3707	201,9889	0,998
	$\hat{Q}_{y,\text{CD},\alpha}$	17,6124	190,0045	499,8199	n.a.	13,5037	100,2106	282,3611	0,566
0,75	$\hat{Q}_{y,\text{cal},\alpha}$	−5,3477	1023,6924	1050,2431	0,826	−4,7339	443,0660	464,5896	0,926
	$\hat{Q}_{y,\text{HT},\alpha}$	−4,6352	3526,8202	3541,2514	0,938	−5,8890	1242,4858	1274,6812	0,940
	$\hat{Q}_{y,\text{ra},\alpha}$	−1,4390	980,5573	980,6669	0,994	−2,0070	555,5135	558,4305	1,000
	$\hat{Q}_{y,\text{diff},\alpha}$	−5,3988	1464,7867	1491,0041	0,996	−3,9008	744,1604	757,8881	1,000
	$\hat{Q}_{y,\text{CD},\alpha}$	49,3038	2753,8212	5179,1826	n.a.	49,4089	1488,9734	3927,2324	0,596

Tableau 6
Résultats des simulation de Monte Carlo pour l'échantillonnage de la population SLID982, sous un plan d'échantillonnage EAS. Nombre de répétitions fixé à $K = 500$

		$n = 100$				$n = 200$			
α	Estimateur	BR_{MC}	V_{MC}	EQM_{MC}	TC	BR_{MC}	V_{MC}	EQM_{MC}	TC
0,25	$\hat{Q}_{y,\text{cal},\alpha}$	0,1360	3,0390	3,0514	0,956	0,2331	1,6787	1,7297	0,934
	$\hat{Q}_{y,\text{HT},\alpha}$	−0,0596	3,6099	3,6062	0,946	0,0499	1,9277	1,9263	0,918
	$\hat{Q}_{y,\text{ra},\alpha}$	0,3067	6,8815	6,9618	0,970	0,0910	3,0743	3,0764	0,958
	$\hat{Q}_{y,\text{diff},\alpha}$	−0,0504	2,9691	2,9657	0,980	0,0198	1,6139	1,6111	0,952
	$\hat{Q}_{y,\text{CD},\alpha}$	1,1042	2,1180	3,3329	0,922	1,1392	1,2937	2,5888	0,826
0,5	$\hat{Q}_{y,\text{cal},\alpha}$	−0,4034	6,3364	6,4865	0,966	−0,1402	2,9940	3,0076	0,940
	$\hat{Q}_{y,\text{HT},\alpha}$	−0,4157	7,4589	7,6168	0,918	−0,1894	3,5865	3,6151	0,928
	$\hat{Q}_{y,\text{ra},\alpha}$	0,7015	41,8314	42,2399	0,958	0,2238	18,7005	18,7131	0,952
	$\hat{Q}_{y,\text{diff},\alpha}$	−0,4859	14,2083	14,4160	0,970	−0,2740	6,6184	6,6803	0,974
	$\hat{Q}_{y,\text{CD},\alpha}$	0,5702	3,5420	3,8601	0,952	0,6697	1,7559	2,2009	0,932
0,75	$\hat{Q}_{y,\text{cal},\alpha}$	−0,4164	12,4657	12,6142	0,952	−0,2384	5,9118	5,9568	0,950
	$\hat{Q}_{y,\text{HT},\alpha}$	−0,5913	12,5456	12,8701	0,930	−0,3519	6,5496	6,6603	0,926
	$\hat{Q}_{y,\text{ra},\alpha}$	0,7404	48,6836	49,1345	0,954	0,2967	18,5786	18,6294	0,966
	$\hat{Q}_{y,\text{diff},\alpha}$	0,3288	53,6456	53,6464	0,954	0,1841	21,7552	21,7456	0,966
	$\hat{Q}_{y,\text{CD},\alpha}$	0,5966	8,3416	8,6809	0,954	0,5413	4,3692	4,6535	0,936

Tableau 7
Résultats des simulations de Monte Carlo pour l'échantillonnage de la population SLID982, sous un plan d'échantillonnage PO et la première règle pour la construction des $\pi_k,\ k \in U$. Nombre de répliques fixé à $K = 500$

		$n = 100$				$n = 200$			
α	Estimateur	BR_{MC}	V_{MC}	EQM_{MC}	TC	BR_{MC}	V_{MC}	EQM_{MC}	TC
0,25	$\hat{Q}_{y,\text{cal},\alpha}$	0,1393	4,8403	4,8500	0,956	0,1603	2,8293	2,8493	0,922
	$\hat{Q}_{y,\text{HT},\alpha}$	−0,0477	5,8276	5,8182	0,934	−0,0227	3,5939	3,5872	0,924
	$\hat{Q}_{y,\text{ra},\alpha}$	0,1648	9,5171	9,5252	0,980	0,1263	4,8687	4,8749	0,972
	$\hat{Q}_{y,\text{diff},\alpha}$	−0,1418	4,7045	4,7152	0,960	−0,0464	2,9213	2,9176	0,936
	$\hat{Q}_{y,\text{CD},\alpha}$	3,9150	3,5279	18,8477	0,584	3,9114	1,9163	17,2112	0,194
0,5	$\hat{Q}_{y,\text{cal},\alpha}$	−0,1746	8,2437	8,2577	0,944	−0,2413	3,6477	3,6986	0,940
	$\hat{Q}_{y,\text{HT},\alpha}$	−0,2824	10,1117	10,1712	0,916	−0,3343	4,5023	4,6050	0,936
	$\hat{Q}_{y,\text{ra},\alpha}$	0,6558	50,4938	50,8228	0,944	0,4263	26,5883	26,7169	0,948
	$\hat{Q}_{y,\text{diff},\alpha}$	−0,5975	17,0315	17,3544	0,972	−0,3496	8,9060	9,0104	0,970
	$\hat{Q}_{y,\text{CD},\alpha}$	4,3173	4,4061	23,0363	0,484	4,0937	2,0711	18,8252	0,184
0,75	$\hat{Q}_{y,\text{cal},\alpha}$	−0,2229	12,1861	12,2114	0,942	−0,2113	6,5823	6,6138	0,952
	$\hat{Q}_{y,\text{HT},\alpha}$	−0,4150	14,2935	14,4371	0,934	−0,2786	7,6597	7,7220	0,934
	$\hat{Q}_{y,\text{ra},\alpha}$	0,7861	47,3844	47,9077	0,980	−0,1344	19,5992	19,5781	0,958
	$\hat{Q}_{y,\text{diff},\alpha}$	0,4347	52,3845	52,4687	0,972	−0,3409	23,8277	23,8962	0,958
	$\hat{Q}_{y,\text{CD},\alpha}$	4,4114	7,7023	27,1478	0,654	4,3549	4,1566	23,1136	0,392

Tableau 8
Résultats des simulations de Monte Carlo pour l'échantillonnage de la population SLID982, sous un plan d'échantillonnage PO et la deuxième règle pour la construction des $\pi_k,\ k \in U$. Nombre de répliques fixé à $K = 500$

		$n = 100$				$n = 200$			
α	Estimateur	BR_{MC}	V_{MC}	EQM_{MC}	TC	BR_{MC}	V_{MC}	EQM_{MC}	TC
0,25	$\hat{Q}_{y,\text{cal},\alpha}$	0,2392	3,4402	3,4906	0,962	0,1674	1,5214	1,5464	0,952
	$\hat{Q}_{y,\text{HT},\alpha}$	0,0267	4,0027	3,9954	0,940	−0,0370	1,6995	1,6975	0,958
	$\hat{Q}_{y,\text{ra},\alpha}$	0,4402	7,4350	7,6139	0,970	0,1850	3,0687	3,0968	0,978
	$\hat{Q}_{y,\text{diff},\alpha}$	0,0528	3,2842	3,2804	0,972	−0,0127	1,4718	1,4690	0,964
	$\hat{Q}_{y,\text{CD},\alpha}$	2,1458	3,0460	7,6444	0,876	1,9785	1,3010	5,2130	0,690
0,5	$\hat{Q}_{y,\text{cal},\alpha}$	−0,1410	6,5627	6,5695	0,942	−0,2850	2,9662	3,0415	0,954
	$\hat{Q}_{y,\text{HT},\alpha}$	−0,2133	7,6604	7,6906	0,928	−0,2876	3,6017	3,6772	0,926
	$\hat{Q}_{y,\text{ra},\alpha}$	1,0245	43,2773	44,2402	0,930	−0,3075	17,7242	17,7833	0,948
	$\hat{Q}_{y,\text{diff},\alpha}$	−0,1973	14,5261	14,5360	0,958	−0,6111	6,2988	6,6596	0,978
	$\hat{Q}_{y,\text{CD},\alpha}$	2,2140	4,5617	9,4543	0,834	1,8882	2,0393	5,6005	0,738
0,75	$\hat{Q}_{y,\text{cal},\alpha}$	−0,1985	12,6334	12,6476	0,952	−0,0022	5,6442	5,6329	0,966
	$\hat{Q}_{y,\text{HT},\alpha}$	−0,4012	13,5045	13,6384	0,922	−0,1078	6,2239	6,2231	0,934
	$\hat{Q}_{y,\text{ra},\alpha}$	0,7968	44,0650	44,6118	0,958	0,3727	19,1830	19,2836	0,960
	$\hat{Q}_{y,\text{diff},\alpha}$	0,4613	49,6620	49,7755	0,960	0,2340	22,1292	22,1397	0,966
	$\hat{Q}_{y,\text{CD},\alpha}$	2,6329	9,6723	16,5850	0,854	2,6729	4,1179	11,2541	0,738

5. Conclusion

Nous avons élaboré des estimateurs des quantiles fondés sur le paradigme du calage. Ces estimateurs sont particulièrement faciles à appliquer et à interpréter, puisqu'ils sont basés sur les pondérations et les contraintes de calage. De surcroît, il nécessite uniquement la connaissance des quantiles de population des variables auxiliaires, qui peuvent être vectorielles. Lorsqu'on adopte la métrique quadratique, il est possible d'obtenir, pour les poids calés, ainsi que pour les estimateurs de la variance, des expressions analytiques semblables à celles établies pour les estimateurs par calage des totaux. Sur le plan pratique, un aspect intéressant de la nouvelle méthodologie est que les estimateurs proposés sont faciles à calculer; il suffit de transformer les variables auxiliaires, puis d'utiliser les logiciels existants pour calculer les estimateurs par calage.

Au moyen d'une petite étude par simulation, nous avons comparé l'estimateur par calage des quantiles sous la métrique quadratique à d'autres estimateurs des quantiles fréquemment utilisés dans la littérature. L'estimateur proposé a donné de raisonnablement bons résultats dans nos expériences empiriques; sa performance était souvent meilleure, ou du moins semblable, à celle d'autres estimateurs utilisant la même quantité d'information. L'estimateur fondé sur un modèle, dans lequel est intégré beaucoup plus d'information au sujet des variables auxiliaires, semble préférable sous échantillonnage aléatoire simple et un modèle spécifié correctement, mais est surpassé par le nouvel estimateur lorsque les probabilités d'inclusion de premier ordre sont inégales. En général, l'estimateur proposé se compare favorablement aux estimateurs fondés sur le plan de sondage de Rao et coll. (1990).

Bien que nous nous soyons concentrés ici sur l'estimation des quantiles par calage sur des quantiles de population connus pour les variables auxiliaires, les estimateurs par calage peuvent être étendus à d'autres problèmes d'estimation importants présentant un intérêt dans le domaine des sondages. Les formulations de ces problèmes mènent toutes à des variables transformées différentes, que nous avons notées $\mathbf{a}_k$ dans le présent article. Par exemple, il est possible de formuler un problème de calage pour le coefficient de Gini bien connu, puis de montrer que la solution de ce problème de calage donnera des poids analogues à ceux dérivés ici; cependant, ces poids ne peuvent être déterminés que numériquement. Les travaux devront se poursuivre dans cette direction, afin d'étendre les estimateurs par calage à un cadre plus général, qui inclurait les totaux, les quantiles et les coefficients de Gini en tant que cas particuliers. Un autre domaine de recherche intéressant est celui du choix de l'estimateur de la fonction de répartition. Dans le présent article, nous avons préconisé un estimateur de cette fonction calculé en utilisant une interpolation linéaire. Nous pourrions aussi considérer un estimateur à noyau de la fonction de répartition (voir, par exemple, Altman et Léger (1995)). L'estimation de la densité par la méthode du noyau dans le cas d'enquêtes complexes est élaborée dans Bellhouse et Stafford (1999). Cela signifie que, dans $\hat{F}_{y,\,\mathrm{cal}}(t)$, la fonction $H_{y,\,s}(t, y_k)$ pourrait être remplacée par un noyau général, qui dépendrait cependant d'un paramètre supplémentaire, c'est-à-dire la fenêtre de l'estimateur à noyau. Notons que l'interpolation linéaire utilisée dans le présent article permet d'éviter le choix d'une fenêtre, qui est souvent une question délicate. L'élaboration d'un cadre général pour les problèmes de calage d'une fonctionnelle particulière et d'un estimateur à noyau de la fonction de répartition sera le sujet de travaux futurs.

Remerciements

Les auteurs remercient deux examinateurs anonymes de leurs suggestions et commentaires constructifs, qui leur ont permis d'améliorer considérablement l'article. Ils remercient également Raymond Chambers, Christian Léger, Éric Rancourt, Ulrich Rendtel et les participants à la XXXII^{e} assemblée de la Société statistique du Canada et à la Joint Statistical Meeting de 2004 de leurs discussions et commentaires. Le premier auteur a bénéficié d'une bourse de l'Office allemand d'échanges universitaires (DAAD) et le deuxième, de bourses du Conseil de recherches en sciences naturelles et en génie du Canada et du Fonds québécois de la recherche sur la nature et les technologies du Québec (Canada).

Bibliographie

Altman, N., et Léger, C. (1995), Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, 46, 195-214.

Bellhouse, D.R., et Stafford, J.E. (1999). Density estimation from complex surveys. *Statistica Sinica*, 9, 407-424.

Cassel, C.M., Särndal, C.-E. et Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrika*, 63, 615-620.

Chambers, R.L., Dorfman, A.H. et Hall, P. (1992). Properties of estimators of finite population distribution functions. *Biometrika*, 79, 577-582.

Chambers, R.L., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Chen, J., et Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective use of auxiliary information. *Biometrika*, 80, 107-116.

Deville, J.-C. (1988). Estimation linéaire et redressement sur information auxiliaire d'enquêtes par sondage. Dans *Essais en l'Honneur d'Edmont Malinvaud*, (Éds, A. Monfort, et J.J. Laffond), *Economica*, Paris, 915-929.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Dorfman, A.H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function. *Australian Journal of Statistics*, 35, 29-41.

Harms, T. (2003), Extensions of the calibration approach: calibration of distribution functions and its link to small area estimators, Chintex working paper #13, Federal Statistical Office, Allemagne.

Kovačević, M.S. (1997). Calibration estimation of cumulative distribution and quantile functions from survey data. *Proceedings of the Survey Methods Section*, *Statistical Society of Canada*, 139-144.

Kovar, J.G., Rao, J.N.K. et Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 16 (Supp.), 25-45.

Kuk, A.Y.C. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. *Biometrika*, 75, 97-103.

Kuk, A.Y.C., et Mak, T.K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society*, Séries B (Méthodologique), 51, 261-269.

Meeden, G. (1995). Estimation de la médiane à l'aide d'informations supplémentaires. *Techniques d'enquête*, 21, 81-88.

Rao, J.N.K., Kovar, J.G. et Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.

Ren, R. (2002). Estimation de la fonction de répartition et des fractiles d'une population finie. *Actes des journées de méthodologie statistique*, *INSEE Méthodes*, Tome 1, 100, 263-289.

Ren, R., et Deville, J.C. (2000). Une généralisation du calage: calage sur les rangs et le calage sur les moments, II^{ème} Colloque Francophone sur les Sondages. Bruxelles.

Rueda, M.M., Arcos A. et Martínez, M.D. (2003). Difference estimators of quantiles in finite populations. *Test*, 12, 481-496.

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

Singh, A.C., et Mohl, C.A. (1996). Comprendre les estimateurs de calage dans les enquêtes par échantillonnage. *Techniques* d'enquête, 22, 107-116.

Thompson, M. (1997). *Theory of Sample Surveys*. Chapman & Hall, New York.

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 625-646.

Wu, C., et Sitter, R.R. (2001). Variance estimation for the finite population distribution function with complete auxiliary information. *The Canadian Journal of Statistics*, 29, 289-308.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, juin 2006
Vol. 32, N° 1, pp. 59-71
Statistique Canada, N° 12-001-XIF au catalogue

Une approche fondée sur un modèle de non-réponse à des fins d'inférence en présence d'imputation pour des données d'enquête manquantes

David Haziza et Jon N.K. Rao [1]

Résumé

En présence de non-réponse partielle, deux approches sont généralement utilisées à des fins d'inférence des paramètres d'intérêt. La première repose sur l'hypothèse que la réponse est uniforme dans les classes d'imputation, tandis que la seconde s'appuie sur l'hypothèse que la réponse est ignorable, mais utilise un modèle pour la variable d'intérêt comme fondement de l'inférence. Dans le présent article, nous proposons une troisième approche qui se fonde sur l'hypothèse d'un mécanisme de réponse précisé ignorable sans que doive être spécifié un modèle de la variable d'intérêt. Dans ce cas, nous montrons comment obtenir des valeurs imputées qui mènent à des estimateurs d'un total approximativement sans biais sous l'approche proposée, ainsi que sous la deuxième des approches susmentionnées. Nous obtenons aussi des estimateurs de la variance des estimateurs imputés qui sont approximativement sans biais en suivant une approche proposée par Fay (1991) dans laquelle sont inversés l'ordre de l'échantillonnage et de la réponse. Enfin, nous effectuons des études par simulation afin d'étudier les propriétés des méthodes dans le cas d'échantillons finis, en termes de biais et d'erreur quadratique moyenne.

Mots clés : Approche basée sur un modèle de non-réponse; approche basée sur un modèle d'imputation; estimateur corrigé pour le biais; estimation de la variance; imputation par la régression aléatoire; imputation par la régression déterministe; non-réponse partielle.

1. Introduction

Il y a non-réponse partielle lors d'une enquête quand une unité échantillonnée participe à l'enquête, mais omet de répondre à une ou à plusieurs variables (Brick et Kalton 1996). Elle est généralement traitée par une forme ou l'autre d'imputation qui consiste à « boucher les trous » dues aux valeurs manquantes pour chaque variable. L'imputation peut effectivement réduire le biais, à condition que l'on dispose d'information auxiliaire appropriée pour toutes les unités échantillonnées et qu'on l'intègre correctement dans le modèle d'imputation et/ou dans le modèle de non-réponse.

L'imputation offre, entre autres, les caractéristiques souhaitables suivantes : i) elle mène à la création d'un fichier de données complet et ii) elle permet d'utiliser les mêmes poids de sondage pour toutes les variables, ce qui assure que les résultats obtenus, après diverses analyses de l'ensemble complet de données, soient cohérents, contrairement aux résultats d'analyses réalisées sur un ensemble de données incomplet. Cependant, l'imputation présente aussi, entre autres, les difficultés suivantes : a) l'imputation marginale pour chaque variable fausse la relation entre les variables et b) traiter les valeurs imputées comme s'il s'agissait de valeurs réelles peut entraîner une sous-estimation importante de la variance des estimateurs imputés, particulièrement quand le taux de non-réponse est appréciable. Des méthodes permettant de résoudre les problèmes (a) et (b) ont été proposées dans la littérature.

Dans le présent article, nous nous concentrons sur l'imputation marginale qui est utilisée communément dans de nombreuses enquêtes. Pour commencer, nous examinons l'imputation par la régression linéaire déterministe qui comprend les cas particuliers de l'imputation par la moyenne et de l'imputation par le ratio. Selon cette méthode, une valeur manquante est remplacée par la valeur prédite obtenue en ajustant un modèle de régression linéaire au moyen des valeurs fournies par les répondants et de celles des variables auxiliaires recueillies pour toutes les unités échantillonnées. Nous examinons aussi le cas de l'imputation par la régression aléatoire qui peut être considérée comme une imputation par la régression déterministe à laquelle est ajouté un résidu aléatoire. Elle comprend le cas particulier de l'imputation hot-deck aléatoire.

Soit U une population finie de taille éventuellement inconnue N. L'objectif est d'estimer le total de population $Y = \sum_U y_i$ d'une qvariable y lorsque l'on a utilisé l'imputation pour traiter la non-réponse pour les valeurs y_i cette variable. Pour être concis, nous utiliserons la notation $\sum_A$ pour $\sum_{i \in A}$, où $A \subseteq U$. Supposons que l'on sélectionne un échantillon probabiliste, s, de taille n conformément à un plan spécifié $p(s)$ à partir de U. Sous des conditions de réponse complète à la variable y, un estimateur de Y sans biais par rapport au plan est donné par l'estimateur d'Horvitz-Thompson bien connu

$$\hat{Y} = \sum_s w_i y_i, \qquad (1)$$

1. David Haziza, Division des méthodes d'enquête auprès des entreprises, Statistique Canada, Ottawa (Ontario) Canada, K1A 0T6; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa (Ontario), Canada. K1S 5B6.

avec les poids d'échantillonnage (ou de sondage) $w_i = 1/\pi_i$, où π_i dénote la probabilité d'inclusion de l'unité désigne i dans l'échantillon $s, i = 1, \ldots, N$. Rao (2005) a suggéré que l'on devrait appeler (1) l'estimateur de Narain-Horvitz-Thompson (NHT) en reconnaissance du fait que Narain (1951) a également découvert (1) indépendamment d'Horvitz et de Thompson (1952).

En présence de non-réponse à la variable y, nous utilisons l'imputation et définissons un estimateur imputé $\hat{Y}_I$ de la forme

$$\hat{Y}_I = \sum_s w_i a_i y_i + \sum_s w_i (1 - a_i) y_i^* = \sum_s w_i \tilde{y}_i, \qquad (2)$$

où y_i^* représente la valeur imputée pour remplacer la valeur manquante y_i, a_i représente l'indicateur de réponse égal à 1 si l'unité i répond à la variable y et égale à 0 autrement, et $\tilde{y}_i = a_i y_i + (1 - a_i) y_i^*$. L'estimateur imputé (2) peut être obtenu à partir du fichier de données imputé contenant les poids de sondage w_i et les $\tilde{y}_i$ uniquement, sans que l'on connaisse les indicateurs de réponse a_i. Cependant, ces derniers seront nécessaires pour estimer la variance. Soit $p_i = P(a_i = 1)$ la probabilité de réponse de l'unité i à la variable y. Dans le présent article, nous supposons que les unités répondent indépendamment les unes des autres, c'est-à-dire que $p_{ij} = P(a_i = 1, a_j = 1) = p_i p_j$ si $i \neq j$.

Comme toute méthode de remplacement des données manquantes, l'imputation nécessite certaines hypothèses au sujet du mécanisme de réponse et/ou du modèle d'imputation. En présence de données imputées, deux approches sont généralement utilisées pour mener des inférences au sujet des totaux, des moyennes et d'autres paramètres d'intérêt, à savoir i) l'approche du modèle d'imputation (MI) et ii) celle du modèle de non-réponse (MN). L'approche (i) est également appelée approche assistée par un modèle (Särndal 1992) et l'approche (ii), approche fondée sur le plan de sondage (Shao et Steel 1999). L'approche MN est basée sur le partitionnement de la population U en J classes d'imputation, suivi de l'imputation des valeurs de y correspondant aux non-répondants compris dans chaque classe en utilisant les valeurs de y des répondants compris dans la même classe comme donneurs, indépendamment dans chacune des J classes. L'hypothèse suivante est formulée :

Hypothèse MN : La probabilité de réponse à une variable d'intérêt est constante dans les classes d'imputation. Autrement dit, $p_i = p_v$, disons, où l'indice inférieur v désigne la classe d'imputation.

Dans l'approche MN, des hypothèses explicites sont formulées au sujet du mécanisme de réponse. Il s'ensuit que l'inférence sous l'hypothèse MN est faite sous les conditions d'échantillonnage répété et d'un mécanisme de réponse uniforme dans les classes. L'approche MN a été étudiée par Rao (1990, 1996), Rao et Shao (1992), Rao et Sitter (1995) et Shao et Steel (1999), entre autres. Pour simplifier, nous supposons qu'il n'y a qu'une seule classe d'imputation, de sorte que $p_i = p$ sous l'hypothèse MN.

L'approche MI est fondée sur l'hypothèse suivante :

Hypothèse MI : Les valeurs d'une variable manquent au hasard (MAR pour *missing at random)* au sens où la probabilité de réponse ne dépend pas de la valeur de la variable qui est imputée, mais des variables auxiliaires utilisées pour l'imputation. En outre, une hypothèse est émise quant au modèle qui génère les valeurs y_i de la variable.

Dans l'approche MI, des hypothèses explicites au sujet de la distribution des valeurs y_i de la variable sont formulées au moyen d'un modèle appelé « modèle d'imputation ». Il s'ensuit que l'inférence sous l'hypothèse MI est faite sous les conditions d'échantillonnage répété et du modèle hypothétique qui génère la population finie de valeurs de y et de non-répondants à la variable y. Contrairement à l'approche MN, le mécanisme de réponse sous-jacent n'est pas spécifié, à part l'hypothèse MAR. L'hypothèse MI concernant le mécanisme de réponse est nettement plus faible que l'hypothèse MN de réponse uniforme dans les classes, mais les inférences sous l'hypothèse MI dépendent du modèle de population hypothétique. L'approche MI a été étudiée, entre autres, par Särndal (1992), Deville et Särndal (1994), ainsi que Shao et Steel (1999).

Sous l'imputation par la régression linéaire, l'approche MI s'appuie sur le modèle d'imputation par la régression linéaire hypothétique suivant :

$$E_m(y_i) = \mathbf{z_i'}\boldsymbol{\gamma},\ V_m(y_i) = \sigma_i^2 = \sigma^2(\boldsymbol{\lambda}'\mathbf{z_i}),$$
$$\mathrm{Cov}_m(y_i, y_j) = 0 \text{ si } i \neq j, \qquad (3)$$

où $\boldsymbol{\gamma}$ est un vecteur de dimension k de paramètres inconnus, $\mathbf{z_i}$ est un vecteur de dimension k de variables auxiliaires disponibles pour toutes $i \in s$, $\boldsymbol{\lambda}$ est un vecteur de dimension k de constantes spécifiées, σ^2 est un paramètre inconnu et E_m, V_m et Cov_m représentent, respectivement, les opérateurs d'espérance, de variance et de covariance par rapport au modèle d'imputation. La contrainte $\sigma_i^2 = \sigma^2(\boldsymbol{\lambda}'\mathbf{z_i})$ ne restreint pas sévèrement la gamme de modèles d'imputation.

Dans le présent article, nous proposons une troisième approche, appelée approche du modèle de non-réponse généralisée (MNG), qui est fondée sur l'hypothèse suivante :

Hypothèse MNG : Les valeurs de la variable manquent au hasard (MAR) et la probabilité de réponse est spécifiée sous forme d'une fonction des variables auxiliaires, $\mathbf{u_i}$, observées sur toutes les unités de l'échantillon, et de paramètres inconnus $\boldsymbol{\eta}$.

Nous supposons ici que la probabilité de réponse, p_i, de l'unité i est liée à un vecteur de dimension l de variables auxiliaires $\mathbf{u_i}$ conformément à un modèle logistique de sorte que

$$p_i = f(\mathbf{u_i}'\boldsymbol{\eta}) = \exp(\mathbf{u_i}'\boldsymbol{\eta}) / \exp(1 + \mathbf{u_i}'\boldsymbol{\eta}), \qquad (4)$$

où $\boldsymbol{\eta}$ est le vecteur de dimension l de paramètres du modèle. Le modèle (4) est le modèle de non-réponse hypothétique. Il peut être validé à partir des valeurs a_i et $\mathbf{u_i}$ pour $i \in s$. Notons que a_i et $\mathbf{u_i}$ sont particuliers à la variable d'intérêt. De plus, notons que l'hypothèse MN est un cas particulier de l'hypothèse MNG. Comme dans l'approche MN, des hypothèses explicites au sujet du mécanisme de réponse sont formulées et l'inférence sous l'hypothèse MNG est faite sous les conditions d'échantillonnage répété et du mécanisme de réponse hypothétique.

Rappelons que l'imputation est utilisée en vue de réduire le biais de non-réponse, en supposant que les variables auxiliaires disponibles permettent d'expliquer la variable pour laquelle des valeurs doivent être imputées et/ou la probabilité de réponse à la variable. Donc, en pratique, le choix de l'approche (MI ou MNG) devrait être dicté par la qualité du modèle d'imputation et du modèle de non-réponse. Le choix entre la modélisation de la probabilité de réponse à la variable et celle de la variable d'intérêt dépendra de la confiance que l'on a dans chacun des modèles. S'il peut paraître intuitivement plus séduisant de modéliser la variable d'intérêt, il existe, en pratique, des cas où il pourrait être plus facile de modéliser la probabilité de réponse (approche MNG). Par exemple, à Statistique Canada, l'Enquête sur les dépenses en immobilisations produit des données sur l'investissement fait au Canada, dans tous les types d'industries. Dans cette enquête, deux variables d'intérêt importantes sont les dépenses d'immobilisations en constructions neuves (CC) et les dépenses d'immobilisations en machines et matériel neufs (CM). Durant une année donnée, un grand nombre d'entreprises ne font aucun investissement en constructions neuves ni en machines neuves. Par conséquent, le fichier de données d'échantillon contient un grand nombre de valeurs nulles pour les variables CC et CM. Le cas échéant, la modélisation de la variable d'intérêt (CC ou CM) peut s'avérer difficile.

En général, les poids de sondage sont utilisés dans l'imputation par la régression linéaire. L'estimateur imputé ainsi obtenu d'un total de population est « robuste » au sens de l'absence de biais approximatif sous l'hypothèse MN ou sous l'hypothèse MI. Toutefois, il contient généralement un biais sous l'hypothèse MNG. Dans le présent article, nous proposons une nouvelle méthode d'imputation par la régression linéaire qui est robuste au sens où elle mène à des estimateurs approximativement sans biais sous l'hypothèse MNG ou sous l'hypothèse MI.

À la section 2, nous décrivons l'élaboration d'une nouvelle méthode d'imputation par la régression linéaire déterministe, ainsi qu'une imputation par la régression linéaire aléatoire, et nous démontrons la propriété de robustesse dans le cas d'un total de population Y. À la section 3, nous présentons les résultats d'une étude par simulation des propriétés dans le cas d'échantillons finis de l'estimateur imputé sous la nouvelle méthode d'imputation. À la section 4, nous développons les estimateurs de la variance, en utilisant l'approche « renversé » de Fay (1991) dans laquelle l'ordre de l'échantillonnage et de la réponse est inversé :

Population → recensement avec non-répondants → échantillon avec non-répondants.

Nous présentons aussi les résultats des simulations concernant les estimateurs de la variance. Enfin, à la section 5, nous examinons le cas des moyennes de domaine.

2. Estimation d'un total

À la présente section, nous étudions le biais de l'estimateur imputé $\hat{Y}_I$. L'erreur totale, $\hat{Y}_I - Y$, peut être décomposée comme il suit :

$$\hat{Y}_I - Y = (\hat{Y} - Y) + (\hat{Y}_I - \hat{Y}). \qquad (5)$$

Dans (5), le terme $\hat{Y} - Y$ est appelé erreur d'échantillonnage, tandis que le terme $\hat{Y}_I - \hat{Y}$ est appelé erreur due à la non-réponse/imputation. Soulignons qu'il n'y a pas d'erreur due à l'imputation dans le cas d'imputation déterministe. Puisque l'erreur d'échantillonnage ne dépend ni de la non-réponse ni de la méthode d'imputation, nous nous concentrons sur l'erreur due à la non-réponse/imputation $\hat{Y}_I - \hat{Y}$ et évaluons ses propriétés étant donné l'échantillon s. Sous l'approche MN ou MNG, nous définissons le biais de non-réponse conditionnel comme étant $E_r(\hat{Y}_I - \hat{Y} \mid s)$, où $E_r(.)$ représente l'espérance par rapport au mécanisme de réponse. Sous l'approche MI, le biais de non-réponse conditionnel est défini comme étant $E_r E_m(\hat{Y}_I - \hat{Y} \mid s)$ sous l'hypothèse MAR.

2.1 Imputation par la régression déterministe

L'imputation par la régression déterministe consiste à utiliser les valeurs imputées

$$y_i^* = \mathbf{z_i}'\hat{\boldsymbol{\gamma}}_r \qquad (6)$$

pour remplacer les valeurs manquantes y_i, où

$$\hat{\boldsymbol{\gamma}}_r = \left(\sum_s w_i a_i \mathbf{z_i}\mathbf{z_i}' / (\boldsymbol{\lambda}'\mathbf{z_i}) \right)^{-1} \sum_s w_i a_i \mathbf{z_i} y_i / (\boldsymbol{\lambda}'\mathbf{z_i}) \qquad (7)$$

est l'estimateur par les moindres carrés pondérés de $\boldsymbol{\gamma}$ sous le modèle (3), basé sur les unités échantillonnées répondant à la question y. Partant de (6), l'estimateur imputé (2) peut s'écrire sous la forme

$$\hat{Y}_I = \hat{Y}_r + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_\mathbf{r})'\hat{\boldsymbol{\gamma}}_\mathbf{r}, \qquad (8)$$

où $\hat{Y}_r = \sum_s w_i a_i y_i$, $\hat{\mathbf{Z}} = \sum_s w_i \mathbf{z_i}$ et $\hat{\mathbf{Z}}_\mathbf{r} = \sum_s w_i a_i \mathbf{z_i}$. Notons que l'estimateur imputé (8) est similaire à un estimateur par la régression dans le cas de l'échantillonnage à deux phases.

Sous l'hypothèse MN, $E_r(a_i \mid s) = p$ et le biais de non-réponse conditionnel, $E_r(\hat{Y}_I - \hat{Y} \mid s)$, est approximativement égal à 0. En outre, sous l'hypothèse MI et le modèle de régression (3), le biais de non-réponse conditionnel $E_r E_m(\hat{Y}_I - \hat{Y} \mid s)$, est nul. Cependant, sous l'hypothèse MNG, le biais de non-réponse conditionnel est donné par

$$E_r(\hat{Y}_I - \hat{Y} \mid s) \approx -\sum_s w_i(1-p_i)(y_i - \mathbf{z_i}'\hat{\boldsymbol{\gamma}}_p) \equiv \mathrm{B}(\hat{Y}_I \mid s), \qquad (9)$$

où

$$\hat{\boldsymbol{\gamma}}_p = \left(\sum_s w_i p_i \mathbf{z_i}\mathbf{z_i}'/(\boldsymbol{\lambda}'\mathbf{z_i})\right)^{-1} \sum_s w_i p_i \mathbf{z_i} y_i/(\boldsymbol{\lambda}'\mathbf{z_i}). \qquad (10)$$

Ce résultat découle du fait que, sous l'hypothèse MNG, $E_r(a_i \mid s) = p_i$. Donc, le choix des valeurs imputées (6) n'est, en général, pas approprié sous cette hypothèse. Pour le cas particulier de l'hypothèse MN avec $p_i = p$, le dernier terme de (9) disparaît, en notant que $(\sum_s w_i \mathbf{z_i}')\hat{\boldsymbol{\gamma}}_p = \boldsymbol{\lambda}'(\sum_s w_i \mathbf{z_i}\mathbf{z_i}'/(\boldsymbol{\lambda}'\mathbf{z_i}))\hat{\boldsymbol{\gamma}}_p = \boldsymbol{\lambda}'(\sum_s w_i \mathbf{z_i} y_i/(\boldsymbol{\lambda}'\mathbf{z_i})) = \sum_s w_i y_i$.

2.2 Estimateur corrigé pour le biais

Nous supposons pour le moment que les probabilités de réponse p_i sont connues. Une approche naturelle en vue d'éliminer le biais présent dans $\hat{Y}_I$ sous l'hypothèse MNG consiste à considérer un estimateur corrigé pour le biais de la forme

$$\hat{Y}_I^a = \hat{Y}_I - \hat{B}(\hat{Y}_I \mid s), \qquad (11)$$

où $\hat{B}(\hat{Y}_I \mid s)$ est un estimateur de $\mathrm{B}(\hat{Y}_I \mid s)$:

$$\hat{B}(\hat{Y}_I \mid s) = -\sum_s w_i a_i \frac{(1-p_i)}{p_i}(y_i - \mathbf{z_i}'\hat{\boldsymbol{\gamma}}_\mathbf{r}). \qquad (12)$$

Soulignons que $E_r[\hat{B}(\hat{Y}_I \mid s) \mid s] \approx \mathrm{B}(\hat{Y}_I \mid s)$ sous l'hypothèse MNG. En introduisant (12) dans (11) par substitution, nous obtenons un estimateur corrigé pour le biais de la forme

$$\hat{Y}_I^a = \sum_s \frac{w_i}{p_i} a_i y_i + \left(\sum_s w_i \mathbf{z_i}' - \sum_s \frac{w_i}{p_i} a_i \mathbf{z_i}'\right)\hat{\boldsymbol{\gamma}}_\mathbf{r}. \qquad (13)$$

Notons que (13) est également sous la forme d'un estimateur par la régression dans le cas de l'échantillonnage à deux phases.

En pratique, les probabilités de réponse p_i sont inconnues. Supposons que nous puissions obtenir des estimateurs $\hat{p}_i$ de p_i par modélisation de p_i conformément au modèle de non-réponse (4). Alors, nous obtenons un estimateur corrigé pour le biais en remplaçant p_i par $\hat{p}_i$ dans (13). Cet estimateur est également approximativement conditionnellement sans biais sous l'hypothèse MI. Donc, l'estimateur corrigé pour le biais (13) est robuste au sens de sa validité sous l'hypothèse MNG. Cependant, contrairement à l'estimateur imputé $\hat{Y}_I$ donné par (2), l'estimateur corrigé pour le biais $\hat{Y}_I^a$ ne peut être calculé sans que l'on connaisse les identificateurs de réponse, a_i, et les probabilités de réponse estimées, $\hat{p}_i$. Par conséquent, pour pouvoir obtenir $\hat{Y}_I^a$, les indicateurs de réponse ainsi que les probabilités estimées de réponse doivent être fournis avec le fichier de données imputé, ce qui n'est pas toujours le cas en pratique. Cet inconvénient de $\hat{Y}_I^a$ peut être éliminé en utilisant la nouvelle méthode d'imputation, décrite à la section 2.3, qui mène à un estimateur approximativement sans biais sous l'hypothèse MNG ou sous l'hypothèse MI sans que l'on connaisse a_i et $\hat{p}_i$ dans le fichier de données imputé. Cependant, il est nécessaire d'avoir accès aux valeurs de a_i et de $\hat{p}_i$ pour estimer la variance.

2.3 Imputation par la régression déterministe modifiée

Nous supposons pour l'instant que les probabilités de réponse p_i sont connues. Nous utilisons alors les valeurs imputées

$$y_i^* = \mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_\boldsymbol{s} \qquad (14)$$

pour remplacer les valeurs manquantes y_i et obtenons la forme de $\tilde{\boldsymbol{\gamma}}_\boldsymbol{s}$ qui mène à un estimateur approximativement sans biais sous l'hypothèse MNG.

2.3.1 Estimateur approximativement sans biais

Le lemme qui suit donne la forme de $\tilde{\boldsymbol{\gamma}}_\boldsymbol{s}$ qui mène à un estimateur approximativement sans biais sous l'hypothèse MNG.

Lemme 1 : *Sous l'hypothèse MNG, le choix de* $\tilde{\boldsymbol{\gamma}}_\boldsymbol{s}$ *qui mène à* $E_r(\hat{Y}_I - \hat{Y} \mid s) = 0$ *est donné par*

$$\tilde{\boldsymbol{\gamma}}_{\boldsymbol{s},N} = \left[\sum_s w_i(1-p_i)\mathbf{z_i}\mathbf{z_i}'/(\boldsymbol{\lambda}'\mathbf{z_i})\right]^{-1} \sum_s w_i(1-p_i)\mathbf{z_i} y_i/(\boldsymbol{\lambda}'\mathbf{z_i}). \qquad (15)$$

Preuve : Le biais de non-réponse conditionnel de $\hat{Y}_I$ avec $y_i^* = \mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_\boldsymbol{s}$ sous l'hypothèse MNG est donné par

$$E_r(\hat{Y}_I - \hat{Y} \mid s) = -\sum_s w_i(1-p_i)(y_i - \mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_\boldsymbol{s}).$$

En notant que $(\boldsymbol{\lambda}'\mathbf{z_i})/(\boldsymbol{\lambda}'\mathbf{z_i}) = 1$, il s'ensuit que $E_r(\hat{Y}_I - \hat{Y} \mid s) = 0$ si $\tilde{\boldsymbol{\gamma}}_\boldsymbol{s}$ satisfait

$$\boldsymbol{\lambda}'\left[\sum_s w_i(1-p_i)\mathbf{z_i}(y_i - \mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_\boldsymbol{s})/(\boldsymbol{\lambda}'\mathbf{z_i})\right] = 0. \qquad (16)$$

Le choix $\tilde{\boldsymbol{\gamma}}_\boldsymbol{s} = \tilde{\boldsymbol{\gamma}}_{\boldsymbol{s},N}$ satisfait (16).

Notons que $\tilde{\boldsymbol{\gamma}}_{s,N}$ est inconnu, puisque les valeurs de y ne sont observées que pour $i \in s_r$ et que les probabilités de réponse p_i sont inconnues. Un estimateur de $\tilde{\boldsymbol{\gamma}}_{s,N}$ fondé sur les unités répondantes et les probabilités de réponse estimées $\hat{p}_i$ est donné par

$$\tilde{\boldsymbol{\gamma}}_r = \left[\sum_s w_i a_i \frac{(1-\hat{p}_i)}{\hat{p}_i}\mathbf{z}_\mathbf{i}\mathbf{z}_\mathbf{i}'/(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i})\right]^{-1} \sum_s w_i a_i \frac{(1-\hat{p}_i)}{\hat{p}_i}\mathbf{z}_\mathbf{i} y_i/(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i}). \quad (17)$$

Nous avons $\mathrm{E}_\mathrm{r}(\tilde{\boldsymbol{\gamma}}_r \mid \mathrm{s}) \approx \tilde{\boldsymbol{\gamma}}_{s,N,}$ de sorte que $\tilde{\boldsymbol{\gamma}}_r$ est conditionnellement approximativement sans biais pour $\tilde{\boldsymbol{\gamma}}_{s,N}$ sous l'hypothèse MNG. Donc, en utilisant les valeurs imputées

$$y_i^* = \mathbf{z}_\mathbf{i}'\hat{\boldsymbol{\gamma}}_r \quad (18)$$

dans (2) avec $\tilde{\boldsymbol{\gamma}}_r$ donné par (17), nous obtenons un estimateur approximativement sans biais du total Y sous l'hypothèse MNG. Notons que $\tilde{\boldsymbol{\gamma}}_r$ est un estimateur par les moindres carrés pondérés de $\boldsymbol{\gamma}$ par rapport à un nouvel ensemble de poids, $\tilde{w}_i/(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i})$, où $\tilde{w}_i = w_i(1-\hat{p}_i)/\hat{p}_i$. Donc, la procédure accroît les poids de sondage w_i des unités pour lesquelles $\hat{p}_i < 1/2$ et diminue ceux des unités pour lesquelles $\hat{p}_i > 1/2$. L'estimateur imputé peut être appliqué au fichier de données imputé contenant les poids d'échantillonnage w_i et les $\tilde{y}_i$ uniquement; les indicateurs de réponse a_i et les probabilités de réponse estimées $\hat{p}_i$ ne sont pas requis. Toutefois, il est nécessaire de connaître a_i et $\hat{p}_i$ pour estimer la variance. Notons que le producteur du fichier de données imputé utilise l'information concernant a_i et $\mathbf{u}_\mathbf{i}$ pour ajuster le modèle de réponse (4) et générer les valeurs imputées y_i^* données par (18).

L'utilisation des valeurs imputées (18) mène également à un estimateur approximativement sans biais de Y sous l'hypothèse MI. Premièrement, sous le modèle de régression (3), en notant que $E_m(y_i \mid s) = \mathbf{z}_\mathbf{i}'\boldsymbol{\gamma}$ et $E_m(\tilde{\boldsymbol{\gamma}}_r \mid s) = \boldsymbol{\gamma}$, nous avons $E_m(\hat{Y}_I - \hat{Y} \mid s) = 0$ et $E_r E_m(\hat{Y}_I - \hat{Y} \mid s) = 0$ sans spécifier le mécanisme de non-réponse MAR sous-jacent. Donc, l'utilisation des valeurs imputées (18) mène à un estimateur imputé robuste au sens de sa validité sous les deux approches. Enfin, il est intéressant de souligner que les valeurs imputées (18) peuvent également être obtenues par la méthode d'imputation par calage (Beaumont 2005). Cette dernière consiste à trouver des valeurs imputées finales aussi proche que possible des valeurs imputées originales conformément à une fonction de distance, sous les contraintes de calage.

Deux cas particuliers de l'imputation par la régression modifiée (18) présentent un intérêt, à savoir i) l'imputation par le ratio modifiée avec $\mathbf{z}_\mathbf{i} = z_i$ et $\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i} = \mathrm{z_i}$ et ii) l'imputation par la moyenne modifiée avec $\mathbf{z}_\mathbf{i} = 1$ et $\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i} = 1$. Dans le cas (i), les valeurs imputées (18) se réduisent à

$$y_i^* = \frac{\sum_s \tilde{w}_i a_i y_i}{\sum_s \tilde{w}_i a_i z_i} z_i. \quad (19)$$

Dans le cas (ii), les valeurs imputées (18) se réduisent à

$$y_i^* = \frac{\sum_s \tilde{w}_i a_i y_i}{\sum_s \tilde{w}_i a_i}. \quad (20)$$

Sous l'hypothèse de réponse uniforme $p_i = p$, les valeurs imputées (19) et (20) se réduisent à $(\sum_s w_i a_i y_i / \sum_s w_i a_i z_i) z_i$ et à $\bar{y}_r = \sum_s w_i a_i y_i / \sum_s w_i a_i$ respectivement, qui sont les valeurs usuelles que les praticiens d'enquête utilisent pour l'imputation par le ratio et par la moyenne (Rao et Sitter 1995).

2.3.2 Choix optimal de $\tilde{\boldsymbol{\gamma}}_s$

Nous examinons maintenant la question du choix « optimal » de $\tilde{\boldsymbol{\gamma}}_s$ par minimisation de l'erreur quadratique moyenne conditionnelle de l'estimateur imputé $\hat{Y}_I$ avec $y_i^* = \mathbf{z}_\mathbf{i}'\tilde{\boldsymbol{\gamma}}_s$. L'erreur quadratique moyenne conditionnelle de l'estimateur $\hat{Y}_I$ est donnée par

$$\begin{aligned}\mathrm{EQM}_r(\hat{Y}_I \mid s) &= V_r(\hat{Y}_I \mid s) + [\mathrm{Biais}(\hat{Y}_I \mid s)]^2 \\ &= \sum_s w_i^2 p_i(1-p_i)(y_i - \mathbf{z}_\mathbf{i}'\tilde{\boldsymbol{\gamma}}_s)^2 \\ &\quad + \left[\sum_s w_i(1-p_i)(y_i - \mathbf{z}_\mathbf{i}'\tilde{\boldsymbol{\gamma}}_s)\right]^2, \end{aligned} \quad (21)$$

où $V_r(.\mid s)$ représente la variance due à la non-réponse conditionnelle par rapport au mécanisme de réponse, étant donné l'échantillon s. Nous recherchons la valeur de $\tilde{\boldsymbol{\gamma}}_s$ qui minimise $\mathrm{EQM}_r(\hat{Y}_I \mid s)$.

Le choix optimal, $\tilde{\boldsymbol{\gamma}}_\mathbf{opt}$, de $\tilde{\boldsymbol{\gamma}}_s$ est complexe, mais, dans le cas particulier de l'imputation par le ratio, $\tilde{\boldsymbol{\gamma}}_\mathbf{opt}$ se réduit à

$$\tilde{\boldsymbol{\gamma}}_\mathbf{opt} = \frac{\sum_s w_i(1-p_i)y_i \sum_s w_i(1-p_i)z_i + \sum_s w_i^2 p_i(1-p_i)y_i z_i}{\left[\sum_s w_i(1-p_i)z_i\right]^2 + \sum_s w_i^2 p_i(1-p_i)z_i^2}. \quad (22)$$

Supposons que les poids d'échantillonnage w_i satisfont $\max(n/Nw_i) = O(1)$ et qu'il existe une constante positive C telle que $C < p_i$. Alors,

$$\begin{aligned}\tilde{\boldsymbol{\gamma}}_\mathbf{opt} &= \frac{\sum_s w_i(1-p_i)y_i}{\sum_s w_i(1-p_i)z_i} + O\left(\frac{1}{n}\right) \\ &= \tilde{\boldsymbol{\gamma}}_{s,N} + O\left(\frac{1}{n}\right).\end{aligned}$$

Donc, pour les grandes tailles d'échantillon, le choix $\tilde{\boldsymbol{\gamma}}_{s,N}$ est presque optimal pour l'imputation par le ratio. De même, $\tilde{\boldsymbol{\gamma}}_{s,N}$ est presque optimal pour l'imputation par la moyenne, qui est un cas particulier de l'imputation par le ratio.

2.4 Imputation par la régression aléatoire

L'imputation aléatoire peut être considérée comme une imputation déterministe avec ajout d'un bruit aléatoire. Soit s_r et s_m les ensembles de répondants et de non-répondants dans l'échantillon, respectivement, et soit $e_j = (y_j - \mathbf{z}_j'\hat{\boldsymbol{\gamma}}_r)/(\boldsymbol{\lambda}'\mathbf{z}_j)^{1/2}$ les résidus standardisés pour les répondants $j \in s_r$ sous l'imputation par la régression déterministe. En outre, $e_i^* = e_j$ avec $P(e_i^* = e_j) = w_j / \sum_s w_i a_i$ indépendamment pour chaque $i \in s_m$. Alors, l'imputation par la régression aléatoire utilise les valeurs imputées $y_i^* = \mathbf{z}_i'\hat{\boldsymbol{\gamma}}_r + \epsilon_i^*$, $i \in s_m$, où $\epsilon_i^* = (\boldsymbol{\lambda}'\mathbf{z}_i)^{1/2}(e_i^* - \bar{e}_r)$ avec $\bar{e}_r = \sum_s w_j a_j e_j / \sum_s w_j a_j$. Soit $E_*(.)$ l'espérance sous le processus d'imputation aléatoire. Nous avons $E_*(\epsilon_i^*) = 0$ et $E_*(\hat{Y}_I)$ égale à (8). Par conséquent, l'estimateur imputé $\hat{Y}_I$ est approximativement sans biais sous l'hypothèse MN ou sous l'hypothèse MI. Il convient de souligner que l'imputation par la régression aléatoire couvre le cas particulier de l'imputation hot-deck (pondérée) aléatoire. Pour le montrer, considérons le modèle d'imputation par la moyenne $E_m(y_i) = \gamma$, $V_m(y_i) = \sigma^2$ et $\mathrm{Cov}_m(y_i, y_j) = 0, i \neq j$. Nous avons $\hat{\boldsymbol{\gamma}}_r = \sum_s w_i a_i y_i / \sum_s w_i a_i = \bar{y}_r$, la moyenne pondérée des valeurs de y fournies par les répondants et $e_j = y_j - \bar{y}_r$. Par conséquent, $y_i^* = \bar{y}_r + \epsilon_i^* = y_j$ correspond à la valeur de y_j pour les répondants tirés aléatoirement avec probabilité $w_j / \sum_s w_i a_i$.

L'estimateur imputé fondé sur l'imputation par la régression aléatoire est asymptotiquement entaché d'un biais sous l'hypothèse MNG. Pour obtenir un estimateur approximativement sans biais pour Y, nous proposons une méthode d'imputation par la régression aléatoire modifiée. Soit $\tilde{e}_j = (y_j - \mathbf{z}_j'\tilde{\boldsymbol{\gamma}}_r)/(\boldsymbol{\lambda}'\mathbf{z}_j)^{1/2}$ et $\tilde{e}_i^* = \tilde{e}_j$ avec $P(\tilde{e}_i^* = \tilde{e}_j) = \tilde{w}_j / \sum_s \tilde{w}_i a_i$ indépendamment pour chaque $i \in s_m$, où $\tilde{\boldsymbol{\gamma}}_r$ est donné par (17) et $\tilde{w}_i = w_i(1 - \hat{p}_i)/\hat{p}_i$. Alors, l'imputation par la régression aléatoire modifiée utilise les valeurs imputées $y_i^* = \mathbf{z}_i'\tilde{\boldsymbol{\gamma}}_r + \tilde{\epsilon}_i^*$, où $\tilde{\epsilon}_i^* = (\boldsymbol{\lambda}'\mathbf{z}_i)^{1/2}(\tilde{e}_i^* - \tilde{e}_r)$ avec $\tilde{e}_r = \sum_s \tilde{w}_j a_j \tilde{e}_j / \sum_s \tilde{w}_j a_j$. Nous avons $E_*(\epsilon_i^*) = 0$ et $E_*(\hat{Y}_I)$ égale à l'estimateur imputé sous l'imputation par la régression déterministe modifiée. Donc, l'estimateur imputé $\hat{Y}_I$ est approximativement sans biais sous l'hypothèse MNG ou sous l'hypothèse MI. Pour le cas particulier du modèle d'imputation par la moyenne, nous avons $\tilde{\gamma}_r = \sum_s \tilde{w}_i a_i y_i / \sum_s \tilde{w}_i a_i$ et $y_i^* = y_j$ correspond à la valeur de y_j pour les répondants tirés aléatoirement avec probabilité $\tilde{w}_j / \sum_s \tilde{w}_i a_i$.

3. Études par simulation

Nous avons effectué deux études par simulation afin d'étudier les propriétés en échantillon fini des méthodes d'imputation par la régression déterministe modifiée et par la régression aléatoire modifiée proposées en termes de biais et de la racine relative de l'erreur quadratique moyenne. La première étude par simulation consiste à comparer les propriétés de l'imputation par la régression déterministe classique et de l'imputation par la régression déterministe modifiée proposée lorsque le modèle d'imputation et/ou le modèle de non-réponse ne sont pas spécifiés correctement. La deuxième a pour but de comparer les propriétés de l'estimateur imputé obtenues en utilisant des classes d'imputation fondées sur les probabilités de réponse estimées et l'imputation par la moyenne pondérée (classique) à celles de l'estimateur imputé obtenu en utilisant la méthode d'imputation par la régression déterministe modifiée proposée.

3.1 Étude par simulation 1

Nous avons généré une population finie de taille $N = 1\,000$ contenant 3 variables : une variable d'intérêt y et deux variables auxiliaires z_1 et z_2. Pour cela, nous avons commencé par générer z_1 et z_2 indépendamment à partir de lois exponentielles de moyenne 4 et 30, respectivement. Puis, nous avons généré les valeurs de y conformément au modèle de régression

$$y_i = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \epsilon_i,$$

où les ϵ_i sont générées à partir d'une loi normale de moyenne 0 et de variance σ^2. Les valeurs des paramètres γ_0, γ_1 et γ_2 ont été fixées, respectivement, à 20, 2 et 0,1, et la variance σ^2 a été choisie de façon que le R^2 du modèle soit approximativement égal à 0,75. L'objectif est d'estimer le total de population $Y = \sum_U y_i$.

Nous avons généré $R = 5\,000$ échantillons aléatoires simples sans remise de taille $n = 100$ à partir de la population finie. Dans chaque échantillon, la non-réponse à la variable y a été générée selon les mécanismes de réponse suivants :

Mécanisme 1 : La probabilité de réponse p_{1i} de l'unité i est donnée par le modèle de régression logistique

$$\log \frac{p_{1i}}{1 - p_{1i}} = \lambda_0 + \lambda_1 z_{1i}.$$

Mécanisme 2 : La probabilité de réponse p_{2i} de l'unité i est donnée par le modèle de régression logistique

$$\log \frac{p_{2i}}{1 - p_{2i}} = \lambda_0 + \lambda_1 y_i.$$

Nous avons choisi les valeurs de λ_0 et λ_1 de façon que le taux de réponse global soit approximativement de 70 %. Les indicateurs de réponse a_{1i} et a_{2i} ont été générés indépendamment à partir d'une loi de Bernoulli avec les paramètres p_{1i} et p_{2i}, respectivement. Notons que, dans le cas du mécanisme de réponse 2, le mécanisme de réponse est non ignorable en ce sens que la probabilité de réponse dépend de la variable d'intérêt y.

Afin de compenser pour la non-réponse à la variable y, nous avons utilisé l'imputation par la régression déterministe classique pour laquelle les valeurs imputées sont données par (6) et l'imputation par la régression déterministe modifiée pour laquelle les valeurs imputées sont données par (18). Les imputations ont été fondées sur les modèles de y et de p énumérés au tableau 1, c'est-à-dire $y_{(1)}, y_{(2)}, y_{(3)}, y_{(4)}$ et $p_{(1)}, p_{(2)}, p_{(3)}$. Notons que $p_{(1)}$ correspond au mécanisme de réponse 1 et $y_{(1)}$, au modèle générant la population.

Tableau 1
Modèles utilisés pour l'imputation

Modèles pour y	Ordonnée à l'origine	z_1	z_2
$y_{(1)}$	Oui	Oui	Oui
$y_{(2)}$	Oui	Non	Oui
$y_{(3)}$	Oui	Oui	Non
$y_{(4)}$	Non	Oui	Oui
Modèles pour p_i	Ordonnée à l'origine	z_1	z_2
$p_{(1)}$	Oui	Oui	Non
$p_{(2)}$	Oui	Non	Oui
$p_{(3)}$	Non	Oui	Non

D'après chaque échantillon simulé, nous avons calculé l'estimateur imputé $\hat{Y}_I$ donné par (2) avec les valeurs imputées (6) et (18), en nous basant sur certaines combinaisons des modèles $y_{(a)}$ et $p_{(b)}; a = 1, \ldots, 4;$ $b = 1, 2, 3$. Comme mesure du biais d'un estimateur imputé $\hat{Y}_I$, nous avons utilisé le biais relatif (BR) simulé exprimé en pourcentage donné par

$$\mathrm{BR}(\hat{Y}_I) = \frac{\mathrm{Biais}(\hat{Y}_I)}{Y} \times 100, \qquad (23)$$

où

$$\mathrm{Biais}(\hat{Y}_I) = \frac{1}{R}\sum_{r=1}^{R} \hat{Y}_I^{(r)} - Y \qquad (24)$$

et $\hat{Y}_I^{(r)}$ représente la valeur de $\hat{Y}_I$ pour le r^{e} échantillon simulé. Comme mesure de la variabilité d'un estimateur impute $\hat{Y}_I$, nous avons utilisé la racine relative de l'erreur quadratique moyenne (RREQM) simulée exprimée en pourcentage donnée par

$$\mathrm{RREQM}(\hat{Y}_I) = \frac{\sqrt{\mathrm{EQM}(\hat{Y}_I)}}{Y} \times 100, \qquad (25)$$

où

$$\mathrm{EQM}(\hat{Y}_I) = \frac{1}{R}\sum_{r=1}^{R} (\hat{Y}_I^{(r)} - Y)^2. \qquad (26)$$

Les résultats concernant le biais relatif et la RREQM sont présentés au tableau 2 pour les échantillons générés selon le mécanisme de réponse 1 et au tableau 3 pour ceux générés selon le mécanisme de réponse 2. L'examen du tableau 2 montre clairement que, si l'imputation est effectuée conformément au modèle correct (c'est-à-dire $y_{(1)}$), l'imputation par la régression déterministe classique mène à un estimateur approximativement sans biais et est plus efficace que l'imputation par la régression déterministe modifiée en ce qui concerne la RREQM. Comme l'a souligné un examinateur, l'imputation par la régression déterministe modifiée peut produire des estimateurs plus efficaces que la régression déterministe classique. Autrement dit, il existe des scénarios (non examinés ici) pour lesquels la méthode d'imputation par la régression déterministe modifiée proposée pourrait être plus efficace que la méthode d'imputation par la régression déterministe classique.

Quand le modèle d'imputation est spécifié incorrectement (c'est-à-dire $y_{(2)}$ et $y_{(4)}$), l'imputation déterministe produit des estimateurs avec biais, tandis que l'imputation déterministe modifiée induit un biais faible à négligeable, à condition que le modèle de non-réponse soit spécifié correctement (c'est-à-dire $p_{(1)}$). Par conséquent, la RREQM est plus grande dans le cas de l'imputation déterministe classique que dans celui de l'imputation par la régression déterministe modifiée. Si les modèles d'imputation et de non-réponse sont tous deux spécifiés incorrectement (c'est-à-dire $y_{(4)} - p_{(2)}$), tous les estimateurs sont biaisés.

Tableau 2
Biais relatif (%) et RREQM (%) des estimateurs imputés sous le mécanisme de réponse 1

Scénario	Biais (classique)	Biais (proposé)	RREQM (classique)	RREQM (proposée)
$y_{(1)} - p_{(1)}$	0,19	−0,01	1,85	2,33
$y_{(2)} - p_{(1)}$	5,20	0,16	5,60	2,66
$y_{(3)} - p_{(1)}$	0,17	−0.04	1,87	2,37
$y_{(4)} - p_{(1)}$	−14,80	−3,50	15,00	6,70
$y_{(1)} - p_{(2)}$	0,19	0,12	1,85	1,86
$y_{(4)} - p_{(2)}$	−14,80	−14,80	15,00	14,60
$y_{(1)} - p_{(3)}$	0,19	0,05	1,85	1,88

Si l'on examine le tableau 3, il est évident que, dans le cas du mécanisme 2, l'estimateur imputé obtenu sous imputation par la régression modifiée donne des résultats aussi bons, voire meilleurs, que l'estimateur imputé obtenu sous imputation par la régression classique dans tous les scénarios. Ce résultat n'est pas étonnant, puisque, pour arriver à réduire efficacement le biais dans le cas de la non-réponse non ignorable, il est nécessaire d'utiliser toute l'information auxiliaire appropriée disponible. Or, l'information auxiliaire utilisée dans le cas de l'imputation par la régression modifiée proposée est plus riche que celle utilisée pour l'imputation par la régression classique, puisque, pour la première, on se sert des variables auxiliaires reliées à la variable d'intérêt y ainsi que celles reliées à la probabilité de réponse, tandis que

pour la seconde, on emploie uniquement les variables auxiliaires reliées avec la variable d'intérêt y.

Tableau 3
Biais relatif (%) et RREQM (%) des estimateurs imputés sous le mécanisme de réponse 2

Scénario	Biais (classique)	Biais (proposé)	RREQM (classique)	RREQM (proposée)
$y_{(1)} - p_{(1)}$	1,84	1,83	2,55	2,54
$y_{(2)} - p_{(1)}$	4,46	1,84	4,89	2,65
$y_{(3)} - p_{(1)}$	2,03	2,02	2,70	2,70
$y_{(4)} - p_{(1)}$	−4,58	−3,04	5,07	3,81
$y_{(1)} - p_{(2)}$	1,84	1,84	2,55	2,55
$y_{(4)} - p_{(2)}$	−4,58	−1,70	5,07	2,88
$y_{(1)} - p_{(3)}$	1,84	1,84	2,55	2,55

3.2 Étude par simulation 2

Nous avons généré une population finie de taille $N = 1\,000$ contenant trois variables : une variable d'intérêt y et trois variables auxiliaires z_1, z_2 et z_3, en commençant par générer z_1, z_2 et z_3 indépendamment à partir d'une loi exponentielle de moyenne 100, puis en générant les valeurs de y selon le modèle de régression

$$y_i = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 z_{1i}^2 + \epsilon_i,$$

où les ϵ_i sont générées à partir d'une loi normale de moyenne 0 et de variance σ^2. Les valeurs des paramètres $\gamma_0, \gamma_1, \gamma_2$ et γ_3 ont été fixées, respectivement, à 20, 10, 0,5 et 10. Nous avons choisi la variance σ^2 de façon que le R^2 du modèle soit approximativement égal à 0,66. L'objectif est d'estimer la moyenne de population $\bar{Y} = \sum_U y_i / N$. Afin de nous concentrer sur l'erreur due à la non-réponse/imputation, nous avons considéré le cas d'un recensement, c'est-à-dire $n = N = 1\,000$. Ensuite, nous avons généré la non-réponse à la variable y pour la population simulée selon les mécanismes de réponse suivants :

Mécanisme 1 : La probabilité de réponse p_{1i} de l'unité i est donnée par le modèle logistique

$$\log \frac{p_{1i}}{1 - p_{1i}} = \lambda_0 + \lambda_1 z_{1i} + \lambda_2 z_{3i}.$$

Mécanisme 2 : La probabilité de réponse p_{2i} de l'unité i est donnée par le modèle logistique

$$\log \frac{p_{2i}}{1 - p_{2i}} = \lambda_0 + \lambda_1 y_i + \lambda_2 z_{3i}.$$

Nous avons choisi les valeurs de λ_0, λ_1 et λ_2 de façon à obtenir un taux de réponse global d'environ 70 %. Les indicateurs de réponse a_{1i} et a_{2i} ont été générés indépendamment $R = 1\,000$ fois à partir d'une loi de Bernoulli avec les paramètres p_{1i} et p_{2i}, respectivement.

Nous avons utilisé deux stratégies pour corriger pour la non-réponse. La première consiste à diviser l'échantillon, s, en classes d'imputation $s_1, s_2, \ldots, s_C$ en nous fondant sur les variables auxiliaires z_1, z_2 et z_3. Pour former les classes, nous avons utilisé la méthode du score qui peut se décrire comme il suit. En utilisant l'information auxiliaire, nous avons d'abord estimé les probabilités de réponse, p_i, pour obtenir $\hat{p}_i$ pour les répondants ainsi que les non-répondants par régression logistique sur z_1, z_2 et z_3. Puis, en utilisant les $\hat{p}_i$, nous avons divisé la population en C classes en suivant la procédure FASTCLUS de SAS (qui utilise pour la classification l'algorithme des k moyennes). La méthode du score mène à une partition de la population telle que, dans les classes, les unités (répondants et non-répondants) sont homogènes par rapport aux valeurs $\hat{p}_i$. La deuxième stratégie s'appuie sur la méthode d'imputation par la régression modifiée proposée basée sur les variables auxiliaires z_1, z_2 et z_3. Le but de l'étude en simulation est de comparer les propriétés des deux estimateurs imputés de la moyenne de population $\bar{Y}$: a) Estimateur imputé fondé sur les C classes d'imputation :

$$\bar{y}_I^C = \sum_{c=1}^{C} \frac{\hat{N}_c}{\hat{N}} \bar{y}_{Ic}, \qquad (27)$$

où

$$\bar{y}_{Ic} = \frac{1}{\hat{N}_c}\left[\sum_{s_c} w_i a_i y_i + \sum_{s_c} w_i (1 - a_i) y_i^*\right],$$

et $\hat{N}_c = \sum_{s_c} w_i$. Nous avons utilisé l'imputation par la moyenne pondérée dans les classes; c'est-à-dire $y_i^* = \sum_{s_c} w_i a_i y_i / \sum_{s_c} w_i a_i$.

b) Estimateur imputé fondé sur l'imputation par la régression modifiée proposée, dénoté $\bar{y}_I$:

$$\bar{y}_I = \frac{1}{\hat{N}}\left[\sum_{s} w_i a_i y_i + \sum_{s} w_i (1 - a_i) y_i^*\right], \qquad (28)$$

où les valeurs imputées y_i^* sont données par (18) en utilisant $\mathbf{z}_\mathbf{i}' = (z_{1i}, z_{2i})'$ et $\hat{N} = \sum_s w_i$. Pour le mécanisme 1, nous avons calculé correctement les probabilités de réponse p_i en utilisant les variables z_1 et z_3, tandis que nous avons utilisé les variables z_1, z_2 et z_3 pour estimer p_i pour le mécanisme 2.

Soulignons que, dans cette étude en simulation, $w_i = 1$ pour tout $i \in U$, parce qu'aucun échantillonnage n'a eu lieu. Enfin, le tableau 4 donne une comparaison de ces estimateurs en ce qui concerne le biais relatif, donné par (23), et de la RREQM, donnée par (25). L'examen du tableau 4 montre clairement que l'estimateur imputé proposé (28) donne de nettement meilleurs résultats que l'estimateur (27) fondés sur les classes d'imputation en ce qui a trait à la RREQM, pour le mécanisme 1 ainsi que le mécanisme 2.

Tableau 4
Biais relatif (%) et RREQM (%) des estimateurs imputés

Estimateur imputé*	Nombre de classes	BR	RREQM
$\bar{y}_I^C$ (mécanisme 1)	1	14,4	14,5
	5	–0,02	4,26
	10	–0,85	7,33
	20	–0,20	8,61
	30	–0,03	8,61
	40	0,03	9,09
	50	0,06	9,44
$\bar{y}_I$ (mécanisme 1)	–	1,11	1,90
$\bar{y}_I^C$ (mécanisme 2)	1	29,0	29,1
	5	21,4	21,4
	10	21,0	21,1
	20	20,9	21,0
	30	20,9	21,0
	40	21,0	21,0
	50	21,0	21,0
$\bar{y}_I$ (mécanisme 2)	–	10,9	10,9

* $\bar{y}_I^C$ est donnée par (27) et $\bar{y}_I$ est donné par (28).

4. Estimation de la variance

À la présente section, nous établissons un estimateur de la variance de l'estimateur imputé $\hat{Y}_I$, en suivant l'approche renversée de Fay (1991). Le variance totale de $\hat{Y}_I$ sous une méthode d'imputation déterministe particulière est donnée par

$$V(\hat{Y}_I - Y) = E_r V_p(\hat{Y}_I - Y \mid \mathbf{a}) + V_r E_p(\hat{Y}_I - Y \mid \mathbf{a}), \quad (29)$$

où $\mathbf{a} = (a_1, \ldots, a_N)'$ est le vecteur des indicateurs de réponse (Shao et Steel 1999). Un estimateur de la variance globale $V(\hat{Y}_I - Y)$ dans (29) est donné par $v_t = v_1 + v_2$, où v_1 est un estimateur de $V_p(\hat{Y}_I - Y \mid \mathbf{a})$ étant donné les indicateurs de réponse a_i, et v_2 est un estimateur de $V_r[E_p(\hat{Y}_I - Y \mid \mathbf{a})]$. L'estimateur v_1 ne dépend pas du mécanisme de réponse ni du modèle d'imputation et, par conséquent, v_1 est valide sous l'hypothèse MNG ou sous l'hypothèse MI.

Sous l'imputation aléatoire correspondante, la variance de l'estimateur imputé $\hat{Y}_I$ est donnée par

$$V(\hat{Y}_I - Y) = E_r V_p E_*(\hat{Y}_I - Y \mid \mathbf{a}) + E_r E_p V_*(\hat{Y}_I - Y \mid \mathbf{a}) + V_r E_p E_*(\hat{Y}_I - Y \mid \mathbf{a}), \quad (30)$$

où $V_*(.)$ représente l'opérateur de variance en ce qui a trait à l'imputation aléatoire. Nous supposons que $E_*(\hat{Y}_I \mid \mathbf{a})$ coincide avec l'estimateur imputé pour le cas déterministe. Donc, $E_r V_p E_*(\hat{Y}_I - Y \mid \mathbf{a})$ est estimé par v_1 dans ce cas. La contribution supplémentaire à la variance due à l'imputation aléatoire vient de la composante $V_r E_p E_*(\hat{Y}_I - Y \mid \mathbf{a})$ qui est estimée par v_2. Par conséquent, il découle de (30) que la variance globale $V(\hat{Y}_I - Y)$ est estimée par $v_t = v_1 + v_* + v_2$. Le terme v_* est absent dans le cas de l'imputation déterministe.

4.1 p_i connues

À la présente section, nous supposons que les probabilités de réponse p_i sont connues. À la section 4.1.1, nous commençons par examiner le cas de l'imputation par la régression déterministe modifiée. Celui de l'imputation par la régression aléatoire modifiée est étudié à la section 4.1.2.

4.1.1 Imputation par la régression déterministe modifiée

Sous l'imputation par la régression déterministe modifiée, l'estimateur imputé quand les p_i sont connues peut s'écrire

$$\hat{Y}_{Ip} = \sum_s w_i a_i y_i + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_\mathbf{r})' \tilde{\boldsymbol{\gamma}}_{\boldsymbol{rp}}, \quad (31)$$

où

$$\tilde{\boldsymbol{\gamma}}_{\boldsymbol{rp}} = \left[\sum_s w_i a_i \frac{(1-p_i)}{p_i} \mathbf{z}_\mathbf{i} \mathbf{z}_\mathbf{i}' / (\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i})\right]^{-1} \left[\sum_s w_i a_i \frac{(1-p_i)}{p_i} \mathbf{z}_\mathbf{i} y_i / (\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i})\right]. \quad (32)$$

Pour obtenir v_1, nous appliquons la méthode de linéarisation de Taylor standard qui donne

$$\hat{Y}_{Ip} - Y \approx \sum_s w_i \tilde{\xi}_{ip}, \quad (33)$$

où

$$\tilde{\xi}_{ip} = a_i y_i + (1 - a_i) \mathbf{z}_\mathbf{i}' \tilde{\boldsymbol{\gamma}}_{\boldsymbol{rp}} + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_\mathbf{r})' \tilde{\mathbf{T}}_\mathbf{p}^{-1} a_i \frac{(1-p_i)}{p_i} \frac{1}{(\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i})} \mathbf{z}_\mathbf{i} (y_i - \mathbf{z}_\mathbf{i}' \tilde{\boldsymbol{\gamma}}_{\boldsymbol{rp}})$$

avec

$$\tilde{\mathbf{T}}_\mathbf{p} = \sum_s w_i a_i \frac{(1-p_i)}{p_i} \mathbf{z}_\mathbf{i} \mathbf{z}_\mathbf{i}' / (\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i}).$$

Si nous dénotons l'estimateur de la variance de l'estimateur en échantillon complet $\hat{Y} = \sum_s w_i y_i$ par $v(y)$, il découle de (33) qu'un estimateur de $V_p(\hat{Y}_I - Y \mid \mathbf{a})$ est donné par

$$v_1 = v(\tilde{\xi}_p), \quad (34)$$

que nous obtenons en remplaçant y_i par $\tilde{\xi}_{ip}$ dans la formule de $v(y)$.

Pour obtenir la deuxième composante v_2, commençons par noter que

$$E_p(\hat{Y}_{Ip} - Y \mid \mathbf{a}) \approx \sum_s a_i y_i + \sum_U (1 - a_i) \boldsymbol{\gamma}_\boldsymbol{p} - Y,$$

où

$$\boldsymbol{\gamma}_\boldsymbol{p} = \left[\sum_U a_i \frac{(1-p_i)}{p_i} \mathbf{z}_\mathbf{i} \mathbf{z}_\mathbf{i}' / (\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i})\right]^{-1} \sum_U a_i \frac{(1-p_i)}{p_i} \mathbf{z}_\mathbf{i} y_i / (\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i}).$$

En appliquant la linéarisation de Taylor, nous pouvons montrer que

$$V_r[E_p(\hat{Y}_{Ip} - Y \mid \mathbf{a})] \approx \sum_U p_i (1 - p_i) \zeta_i^2, \quad (35)$$

où

$$\zeta_i = \left[1 + \frac{(1-p_i)}{p_i}\frac{1}{(\boldsymbol{\lambda}'\mathbf{z_i})}(\mathbf{Z}-\mathbf{Z_r})'\mathbf{T_p^{-1}}\mathbf{z_i}\right](y_i - \mathbf{z_i}'\boldsymbol{\gamma}_p)$$

avec $\mathbf{Z} = \sum_U \mathbf{z_i}$, $\mathbf{Z_r} = \sum_U a_i\mathbf{z_i}$ et

$$\mathbf{T_p} = \sum_U a_i \frac{(1-p_i)}{p_i}\mathbf{z_i}\mathbf{z_i}'/(\boldsymbol{\lambda}'\mathbf{z_i}).$$

Nous obtenons alors la composante v_2 par estimation des quantités inconnues dans (35), ce qui mène à

$$v_2 = \sum_s w_i a_i (1-p_i)\hat{\xi}_i^2, \qquad (36)$$

où

$$\hat{\zeta}_i = \left[1 + \frac{(1-p_i)}{p_i}\frac{1}{(\boldsymbol{\lambda}'\mathbf{z_i})}(\hat{\mathbf{Z}}-\hat{\mathbf{Z}}_\mathbf{r})'\tilde{\mathbf{T}}_\mathbf{p}^{-1}\mathbf{z_i}\right](y_i - \mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_{rp}).$$

Un estimateur de la variance totale v_t est obtenu par sommation de (34) et de (36) : $v_t = v_1 + v_2$. En pratique, les probabilités de réponse sont inconnues. Par conséquent, il est impossible de calculer l'estimateur de la variance v_t. Une solution simple consiste à remplacer p_i par les probabilités de réponse estimées $\hat{p}_i$ dans (34) et (36), puis à utiliser l'estimateur résultant v_t comme estimateur de la variance de $\hat{Y}_I$. Comme nous le montrons à l'aide d'une étude en simulation à la section 4.3, cette méthode simple donne des résultats acceptables.

4.1.2 Imputation par la régression aléatoire modifiée

Nous commençons par noter que

$$V_*(y_i^*) = (\boldsymbol{\lambda}'\mathbf{z_i})\sum_s w_j \frac{(1-p_j)}{p_j}a_i(\tilde{e}_j - \tilde{e}_r)^2 \Big/ \sum_s w_j\frac{(1-p_j)}{p_j}a_j \equiv \tilde{s}_e^2$$

et $\mathrm{Cov}_*(y_i^*, y_j^*) = 0, i \neq j$. Donc, d'après (2), la composante v_*, due à l'imputation aléatoire, est donnée par

$$v_* = \sum_s w_i^2(1-a_i)V_*(y_i^*) = \sum_s w_i^2(1-a_i)\tilde{s}_e^2. \qquad (37)$$

Un estimateur de la variance totale est obtenu par sommation de (34), (36) et (37) : $v_t = v_1 + v_2 + v_*$. De nouveau, puisque les probabilités de réponse p_i sont inconnues, il est impossible de calculer v_* dans (37). Nous proposons de remplacer dans cette équation les p_i par les probabilités de réponse estimées $\hat{p}_i$.

4.2 p_i inconnues

Nous utilisons la méthode de Binder (Binder, 1983) pour dériver la composante v_1 lorsque les probabilités de réponse p_i sont estimées. Nous supposons que $p_i = f(\mathbf{u_i}'\boldsymbol{\eta})$, où $\boldsymbol{\eta}$ est un vecteur de dimension l de paramètres inconnus, $\mathbf{u_i}$ est un vecteur de dimension l de variables auxiliaires disponibles pour tout $i \in s$. Par exemple, dans le cas de la régression logistique, $f(\mathbf{u_i}'\boldsymbol{\eta}) = \exp(\mathbf{u_i}'\boldsymbol{\eta})/\exp(1+\mathbf{u_i}'\boldsymbol{\eta})$. Les probabilités de réponse estimées sont données par $\hat{p}_i = f(\mathbf{u_i}'\hat{\boldsymbol{\eta}})$, où $\hat{\boldsymbol{\eta}}$ est un estimateur convergent de $\boldsymbol{\eta}$. Soit $\boldsymbol{\theta} = (\boldsymbol{\eta}_N', \boldsymbol{\gamma}_N', Y)'$, où $\boldsymbol{\eta}_N$ et $\boldsymbol{\gamma}_N$ sont les paramètres de recensement correspondant à $\boldsymbol{\eta}$ et $\boldsymbol{\gamma}$, respectivement. Un estimateur de $\boldsymbol{\theta}$ donné par $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\eta}}', \tilde{\boldsymbol{\gamma}}_r', \hat{Y}_I)'$ peut être exprimé comme une solution des équations d'estimations au niveau de l'échantillon

$$\hat{\mathbf{S}}(\boldsymbol{\theta}) = 0,$$

où $\hat{\mathbf{S}}(\boldsymbol{\theta}) = (\hat{\mathbf{S}}_1(\boldsymbol{\theta}), \hat{\mathbf{S}}_2(\boldsymbol{\theta}), \hat{S}_3(\boldsymbol{\theta}))'$ avec

$$\hat{\mathbf{S}}_\mathbf{1}(\boldsymbol{\theta}) = \sum_s w_i\mathbf{u_i}[a_i - f(\mathbf{u_i}'\boldsymbol{\eta}_N)] = \mathbf{0},$$

$$\hat{\mathbf{S}}_\mathbf{2}(\boldsymbol{\theta}) = \sum_s w_i a_i \mathbf{z_i}\frac{(1-f(\mathbf{u_i}'\boldsymbol{\eta}_N))}{f(\mathbf{u_i}'\boldsymbol{\eta}_N)}(y_i - \mathbf{z_i}'\boldsymbol{\gamma}_N)/(\boldsymbol{\lambda}'\mathbf{z_i}) = \mathbf{0}$$

et

$$\hat{S}_3(\boldsymbol{\theta}) = Y - \sum_s w_i\mathbf{z_i}'\boldsymbol{\gamma}_N - \sum_s w_i a_i(y_i - \mathbf{z_i}'\boldsymbol{\gamma}_N) = 0.$$

Soit $\hat{\mathbf{J}}(\boldsymbol{\theta}) = (\partial\hat{\mathbf{S}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta})$ la matrice de dimensions $(k+l+1)\times(k+l+1)$ des dérivées partielles. Nous avons

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) = [\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta})]\boldsymbol{\Sigma}(\boldsymbol{\theta})[\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta})]',$$

où $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ dénote la matrice symétrique de dimensions $(k+l+1)\times(k+l+1)$ dont l'élément ij est la covariance entre $\hat{S}_i(\boldsymbol{\theta})$ et $\hat{S}_j(\boldsymbol{\theta})$ en ce qui a trait à l'échantillonnage, sachant le vecteur des indicateurs de réponse **a**. Si $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ est remplacé par un estimateur convergent $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$, disons, nous obtenons un estimateur de la variance convergent $\mathbf{v}(\hat{\boldsymbol{\theta}})$ donné par

$$\mathbf{v}(\hat{\boldsymbol{\theta}}) = [\hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}})]\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})[\hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}})]'.$$

Puisque nous nous intéressons à l'estimateur de la variance, v_1, de $\hat{Y}_I$, nous avons besoin de la ligne finale, **b**, disons, de $\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta})$, évaluée à $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Il s'ensuit que

$$v_1 = \mathbf{b}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\mathbf{b}'. \qquad (38)$$

Pour obtenir la composante v_2, nous supposons que les poids d'échantillonnage w_i satisfont $\max(n/Nw_i) = O(1)$ et qu'il existe une constante positive C telle que $C < p_i$. En outre, nous supposons que $\hat{\boldsymbol{\eta}} - \boldsymbol{\eta} = O_p(n^{-1/2})$. Par linéarisation de Taylor, nous obtenons

$$\hat{Y}_I = \hat{Y}_{Ip} + (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})\sum_s p_i^{-1}(y_i - \tilde{\boldsymbol{\gamma}}_a)\partial\frac{f(\mathbf{u_i}'\boldsymbol{\eta})}{\partial\boldsymbol{\eta}} + O_p(N/n),$$

où

$$\tilde{\boldsymbol{\gamma}}_a = \left[\sum_U (1-a_i)\mathbf{z_i}\mathbf{z_i}'/(\boldsymbol{\lambda}'\mathbf{z_i})\right]^{-1}\left[\sum_U (1-a_i)\mathbf{z_i}y_i/(\boldsymbol{\lambda}'\mathbf{z_i})\right].$$

En supposant que $f(\mathbf{u_i}'\boldsymbol{\eta})/\partial\boldsymbol{\eta}$ est bornée uniformément, nous avons

$$E_p(\hat{Y}_I) = E_p(\hat{Y}_{Ip}) + O_p(N/n^{1/2}).$$

Donc, la composante $V_r[E_p(\hat{Y}_{Ip} - Y \mid \mathbf{a})]$ est donnée approximativement par (35) et v_2 est donné par (36) avec p_i remplacé par $\hat{p}_i$. Dans le cas de l'imputation par la régression aléatoire modifiée, la composante due à

l'imputation aléatoire serait estimée au moyen de (37) avec p_i remplacé par $\hat{p}_i$.

4.3 Étude par simulation

Nous avons réalisé une petite étude par simulation afin d'évaluer les propriétés des estimateurs de la variance envisagés aux sections 4.1 et 4.2. Nous avons généré une population de taille $N = 2\,500$ contenants deux variables y et z. Nous avons d'abord généré la variable z à partir d'une loi Gamma avec un paramètre d'échelle égal à 4 et un paramètre de forme égal à 10. Puis, nous avons produit les valeurs de y conformément au modèle du ratio

$$y_i = \gamma z_i + \epsilon_i,$$

où les ϵ_i sont générées à partir d'une loi normale de moyenne 0 et de variance σ^2. Nous avons fixé la valeur du paramètre γ à 2 et avons choisi la variance σ^2 de façon que le R^2 du modèle soit environ égal à 0,81. L'objectif est d'estimer le total de population $Y = \sum_U y_i$.

Nous avons généré $R = 10\,000$ échantillons aléatoires simples sans remise à partir de la population finie en utilisant les fractions d'échantillonnage n/N suivantes : 0,05, 0,1 et 0,25. Dans chaque échantillon, nous avons généré la non-réponse à la question y selon le mécanisme de réponse suivant : la probabilité de réponse p_i pour l'unité i est donnée par le modèle logistique

$$\log \frac{p_i}{1 - p_i} = \lambda_0 + \lambda_1 z_i.$$

Nous avons choisi les valeurs de λ_0 et λ_1 de façon à obtenir un taux de réponse global d'environ 70 %. Nous avons généré les indicateurs de réponse a_i indépendamment à partir d'une loi de Bernoulli de paramètre p_i.

Pour corriger pour la non-réponse à la variable y, nous avons utilisé l'imputation par le ratio déterministe modifiée pour laquelle les valeurs imputées sont données par (19). D'après chaque échantillon simulé, nous avons calculé l'estimateur imputé $\hat{Y}_I$ donné par (2) avec les valeurs imputées (19). Comme mesure du biais de l'estimateur de la variance v, nous avons utilisé le biais relatif $[E(v) - \text{EQM}(\hat{Y}_I)]/\text{EQM}(\hat{Y}_I)$. Nous désignons par v_{naive} l'estimateur de la variance totale obtenu par sommation de (34) et (36) quand les probabilités de réponse p_i sont remplacées par les probabilités de réponse estimées $\hat{p}_i$, et par v_{correct} l'estimateur de la variance totale obtenue par sommation de (38) et (36) avec p_i remplacé par $\hat{p}_i$. Le tableau 5 donne le biais relatif (en %) des deux estimateurs de la variance. Il montre clairement que ces estimateurs donnent lieu à une sous-estimation, mais que celle-ci est un peu moins prononcée dans le cas de v_{correct}. En outre, ils donnent tous deux de bons résultats, le biais relatif étant inférieur à –10 %. Donc, l'estimateur de la variance le plus simple v_{naive} pourrait convenir en pratique.

Tableau 5
Biais relatif (%) des estimateurs de la variance

f	$\text{BR}(v_{\text{naive}})$	$\text{BR}(v_{\text{correct}})$
0,05	–6,3	–5,1
0,10	–5,8	–4,1
0,25	–4,3	–3,2

5. Estimation de moyennes de domaine

En pratique, il est souvent nécessaire de produire des estimations pour divers domaines (sous-populations). Par exemple, dans le cas de l'Enquête sur la population active du Canada, des estimations du chômage sont requises selon le groupe âge-sexe et selon l'industrie au niveau provincial. Pour corriger pour la non-réponse partielle, on pourrait utiliser la méthode d'imputation par la régression modifiée proposée. Cependant, les domaines doivent être spécifiés d'avance à l'étape de l'imputation. Autrement dit, les indicateurs de domaine doivent faire partie du modèle d'imputation. Or, en pratique, les domaines ne sont généralement pas spécifiés à l'étape de la vérification et de l'imputation, si bien que les estimations par domaine sont calculées d'après des données imputées fondées sur des modèles d'imputation ne contenant pas les indicateurs de domaine. Par conséquent, les estimateurs imputés utilisés pour les domaines sont généralement biaisés. Nous proposons un estimateur corrigé pour le biais, s'inspirant de la section 2.2, pour remédier à ce problème. L'estimateur corrigé pour le biais peut être obtenu à l'étape de l'estimation et ne nécessite pas la spécification des domaines à l'étape de l'imputation.

Nous pouvons exprimer un vecteur de moyennes de domaine sous la forme

$$\overline{\mathbf{Y}}_{(\mathbf{d})} = \left(\sum_U \mathbf{x_i}\, \mathbf{x_i'} \right)^{-1} \sum_U \mathbf{x_i}\, y_i, \qquad (39)$$

où $\mathbf{x} = (x_{1i}, \ldots, x_{di}, \ldots, x_{Di})'$ est un vecteur d'indicateurs de domaine, x_{di}, tel que $x_{di} = 1$ si $i \in$ domaine d et $x_{di} = 0$, autrement. Nous supposons que $\mathbf{x}$ est connu pour toutes les unités $i \in s$. Autrement dit, seule la réponse à la variable y peut être manquante. En l'absence de non-réponse, un estimateur approximativement sans biais de $\overline{\mathbf{Y}}_{(d)}$ est donné par

$$\hat{\overline{\mathbf{Y}}}_{(\mathbf{d})} = \left(\sum_s w_i\, \mathbf{x_i}\, \mathbf{x_i'} \right)^{-1} \sum_s w_i\, \mathbf{x_i}\, y_i. \qquad (40)$$

En présence de non-réponse à la question y, un estimateur imputé de $\overline{\mathbf{Y}}_{(d)}$ est donné par

$$\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}} = \hat{\mathbf{T}}^{-1}\left[\sum_s w_i a_i \mathbf{x_i} y_i + \sum_s w_i (1-a_i)\mathbf{x_i} y_i^*\right] = \hat{\mathbf{T}}^{-1}\sum_s w_i a_i \mathbf{x_i} \tilde{y}_i, \quad (41)$$

où $\hat{\mathbf{T}} = \sum_s w_i \mathbf{x_i}\mathbf{x_i}'$. Notons que l'estimateur imputé $\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}}$ dans (41) ne nécessite pas les identificateurs de réponse, a_i. Haziza et Rao (2005) ont montré que l'estimateur imputé $\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}}$ est biaisé sous l'hypothèse MN. Ils ont proposé un estimateur corrigé pour le biais qui est approximativement sans biais sous l'hypothèse MN ou sous l'hypothèse MI. Nous proposons ici une extension de l'estimateur corrigé pour le biais de Haziza-Rao qui est approximativement sans biais sous l'hypothèse MNG ou sous l'hypothèse MI.

Il est facile de voir que, sous l'hypothèse MNG, le biais de non-réponse conditionnel de l'estimateur imputé (41) basé sur l'imputation par la régression déterministe modifiée (18) est de la forme

$$\text{Biais}\left(\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}} \mid s\right) \approx -\hat{\mathbf{T}}^{-1}\left[\sum_s w_i (1-p_i)\mathbf{x_i}(y_i - \mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_{s,N})\right], \quad (42)$$

où $\tilde{\boldsymbol{\gamma}}_{s,N}$ est donné par (15). Un estimateur approximativement conditionnellement sans biais du biais exprimé par (42) est de la forme

$$\hat{B}\left(\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}} \mid s\right) \approx -\hat{\mathbf{T}}^{-1}\left[\sum_s \tilde{w}_i a_i \mathbf{x_i}(y_i - \mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}})\right], \quad (43)$$

où $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}}$ est donné par (17). Un estimateur corrigé pour le biais, $\hat{\bar{\mathbf{Y}}}^{\mathbf{a}}_{\mathbf{I(d)}}$, est alors obtenu sous la forme $\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}} - \hat{B}(\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}} \mid s)$, qui mène à

$$\hat{\bar{\mathbf{Y}}}^{\mathbf{a}}_{\mathbf{I(d)}} = \hat{\mathbf{T}}^{-1}\left[\sum_s \frac{w_i}{\hat{p}_i} a_i \mathbf{x_i}(y_i - \mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}}) + \sum_s w_i \mathbf{x_i}\mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}}\right]. \quad (44)$$

L'estimateur corrigé pour le biais (44) est approximativement sans biais sous l'hypothèse MI ou sous l'hypothèse MNG. Donc, il est robuste au sens de sa validité sous ces deux hypothèses. Cependant, il nécessite les identificateurs de réponse a_i ainsi que les probabilités de réponse $\hat{p}_i$, contrairement à l'estimateur imputé $\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}}$ dans (41).

Il est possible d'obtenir un estimateur corrigé pour le biais de la forme (44) si nous utilisons l'imputation par la régression déterministe classique au lieu de la méthode modifiée. Il est intéressant de constater que l'estimateur corrigé pour le biais est identique à l'estimateur obtenu en utilisant l'imputation par calage (Beaumont 2005). Ce dernier estimateur ne nécessite pas la connaissance de a_i ni de $\hat{p}_i$ dans le fichier de données imputé, mais les domaines doivent être spécifiés à l'étape de l'imputation, ce qui n'est pas toujours faisables en pratique.

Si le modèle de non-réponse (4) contient uniquement l'ordonnée à l'origine, nous avons $\hat{p}_i = \hat{p}$, où $\hat{p}$ désigne le taux global de réponse. Dans ce cas, l'estimateur corrigé pour le biais (44) se réduit à

$$\hat{\bar{\mathbf{Y}}}^{\mathbf{a}}_{\mathbf{I(d)}} = \hat{p}^{-1}\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}} + (1-\hat{p}^{-1})\hat{\mathbf{T}}^{-1}\sum_s w_i \mathbf{x_i}\mathbf{z_i}'\hat{\boldsymbol{\gamma}}_{\boldsymbol{I}}, \quad (45)$$

en notant que $\hat{\boldsymbol{\gamma}}_{\boldsymbol{r}} = \hat{\boldsymbol{\gamma}}_{\boldsymbol{I}}$, où, sous l'imputation par la régression déterministe,

$$\hat{\boldsymbol{\gamma}}_{\boldsymbol{I}} = \left(\sum_{i\in s} w_i \mathbf{z_i}\mathbf{z_i}'/(\boldsymbol{\lambda}'\mathbf{z_i})\right)^{-1} \times \left[\sum_{i\in s} w_i a_i \mathbf{z_i} y_i/(\boldsymbol{\lambda}'\mathbf{z_i}) + \sum_{i\in s} w_i (1-a_i)\mathbf{z_i} y_i^*/(\boldsymbol{\lambda}'\mathbf{z_i})\right] = \hat{\boldsymbol{\gamma}}_{\boldsymbol{r}}.$$

Haziza et Rao (2005) ont obtenu l'estimateur corrigé pour le biais (45).

Conclusion

Par souci de simplicité, nous avons considéré le cas d'une classe d'imputation unique, mais notre méthode MNG s'étend facilement à plusieurs classes d'imputation au moyen d'imputations distinctes dans le cas de plusieurs classes. Par exemple, nous pourrions procéder à l'imputation par la moyenne pondérée dans les classes en utilisant nos poids modifiés $\tilde{w}_i$. En outre, notre méthode peut être étendue au cas de l'imputation composite (Sitter et Rao 1997; Shao et Steel 1999) qui repose sur diverses imputations pour les réponse manquantes à une question selon l'information auxiliaire disponible. Par exemple, on peut recourir à l'imputation par le ratio lorsqu'une variable auxiliaire x est observée et à une autre méthode quand x n'est pas observée. Dans ce cas, l'approche MI fondée sur le modèle du ratio reliant y à x ne sera pas applicable, contrairement au cas où la variable x est observée sur toutes les unités échantillonnées.

Remerciements

Les travaux de recherche de J.N.K. Rao ont été financés par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada. Les auteurs remercient les examinateurs de leurs commentaires et suggestions utiles.

Bibliographie

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society*, B, 67, 445-458.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 15, 279-292.

Brick, J.M., et Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.

Deville, J.C., et Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference*, *US Bureau of the Census*, 429-440.

Haziza, D., et Rao, J.N.K. (2005). Inference for domains under imputation for missing survey data. *Canadian Journal of Statistics*, 33, 149-161.

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.

Narain, R.D. (1951). On sampling without replacement with variying probabiltities. *Journal of the Indian Society of Agricultural Statistics*, 2, 169-174.

Rao, J.N.K. (1990). Variance estimation under imputation for missing data. Rapport technique, Statistique Canada, Ottawa.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of American Statistical Association*, 91, 499-506.

Rao, J.N.K. (2005). Évaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage. *Techniques d'enquête*, 31, 127-151.

Rao, J.N.K., et Shao, J. (1992). On variance estimation under imputation for missing data. *Biometrika*, 79, 811-822.

Rao, J.N.K., et Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

Särndal, C.-E. (1992). Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation. *Techniques d'enquête*, 18, 257-268.

Shao, J., et Steel, P. (1999). Variance Estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Sitter, R., et Rao, J.N.K. (1997). Imputation for missing values and corresponding variance estimation. *Canadian Journal of Statistics*, 25, 61-73.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, juin 2006
Vol. 32, N° 1, pp. 73-85
Statistique Canada, N° 12-001-XIF au catalogue

Un modèle d'estimation et d'imputation des ménages du recensement non-répondants sous échantillonnage pour le suivi des cas de non-réponse

Elaine L. Zanutto et Alan M. Zaslavsky [1]

Résumé

L'échantillonnage pour le suivi des cas de non-réponse (échantillonnage SCNR) est une innovation qui a été envisagée lors de l'élaboration de la méthodologie du recensement décennal des États-Unis de 2000. L'échantillonnage SCNR consiste à envoyer des recenseurs auprès d'un échantillon seulement des ménages qui n'ont pas répondu au questionnaire initial envoyé par la poste; ce qui réduit les coûts, mais crée un problème important d'estimation pour petits domaines. Nous proposons un modèle permettant d'imputer les caractéristiques des ménages qui n'ont pas répondu au questionnaire envoyé par la poste, afin de profiter des économies importantes que permet de réaliser l'échantillonnage SCNR, tout en obtenant un niveau de précision acceptable pour les petits domaines. Notre stratégie consiste à modéliser les caractéristiques des ménages en utilisant un petit nombre de covariables aux niveaux élevés de détail géographique et des covariables plus détaillées (plus nombreuses) aux niveaux plus agrégés de détail géographique. Pour cela, nous commençons par classer les ménages en un petit nombre de types. Puis, au moyen d'un modèle loglinéaire hiérarchique, nous estimons dans chaque îlot la distribution des types de ménage parmi les ménages non-répondants non échantillonnés. Cette distribution dépend des caractéristiques des ménages répondants qui ont retourné le questionnaire par la poste appartenant au même îlot et des ménages non-répondants échantillonnés dans les îlots voisins. Nous pouvons alors imputer les ménages non-répondants non échantillonnés d'après cette distribution estimée des types de ménage. Nous évaluons les propriétés de notre modèle loglinéaire par simulation. Les résultats montrent que, comparativement aux estimations produites par des modèles de rechange, notre modèle loglinéaire produit des estimations dont l'EQM est nettement plus faible dans de nombreux cas et à peu près la même dans la plupart des autres cas. Bien que l'échantillonnage SCNR n'ait pas été utilisé lors du recensement de 2000, notre stratégie d'estimation et d'imputation peut être appliquée lors de tout recensement ou enquête recourant cet échantillonnage où les unités forment des grappes telles que les caractéristiques des non-répondants sont reliées aux caractéristiques des répondants vivant dans le même secteur, ainsi qu'aux caractéristiques des non-répondants échantillonnés dans les secteurs voisins.

Mots clés : Données manquantes; estimation pour petits domaines; ajustement proportionnel itératif; modèle loglinéaire; ECM.

1. Introduction

L'échantillonnage pour le suivi des cas de non-réponse (SCNR) est une innovation qui a été envisagée lors de l'élaboration de la méthodologie du recensement décennal des États-Unis de 2000 (U.S. Bureau of the Census 1997a, b). Selon les procédures suivies à l'heure actuelle pour 99 % des ménages, le Census Bureau commence par envoyer par la poste ou livrer sur place un questionnaire qui doit être retourné par la poste dûment rempli. Puis, des recenseurs essaient de prendre contact avec tous les ménages qui n'ont pas répondu par la poste (environ 35 % de ceux auxquels le questionnaire a été envoyé par la poste). La charge de travail que représente la communication avec ces quelques 42 millions de ménages fait de cette opération de suivi l'une des plus coûteuses du recensement.

L'échantillonnage SCNR comporte l'envoi de recenseurs auprès d'un échantillon seulement de ménages non-répondants. Cet échantillon est un échantillon non mis en grappes de logements non-répondants (l'« unité d'échantillonnage ») ou un échantillon en grappes comprenant toutes les unités non-répondantes dans un échantillon d'îlots de recensement (petites zones correspondant à peu près à un îlot urbain ou à une région rurale compacte comptant environ 15 logements). Cette deuxième étape de suivi se solde par l'obtention d'un questionnaire rempli (par procuration ou imputation, au besoin) pour tous les logements échantillonnés, sauf ceux pour lesquels il est déterminé qu'ils sont inoccupés.

L'économie que permet de réaliser l'échantillonnage est importante, mais l'approche nécessite l'estimation des caractéristiques d'un très grand nombre de ménages non-répondants non échantillonnés, ce qui pose un problème considérable d'estimation pour petits domaines (Ghosh et Rao 1994; Rao 2003). Nous montrons qu'en utilisant des modèles appropriés pour imputer les caractéristiques des ménages non-répondants non échantillonnés, nous pouvons profiter des économies importantes de l'échantillonnage SCNR, tout en obtenant un niveau de précision acceptable pour les estimations sur petits domaines. Notre stratégie consiste à modéliser les caractéristiques des ménages en utilisant un petit nombre de covariables aux niveaux élevés de détail géographique et des covariables plus détaillées

1. Elaine L. Zanutto, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, États-Unis. Courriel : zanutto@wharton.upenn.edu; Alan M. Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, États-Unis. Courriel : zaslavsky@hcp.med.harvard.edu.

(plus nombreuses) aux niveaux plus agrégés de détail géographique. Pour cela, nous commençons par classer les ménages en un petit nombre de types. Puis, nous utilisons un modèle loglinéaire hiérarchique pour estimer dans chaque îlot la distribution des types de ménage parmi les ménages non-répondants non échantillonnés. Cette distribution dépend des caractéristiques des ménages répondants du même îlot qui ont répondu par la poste et des ménages non-répondants échantillonnés dans les îlots voisins. Nous pouvons alors imputer les ménages non-répondants non échantillonnés d'après cette distribution estimée des types de ménage.

Bien que l'échantillonnage SCNR n'ait pas été utilisé lors du Recensement de 2000 pour des raisons juridiques complexes, notre stratégie d'estimation et d'imputation peut être adoptée pour l'estimation pour petits domaines ou l'imputation lors de tout recensement ou enquête avec échantillonnage SCNR où les unités sont mises en grappes de telle sorte que les caractéristiques des non-répondants soient reliées à celles des répondants compris dans le même domaine, ainsi qu'à celles des non-répondants échantillonnés dans les domaines voisins. Les méthodes apparentées de Purcell et Kish (1980) et de Zhang et Chambers (2004) reposent aussi sur l'utilisation de modèles loglinéaires pour estimer des dénombrements croisés sur petits domaines en supposant que les totaux de population sont connus et que des données auxiliaires croisées sont disponibles au niveau du petit domaine. Nous possédons une source supplémentaire d'information, c'est-à-dire les caractéristiques des ménages non-répondants dans l'échantillon SCNR. Ceci nous permet de modéliser directement la relation entre les ménages répondants et non-répondants dans certains îlots.

À la section 2, nous résumons les stratégies proposées pour imputer les données manquantes dans cette situation. À la section 3.1, nous décrivons notre méthode générale d'échantillonnage et d'estimation. À la section 3.2, nous présentons notre modèle d'estimation et d'imputation et à la section 3.3, nos méthodes de lissage et d'estimation. À la section 4, nous évaluons notre modèle par simulation. Enfin, à la section 5, nous résumons les méthodes d'estimation de l'EQM et à la section 6, nous présentons nos conclusions.

2. Propositions antérieures pour l'imputation des non-répondants au recensement

Plusieurs méthodes ont été proposées pour imputer les caractéristiques des logements non-répondants. Les approches « descendantes » consistent à estimer d'abord des dénombrements pour des agrégats de ménages, puis à les répartir entre les petits domaines en maintenant la convergence avec les agrégats. Les modèles de ratio simple (Fuller, Isaki et Tsay 1994, nommés ci-après « FIT »), les modèles de régression de Poisson (Bell et Otto 1994), ou les modèles loglinéaires plus complexes (tels que ceux proposés ici et dans Zanutto et Zaslavsky 1995b, a) sont utilisés pour estimer les dénombrements pour de petits domaines et des groupes démographiques de faible niveau d'agrégation pour lesquels des estimations directes sont impossibles. Comme nous, FIT classifient les ménages en un nombre modéré de types définis d'après des caractéristiques importantes (par exemple, nombre de personnes, race, mode d'occupation du logement), puis estiment le nombre de ménages de chaque type parmi les non-répondants non échantillonnés. Ils produisent ensuite une liste de recensement complète en imputant le nombre estimé de ménages de chaque type. La différence principale entre notre approche et celle de FIT tient au fait que l'utilisation d'un modèle loglinéaire plutôt qu'un modèle de ratio stratifié nous donne plus de souplesse en ce qui concerne la finesse des contraintes imposées à divers niveaux de détail géographique. Bell et Otto (1994) estiment le nombre de personnes de 18 ans et plus de chaque race (Hispaniques, non-Hispaniques, Noirs, Autre) dans chaque logement non-répondant non échantillonné, mais n'examinent pas la façon dont il faut grouper les personnes imputées dans les ménages ni la façon d'imputer les caractéristiques du ménage, comme le mode d'occupation du logement. Ces modèles « descendants » *ad hoc* ne contiennent, au plus, que quelques caractéristiques des ménages et, par conséquent, ne modélisent pas explicitement leur structure, mais ils sont conçus de façon à assurer la cohérence des agrégats jugés les plus importants.

Schafer (1995) élabore une stratégie « ascendante » dans laquelle les ménages sont constitués en partant des individus, de leurs caractéristiques et de leurs relations, qui doivent toutes être exprimées par un modèle particulier. Ces modèles décrivent la population de façon plus détaillée et permettent de faire des inférences probabilistes (c'est-à-dire bayésiennes) complètes au sujet des caractéristiques inobservées. Cependant, contrairement à l'autre, cette approche oblige à construire un ensemble assez complexe de modèles avant que toute imputation puisse être faite. De surcroît, dans ce cadre, il est plus difficile d'assurer la convergence entre les microdonnées et les contrôles agrégés. Cependant, une stratégie combinée permettrait d'utiliser nos modèles pour produire des estimations quasiment sans biais selon le type de ménage et ceux de Schafer pour exécuter les imputations.

3. Méthodes d'estimation et modèles

3.1 Vue d'ensemble

À la première étape de la méthode d'imputation, nous prédisons pour chaque îlot les nombres de ménages

non-répondants non échantillonnés de chaque type au moyen d'une combinaison de modèles logistiques et loglinéaires. Cette étape est celle qui est l'objet du présent article (et de celui de FIT).

Pour la modélisation, nous avons classé les ménages par types en nous basant sur quelques caractéristiques importantes. Ici, nous utilisons 19 types, dont l'un est « logement inoccupé ». Les 18 autres sont définis par classification croisée des ménages pour trois catégories de taille (1 à 2 personnes, 3 à 4 personnes, 5 personnes et plus), trois catégories de race (Hispaniques, non-Hispaniques noirs et Autre) et deux catégories de mode d'occupation du logement (propriétaire, locataire).

Pour prédire le nombre de logements inoccupés parmi les logements non-répondants non échantillonnés dans chaque îlot, nous (et FIT) avons ajusté un modèle de régression logistique, en tenant compte du fait que la relation entre les ménages répondants et non-répondants n'est pas la même pour les logements inoccupés que pour ceux qui sont occupés. Les logements inoccupés répondants sont simplement ceux qui ont été considérés comme étant inoccupés par un facteur des services postaux, ce qui a entraîné le retour du questionnaire original. Leur distribution dépend vraisemblablement en grande partie des caractéristiques du logement associées à la distribution du courrier, ce qui nous renseigne peu sur la distribution des logements inoccupés non-répondants.

Après avoir modélisé les logements inoccupés, nous avons ajusté un modèle loglinéaire afin de prédire la distribution des types de ménage des logements occupés parmi les ménages non-répondants non échantillonnés restants à trois niveaux de détail géographique. L'îlot est la plus petite unité pour laquelle les dénombrements sont estimés. Le « domaine d'estimation » est l'unité la plus grande, c'est-à-dire la région dans laquelle l'estimation est faite indépendamment d'autres domaines de ce genre; dans notre application aux données de Recensement de 1990, il s'agit de la région pour laquelle le recensement a été réalisé à partir de l'un des 449 bureaux locaux de district, ou district de recensement (DR), comptant environ 200 000 ménages, en moyenne. Enfin, nous appelons « secteur » un niveau d'agrégation géographique intermédiaire comprenant un ensemble relativement homogène d'îlots contigus à l'intérieur d'un domaine d'estimation. Dans le contexte de la classification géographique type du Census Bureau, il pourrait s'agir de secteurs de recensement, de groupes d'îlots ou de secteurs du registre des adresses.

Nous exposons brièvement les dernières étapes qui seraient suivies pour obtenir les produits du recensement au moyen des estimations. À la deuxième étape de la méthode d'imputation, les dénombrements prévus seraient arrondis en nombres entiers. Des mécanismes sans biais (c'est-à-dire des procédures stochastiques qui, en prédiction, imputent le nombre prédit d'unités dans chaque cellule) d'« arrondissement contrôlé » (c'est-à-dire arrondissement dans un tableau à double entrée en préservant les totaux de marge) ont été établis par Cox (1987), ainsi que par George et Penny (1987). Cependant, d'autres études doivent être réalisées pour déterminer s'il est possible de modifier ces méthodes de sorte que les nombres de ménages soient arrondis en assurant le maintien de toutes les valeurs de marge correspondant aux effets inclus dans le modèle loglinéaire. Il s'agit d'un domaine où la recherche est active étant donné son importance en ce qui concerne la non-divulgation statistique.

Ensuite, les renseignements détaillés sur les individus et les ménages seraient imputés pour les ménages non-répondants en leur substituant des ménages donneurs ayant les mêmes caractéristiques. Les donneurs peuvent être choisis parmi les non-répondants échantillonnés, les répondants, ou une combinaison des deux sources. Enfin, des totalisations et des échantillons de microdonnées seraient préparés à partir des listes complètes.

3.2 Modèle loglinéaire

Nous avons ajusté un modèle loglinéaire pour estimer la prévalence des divers types de ménage parmi les ménages non-répondants non échantillonnés dans un DR, en utilisant des données provenant des répondants ainsi que des non-répondants dans l'échantillon SCNR pour ce DR. Le modèle prédit, pour chaque îlot, les types de ménage parmi les ménages non-répondants non échantillonnés, d'après des renseignements sur les caractéristiques des ménages répondants dans le même îlot et les caractéristiques des ménages non-répondants, déterminées d'après l'échantillon SCNR, dans les îlots voisins. À cette fin, le modèle loglinéaire contient des interactions entre les caractéristiques des ménages qui définissent le type de ménage et la situation de réponse à divers niveaux de détail géographique.

Cette stratégie de modélisation est motivée par le fait que, quand un modèle loglinéaire hiérarchique (c'est-à-dire un modèle dans lequel, pour chaque effet d'interaction inclus, les effets ou interactions principaux qui lui sont marginaux sont également inclus) est ajusté par la méthode du maximum de vraisemblance, les valeurs ajustées pour chaque total de marge ou moyenne correspondant à un effet dans le modèle sont égales aux totaux de marge ou aux moyennes observés correspondants (Birch 1963). Par conséquent, les prédictions pour les types de ménage concordent avec les taux observés pour les caractéristiques incluses dans le modèle, aux niveaux de détail géographique et situations de réponse correspondant aux interactions incluses dans le modèle. En outre, comme les prédictions du modèle pour les effets inclus sont contraintes de concorder avec les

taux observés sur un échantillon probabiliste (échantillon SCNR), les estimations correspondantes sont convergentes et approximativement sans biais. (Nous n'obtenons pas l'absence de biais parfaite, parce que le modèle de prédiction n'est pas linéaire et que le nombre de ménages non répondants non échantillonnés dans un îlot pourrait être associé à certaines caractéristiques des ménages non-répondants de l'îlot.)

Le modèle loglinéaire contient des facteurs géographiques emboîtés pour les îlots et les secteurs. Il contient aussi des facteurs croisés représentant les caractéristiques démographiques des ménages, à savoir un indicateur de réponse à la première étape (ménage répondant ou non-répondant), un indice de type de ménage et des expressions du modèle en les variables qui définissent les types de ménage. Ces expressions du modèle sont des sous-modèles contenant toutes les interactions qui définissent le type de ménage (c'est-à-dire race × taille × mode d'occupation).

Nous utilisons la notation suivante :

i = indice d'îlot ($i = 1, ...,$ nombre d'îlots dans le DR);

j = indice de type de ménage ($j = 1, ...,$ nombre de types);

r = indicateur de réponse à la première étape (envoi postal), $r = 0$ pour les ménages non-répondants et $r = 1$ pour les répondants;

$a = a(i)$ = indice indiquant le secteur qui contient l'îlot i ($a = 1, ...,$ nombre de secteurs);

$x_k = x_k(j)$, $k = 1, 2, 3, 4$ = expressions du modèle en les variables qui définissent les types de ménage où x_1 représente la classification croisée complète définissant les types de ménage, x_2 et x_3 représentent les expressions du modèle qui sont marginales à x_1, et x_4 est une expression du modèle qui est marginale à x_3. (Cette terminologie est expliquée plus bas).

Nous supposons que le modèle loglinéaire a la forme suivante :

$$n_{ijr} \sim \text{Poisson}(m_{ijr}),\ \log(m_{ijr}) = z_{ijr}^T \beta \qquad (1)$$

où n_{ijr} et m_{ijr} sont, respectivement, les dénombrements observés et prévus pour l'îlot i, le type de ménage j et la situation de réponse r, et Z est la matrice de plan d'expérience correspondent à la formule du modèle suivante:

$$x_1 + i * x_2 + i * r + r * x_3 + r * a * x_4. \qquad (2)$$

Conformément à la notation type de Wilkinson et Rogers (1973) pour les modèles linéaires généralisés, l'opérateur « * » indique que tous les effets principaux et toutes les interactions marginales à l'interaction donnée sont inclus dans le modèle, de sorte que celui-ci contient les effets principaux pour l'expression du modèle x_1, l'indicateur de réponse r, les indices d'îlot i et les interactions $i * x_2$, $i * r$, $r * x_3$ et $r * a * x_4$.

Puisque, dans (1), x_4 interagit avec le secteur, c'est-à-dire le niveau le plus faible d'agrégation pour lequel on dispose de données sur les non-répondants, cette expression du modèle devrait représenter une classification assez grossière des ménages n'incluant que les caractéristiques des ménages les plus importantes pour faire des imputations exactes au niveau du secteur. L'expression x_3 peut inclure des termes non compris dans x_4, puisqu'elle est ajustée à un niveau plus élevé d'agrégation géographique pour lequel un plus grand nombre de données sont disponibles. De même, l'expression x_1 pourrait inclure le plus grand nombre d'interactions, y compris l'interaction de toutes les variables qui définissent le type de ménage, puisqu'elle est ajustée au niveau d'agrégation géographique le plus élevé, en utilisant toutes les données disponibles. Enfin, x_2, qui peut différer de x_3, puisqu'elle interagit avec i au lieu de r, devrait être moins détaillée que x_1, puisqu'elle interagit avec l'îlot, c'est-à-dire un niveau d'agrégation géographique beaucoup plus faible. Ces lignes directrices sont motivées par le fait que les estimations des interactions avec i, r ou a sont déterminées d'après un nombre relativement faible d'observations et devrait demeurer simple. Le choix de x_2, x_3 et x_4 comme il est décrit plus haut devrait améliorer la précision des estimations par modèle, tout en maintenant les totaux de marge les plus importants.

Pour illustrer ce que pourraient être les termes $x_1, ..., x_4$ supposons que nous définissions le type de ménage par une classification croisée race × taille × mode d'occupation. Alors, une spécification possible de x_1, x_2 et x_3 est x_1 = race * taille * mode d'occupation, x_2 = race * taille + mode d'occupation, x_3 = taille * mode d'occupation, et x_4 = race + taille + mode d'occupation. Permettre que les termes $x_1, ..., x_4$ soient des expressions du modèle, plutôt que de simples interactions nous donne un moyen concis de représenter un modèle contenant toutes les interactions souhaitées. Par exemple, un modèle contenant un terme $i * x_2$, où x_2 correspond à la spécification susmentionnée, comprend à la fois une interaction îlot × race × taille et une interaction îlot × mode d'occupation.

Une interprétation heuristique de notre modèle loglinéaire est que nous estimons la distribution détaillée des types de ménage sur la région entière (x_1), puis que nous déplaçons cette distribution pour tenir compte des caractéristiques générales de l'îlot (x_2), des différences générales entre les ménages répondants et non-répondants (x_3), ainsi que des différences les plus importantes entre les ménages répondants et non-répondants dans le secteur étudié (x_4).

Toutes les interactions pourraient être incluses, sauf celles de la forme $r * i * x$, où x représente une expression du modèle en les variables qui définissent le type de ménage (c'est-à-dire telles que x_1, x_2, x_3, ou x_4). Les interactions de cette forme dépendent des totaux de marge déterminés uniquement d'après les ménages non-répondants dans un îlot unique, totaux qui ne sont pas disponibles pour les îlots non échantillonnés sous le plan d'échantillonnage des îlots et sont fondés sur un très petit échantillon sous le plan d'échantillonnage des unités de logement. Par conséquent, notre spécification du modèle exclut tous les effets $r * i * x$, qu'il est toujours impossible d'estimer (ou qui sont estimés médiocrement dans le cas du plan d'échantillonnage des ménages). Ce modèle généralise deux théories simples qui sont intégrées sous forme de sous-modèles. En premier lieu, s'il n'y a aucune différence entre les îlots (c'est-à-dire que les interactions loglinéaires $i * x_2$ et $a * x_4$ sont nulles), alors les ménages non-répondants dans chaque îlot sont imputés d'après la proportion globale de ménages non-répondants dans chacune des catégories x_3 dans l'échantillon SCNR, grâce à l'effet $r * x_3$. Autrement dit, les imputations sont faites en utilisant les mêmes proportions dans chaque îlot. En deuxième lieu, s'il n'existe aucune différence entre les répondants et les non-répondants (c'est-à-dire pas d'interaction $r * x_3$ ou $r * x_4$), alors les non-répondants sont imputés en utilisant les mêmes proportions que celles observées pour les répondants dans chaque îlot.

La formulation générale de notre modèle permet de tenir compte de nombreuses définitions du secteur et du type de ménage, et d'un choix nombreux d'expressions du modèle. Les secteurs devraient être définis de façon qu'ils soient suffisamment grands pour contenir des données adéquates pour l'estimation des interactions correspondantes, mais être aussi relativement homogènes. Par exemple, ils pourraient être définis par une combinaison de contigüités géographiques et de stratifications selon les covariables au niveau de l'îlot (comme le pourcentage de minorités), afin d'obtenir des secteurs plus homogènes dont les différences pourraient être décrites par modélisation. La généralisation a plus de deux niveaux d'agrégation géographique dans le domaine d'estimation est également simple. Donc, nous pourrions par exemple ajouter l'interaction d'une autre expression du modèle x_5 avec une unité géographique de niveau compris entre celui du secteur et de l'îlot.

Lors de l'ajustement du modèle par la méthode du maximum de vraisemblance, nous égalons aux valeurs observées correspondantes les quantités suivantes : 1) dénombrements ajustés d'îlot (au moyen de l'effet principal pour l'îlot, i), 2) taux de réponse par îlot (au moyen du terme $r * i$), 3) caractéristiques moyennes des ménages pour l'ensemble des ménages (pour les caractéristiques x_1 au moyen du terme d'effet principal pour x_1) et 4) par îlot (pour les caractéristiques x_2 au moyen du terme $i * x_2$), et 5) caractéristiques moyennes des ménages pour l'ensemble des non-répondants (pour les caractéristiques x_3 au moyen du terme $r * x_3$) et 6) pour les répondants par secteur (pour les caractéristiques x_4, au moyen du terme $r * a * x_4$). Donc, ce modèle généralise le modèle d'indépendance îlot × type utilisé par FIT et donne des résultats sans biais à des niveaux d'agrégation plus faibles, en supposant que les totaux de marges et les moyennes sont estimés sans biais d'après les données. L'estimation pour le secteur n'est pas exactement la même que l'estimation sans biais usuelle obtenue par estimation directe d'après l'échantillon SCNR, parce que le modèle fait concorder les valeurs de marge observées et ajustées pour les ménages compris dans l'échantillon. En effet, il existe un ajustement pour la covariance (régression) qui déplace l'agrégat de façon à tenir compte des différences observées entre les ménages répondants dans les îlots échantillonnés et les ménages répondants dans les îlots non échantillonnés, ou dans le cas du plan d'échantillonnage des logements, entre les ménages répondants dans les îlots pour lesquels des ménages sont sélectionnés dans l'échantillon SCNR et ceux pour lesquels l'échantillon SCNR ne contient aucun ménage.

L'idée de modéliser les caractéristiques des ménages en utilisant des covariables de niveau peu détaillé au niveau de l'îlot et des covariables de niveau plus détaillé aux niveaux géographique plus agrégés est semblable du point de vue conceptuel, mais non dans les détails, au modèle décrit dans Zaslavsky (2004). Pour une description de l'utilisation de poids loglinéaires pour faire concorder les estimations d'échantillon aux agrégats, voir Brackstone et Rao (1976), Oh et Scheuren (1983), et Zaslavsky (1988).

3.3 Estimation et lissage

Nous ajustons le modèle par estimation du maximum de vraisemblance sous échantillonnage de Poisson, ce qui équivaut à ajuster un modèle de régression logistique multinomial. Le fait que les données ne forment pas un tableau îlot × réponse × type complet, parce que nous disposons des dénombrements par îlot, mais non des caractéristiques des ménages non-répondants non échantillonnés, complique l'ajustement du modèle. Dans le cas du plan d'échantillonnage des îlots, les renseignements sur les caractéristiques manquent pour tous les non-répondants dans certains îlots et, dans le cas du plan d'échantillonnage des logements, les renseignements sur les caractéristiques manquent pour certains non-répondants dans presque tous les îlots. Pour ajuster le modèle, nous utilisons un algorithme d'ajustement proportionnel itératif (API) modifié adapté aux données qui sont classées partiellement dans une partie de l'ensemble de données (voir l'annexe).

Dans le cas de certains ensembles de données, il se peut que certains paramètres ne puissent être estimés parce que les estimations de vraisemblance se situent sur la frontière de l'espace du paramètre (infinie dans le cas de l'échelle loglinéaire, ce qui donne une valeur nulle sur l'échelle de dénombrement) ou parce qu'il n'existe aucun renseignement pour le paramètre. Adapter la spécification du modèle à chaque domaine d'estimation pour éliminer les paramètres qui ne peuvent être estimés est irréaliste dans le contexte de production d'un recensement.

Grâce à l'introduction d'une petite quantité d'information a priori, il est possible d'assurer que tous les paramètres puissent être estimés. Pour cela, nous annexons aux données dont nous disposons pour chaque secteur une petite quantité de « pseudo-données » dont les proportions selon le type sont égales à celles observées pour une région environnante (le DR dans nos simulations), en ajoutant ces dénombrements au tableau de données avant d'ajuster le modèle. Cette étape correspond à l'application d'une analyse bayésienne empirique à des données multinomiales de loi $f(n_1, ..., n_H \mid p_1, ..., p_H) \propto \prod_{i=1}^{H} p_i^{n_i}$, où $n_1, ..., n_H$ sont les nombres observés de ménages de chaque type dans un îlot ou un secteur. Si $\{p_i\}$ suit une loi a priori conjointe de Dirichlet, $f(p_1, ..., p_H) \propto \prod_{i=1}^{H} p_i^{\alpha_i - 1}$, $\alpha_i \geq 0$, la loi a posteriori résultante des p_i est une Dirichlet de paramètres $\alpha_i + x_i$ (Gelman, Carlin, Stern et Rubin 1995, page 76) et de mode a posteriori proportionnel aux paramètres. Donc, cette méthode bayésienne empirique équivaut à ajouter $\sum \alpha_i$ ménages au secteur, où les α_i de ces ménages sont du i^{e} type. Nous posons que les α_i sont proportionnels aux proportions observées pour chaque type de ménage dans une région environnante, de sorte que la distribution observée des types de ménage soit lissée par le mélange avec celle observée pour une région environnante, ce qui évite l'introduction d'un biais au niveau de la région plus grande. Cette spécification a priori induit une loi a priori sur les paramètres du modèle loglinéaire. Voir Rubin et Schenker (1987), Zaslavsky (1988), ainsi qu'un exemple et des références historiques dans Clogg, Rubin, Schenker, Schultz et Weidman (1991) pour une utilisation semblable du lissage.

Quand les paramètres du modèles sont estimés, l'étape suivante consiste à calculer les dénombrements prévus de chaque type de ménage pour les ménages non-répondants qui ne font pas partie de l'échantillon SCNR. Au moyen de l'algorithme d'ajustement proportionnel itératif, nous obtenons automatiquement les prédictions pour les ménages non-répondants non échantillonnés en appliquant les mêmes proportions d'ajustement à la partie partiellement observée du tableau qu'à la partie entièrement observée, de sorte qu'aucun calcul supplémentaire n'est nécessaire (voir l'annexe).

4. Simulations

4.1 Vue d'ensemble

Notre étude en simulation a pour but d'évaluer le biais, la variance et l'erreur quadratique moyenne (EQM) des estimations des agrégats démographiques étudiés (comme le nombre de ménages selon la race, la taille et le mode d'occupation du logement) à divers niveaux de détail géographique, en utilisant les compositions estimées des ménages pour les ménages non-répondants non compris dans l'échantillon SCNR. Les évaluations analytiques sont infaisables étant donné la complexité des modèles et du plan d'échantillonnage, la relation de dépendance entre les propriétés du modèle et la répartition géographique réelle des types de ménage, ainsi que le nombre de variantes du modèle qui pourraient être examinées.

Nous avons utilisé des données au niveau de l'îlot du recensement décennal des États-Unis de 1990 provenant de trois districts de recensement (DR); ces îlots représentent nos domaines d'estimation. Du point de vue structurel, ces simulations sont semblables à celles décrites par Schindler (1993) ou par FIT.

Les étapes de la simulation sont les suivantes :

1. Échantillonner les îlots ou les logements non-répondants conformément au plan d'échantillonnage SCNR.
2. Ajuster un modèle de régression logistique pour les ménages des logements inoccupés aux caractéristiques des ménages répondants et à celles des ménages non-répondants échantillonnés.
3. Calculer le nombre prévu de ménages non-répondants correspondant à un logement inoccupé pour chaque îlot.
4. Ajuster un modèle pour les types de ménages parmi les logements occupés en utilisant les caractéristiques des ménages répondants et des ménages non-répondants échantillonnés.
5. Calculer, pour chaque îlot, le nombre prévu de ménages non-répondants non échantillonnés pour chaque type de ménage parmi les logements occupés.
6. Calculer les agrégats d'intérêt d'après les dénombrements prévus et les comparer aux valeurs réelles au moyen de fonctions de perte.

Lors de l'exécution de nos simulations, répéter ces étapes 30 fois a produit des estimations de la racine carrée de l'erreur quadratique moyenne (RMSE pour Root mean square error) (définie à la section 4.3) d'une précision suffisante pour évaluer les propriétés de notre modèle comparativement à des modèles de rechange. Plus précisément, les coefficients de variation estimés des différences estimées de RMSE pour la méthode par le ratio stratifiée (décrite plus

bas) et la méthode loglinéaire sont inférieurs à 0,05, sauf quand la différence entre les RMSE estimées est très faible, auquel cas le coefficient de variation est très grand.

Nous comparons les propriétés de notre modèle à deux méthodes d'estimation de rechange, sous échantillonnage des unités de logement ainsi que des îlots. Chaque méthode commence par l'ajustement d'un modèle de régression logistique afin d'estimer, pour chaque îlot, le nombre de ménages non-répondants qui correspondent à des logements inoccupés. La première méthode, c'est-à-dire la « méthode du ratio non stratifiée », consiste à imputer des ménages dans chaque îlot pour remplacer les ménages non-répondants non échantillonnés proportionnellement à la répartition des types de ménage entre les ménages non-répondants dans l'échantillon de suivi pour le DR complet. La deuxième option, c'est-à-dire la « méthode du ratio stratifiée », est une variante de celle de FIT. Nous commençons par former des strates d'environ 82 îlots d'après la composition raciale de ces derniers, comme l'on décrit FIT. (Nous utilisons les données sur les répondants ainsi que les non-répondants pour former les strates, en supposant, comme FIT, que des renseignements semblables pourraient être tirés de dossiers administratifs. La stratification fondée uniquement sur l'information sur les répondants a donné des résultats comparables.) Puis, dans chaque strate, nous imputons des ménages non-répondants non échantillonnés aux divers types de ménage dans les logements occupés proportionnellement à la fréquence du type concerné dans l'échantillon de suivi pour la strate en question.

Nous simulons chaque méthode d'estimation en utilisant un taux d'échantillonnage SCNR de 30 %. Dans chaque strate, nous simulons cet échantillonnage en sélectionnant un échantillon aléatoire simple à 30 % d'îlots dans le cas du plan d'échantillonnage des îlots et un échantillon aléatoire simple à 30 % de ménages non-répondants dans chaque strate dans le cas du plan d'échantillonnage des logements. Nous supposons que les caractéristiques des ménages non-répondants dans ces échantillons sont connues (c'est-à-dire à la suite des opérations de suivi). Aussi bien pour la méthode du modèle loglinéaire que pour la méthode du ratio stratifiée, nous avons sélectionné un échantillon à 30 % d'îlots ou de ménages non-répondants par échantillonnage aléatoire simple sans remise dans que région.

Nous considérons plusieurs formulations du modèle loglinéaire. Tant pour l'échantillonnage des îlots que pour celui des logements, d'après les critères décrits à la section 4.3, le meilleur modèle est celui qui utilise $x_1 =$ taille* race*mode d'occupation, $x_2 =$ race * mode d'occupation + taille, $x_3 =$ race*taille, $x_4 =$ mode d'occupation. Ce modèle est celui que nous avons utilisé dans les simulations.

Pour être certain que le modèle puisse être ajusté à chaque cas et pour accélérer la convergence de l'ajustement proportionnel itératif, nous lissons les données en ajoutant à chaque îlot un ménage répondant hypothétique (« pseudo-données »). Ce ménage est réparti entre les 18 types de ménages correspondants aux logements occupés conformément aux proportions globales de ménages répondants dans le DR. Les estimations obtenues en utilisant cinq ménages pour le lissage étaient à peu près de la même précision que celles obtenues avec un ménage, et un lissage plus agressif (par ajout de 10, 15, 20 ou 25 ménages par îlot) augmentaient les erreurs d'estimation. En outre, même si l'ajout d'une petite fraction uniquement d'un ménage à chaque îlot suffit à assurer que le modèle puisse être ajusté à chaque cas, l'utilisation de moins d'un ménage par îlot ralentit considérablement la convergence et augmente légèrement l'erreur dans les estimations.

Les trois méthodes d'estimation s'appuient sur le même modèle de régression logistique pour les logements inoccupés. Pour chaque îlot, les covariables sont le taux de non-réponse par la poste, les pourcentages de ménages répondants qui sont (séparément) locataires, occupants d'un appartement et membres d'une race minoritaire (Noirs ou Hispaniques), la valeur moyenne des logements occupés par leur propriétaire, le loyer mensuel moyen pour les logements locatifs, des variables indicatrices pour chacun des secteurs et les interactions entre le pourcentage de locataires répondants et le loyer mensuel moyen, le pourcentage de locataires répondants et le carré du loyer mensuel moyen (centré sur la moyenne), le pourcentage de propriétaires répondants et la valeur moyenne des logements, et le pourcentage de propriétaires répondants et le carré de la valeur moyenne des logements (centré sur la moyenne). Afin d'éviter les problèmes de calcul causé par les îlots ne contenant pas de ménages non-répondants correspondants à des logements inoccupés, un ménage non-répondant hypothétique est ajouté à chaque îlot et réparti entre les logements inoccupés et occupés conformément à leur proportion dans les ménages non-répondants échantillonnés dans le DR.

4.2 Données

Nous utilisons les données du questionnaire abrégé du Recensement de 1990 pour trois DR dont les caractéristiques sont décrites au tableau 1. La race d'un ménage est déterminée par celle dont la prévalence est la plus élevée dans le ménage, habituellement la seule qui existe (98 % des ménages). Dans le DR 1, nous avons fusionné les groupes d'îlots consécutifs (et par conséquent contigus) (grappes d'îlots contigus) en 94 secteurs contenant, en moyenne, 52 îlots et 1 100 ménages. Pour les DR 2 et 3, nous ne disposions pas de renseignements sur les groupes d'îlots, si bien que nous avons formé des secteurs en groupant des îlots consécutifs en grappes contenant en moyenne 50 îlots

(en moyenne, 548 ménages par secteur dans le DR 2 et 918 ménages par secteur dans le DR 3).

Tableau 1
Caractéristiques des districts de recensement utilisés dans les simulations

	DR1	DR2	DR3
Ménages	112 966	169 321	149 567
Îlots	4 907	15 470	8 167
Pseudo-secteurs	94	309	163
Noir non-hispanique	14,4 %	28,5 %	1,3 %
Hispanique	6,1 %	1,0 %	6,6 %
Autre	73,5 %	59,4 %	81,5 %
Propriétaire	63,8 %	59,5 %	52,6 %
Locataire	30,2 %	29,4 %	36,7 %
Logement inoccupé	6,0 %	11,1 %	10,7 %
Taille 1 (1 à 2 personnes)	50,4 %	46,9 %	55,2 %
Taille 2 (3 à 4 personnes)	31,6 %	31,6 %	26,2 %
Taille 3 (5 personnes et plus)	12,0 %	10,4 %	7,9 %
Taux de réponse	72,6 %	65,3 %	56,7 %

4.3 Mesures du biais, de la variance et de l'erreur quadratique moyenne

Les fonctions de perte que nous utilisons pour nos évaluations sont fondées sur l'erreur relative pour la catégorie de ménage j (un type ou une combinaison de types) dans l'unité géographique i (un îlot ou un groupe d'îlots) :

$$d_{ijs} = \frac{\hat{Y}_{ijs} - Y_{ij}}{Y_{i+}} \qquad (3)$$

où Y_{ij} est le nombre réel de ménages dans la catégorie j dans l'unité géographique i, $\hat{Y}_{ijs}$ est le nombre correspondant de ménages estimés d'après l'échantillon s (y compris ceux observés dans l'échantillon et ceux estimés par le modèle), et Y_{i+} est le nombre total de ménages dans l'unité géographique i.

Nous résumons le biais dans les dénombrements estimés pour la catégorie j et un niveau de détail géographique (îlot, secteur, DR) au moyen de la racine carrée du biais quadratique moyen pondéré (RMWSB pour *Root Mean Weighted Squared Bias*) :

$$\hat{\mathrm{RMWSB}}_j^2 = \frac{\sum_i Y_{i+} \left\{ \left(\frac{1}{s}\sum_s d_{ijs} \right)^2 - \frac{1}{S(S-1)} \left(\sum_s d_{ijs}^2 - \frac{1}{S} \left(\sum_s d_{ijs} \right)^2 \right) \right\}}{\sum_i Y_{i+}} \qquad (4)$$

où S est le nombre d'échantillons tirés et $i = 1, \ldots, I$ où I est le nombre d'unités géographiques. Le deuxième terme du numérateur élimine un biais dû à la finitude de la simulation. Sous l'angle du plan de sondage, nous considérons la composition de chaque secteur comme une quantité fixe et seul l'échantillonnage est aléatoire. Alors on définit le biais comme la différence moyenne, sur tous les échantillons possibles, entre la valeur réelle pour un secteur donné et la valeur estimée correspondante; il s'agit essentiellement de l'erreur de modèle pour le secteur. Ce genre d'erreur est inévitable puisque la composition des non-répondants ne peut être entièrement prédite dans tout îlot. Un type plus grave de biais correspondrait à une erreur systématique dans les estimations pour un ensemble d'îlots de composition comparable. Nous n'avons pas recherché tous les types possibles de biais au sens susmentioné, mais la spécification du modèle nous protège contre ce biais aux niveaux élevés d'agrégation, parce que les estimations d'après le modèle sont contraintes de concorder (approximativement) avec des estimations sans biais pour les secteurs et les DR.

À titre de mesure de l'erreur globale, nous calculons la racine carrée de l'erreur quadratique moyenne pondérée (RMWMSE pour Root Mean Weighted Mean Squared Error) pour chaque catégorie de ménage j, qui est donnée par

$$\hat{\mathrm{RMWMSE}}_j^2 = \frac{\sum_i Y_{i+} \left(\frac{1}{s}\sum_s d_{ijs}^2 \right)^2}{\sum_i Y_{i+}} \qquad (5)$$

où Y_{ij}, $\hat{Y}_{ijs}$, Y_{i+}, i et S sont définis de la même façon que précédemment. (Les deux « moyennes » sont des moyennes sur les unités géographiques *i* et sur les échantillons *s*.) Nous obtenons une mesure de l'écart-type des estimations pour la catégorie de ménage j en calculant la racine carrée de la variance moyenne pondérée (RMWV pour Root Mean Weighted Variance) :

$$\hat{\mathrm{RMWV}}_j^2 = \frac{\sum_i Y_{i+} \left\{ \frac{1}{S-1} \left(\sum_s d_{ijs}^2 - \frac{1}{S} \left(\sum_s d_{ijs} \right)^2 \right) \right\}}{\sum_i Y_{i+}} = \hat{\mathrm{RMWMSE}}_j^2 - \hat{\mathrm{RMWSB}}_j^2 . \qquad (6)$$

Il convient de souligner que ces mesures de l'erreur quadratique moyenne, du biais et de l'écart-type sont toutes des estimations des espérances se rapportant aux échantillonnages SCNR répétés à partir de la population finie d'îlots. Ces fonctions de perte peuvent être appliquées à divers niveaux de détail géographique, ce qui reflète le fait que l'utilisation principale des estimations au niveau de l'îlot est l'agrégation pour former des estimations de niveau géographique plus élevée. En tenant compte de cet aspect, nous avons également choisi ces mesures parce qu'elles pondèrent les erreurs par la taille de l'unité géographique.

Cela nous donne des estimations convergentes de l'erreur lors de l'agrégation sur les unités géographiques, ce qui est approprié étant donné le caractère arbitraire des limites des unités ((Zaslavsky 1993). Nous fondons nos mesures sur les erreurs de mesure relatives à la population totale de la région géographique i plutôt qu'à la population dans la catégorie cible uniquement, parce que ce dernier dénominateur exagère l'importance des petites erreurs pour les îlots dans lesquels la catégorie n'apparaît que rarement ou jamais.

4.4 Résultats

Dans le cas de la simulation de l'échantillonnage SCNR selon un plan d'échantillonnage des îlots ainsi qu'un plan d'échantillonnage d'unités de logement, nous estimons le nombre de ménages possédant chaque caractéristique aux niveaux de l'îlot, du secteur et du DR au moyen de chacune des trois méthodes d'estimation. À la figure 1, les résultats pour chaque méthode sont représentés par les diverses barres ombrées pour l'échantillonnage des logements. (Nous ne présentons pas les résultats pour l'échantillonnage des îlots, mais le profil des résultats est comparable, la RMWMSE étant environ 10 % plus élevée pour toutes les estimations.) Dans cette figure, chaque ligne de graphiques à barres donne la RMWMSE pour les estimations aux niveaux de l'îlot, du secteur et du DR, pour l'un des trois DR. Chaque groupe de trois barres représente la RMWMSE pour les estimations du nombre total de ménages selon chaque catégorie de mode d'occupation du logement, chaque catégorie de taille du ménage et chaque catégorie de race en utilisant chacune des trois méthodes. Comme les trois méthodes s'appuient sur le même modèle de régression logistique pour prédire le nombre de logements non-répondants non échantillonnés inoccupés dans chaque bloc, nous avons omis la catégorie des logements inoccupés dans les graphiques.

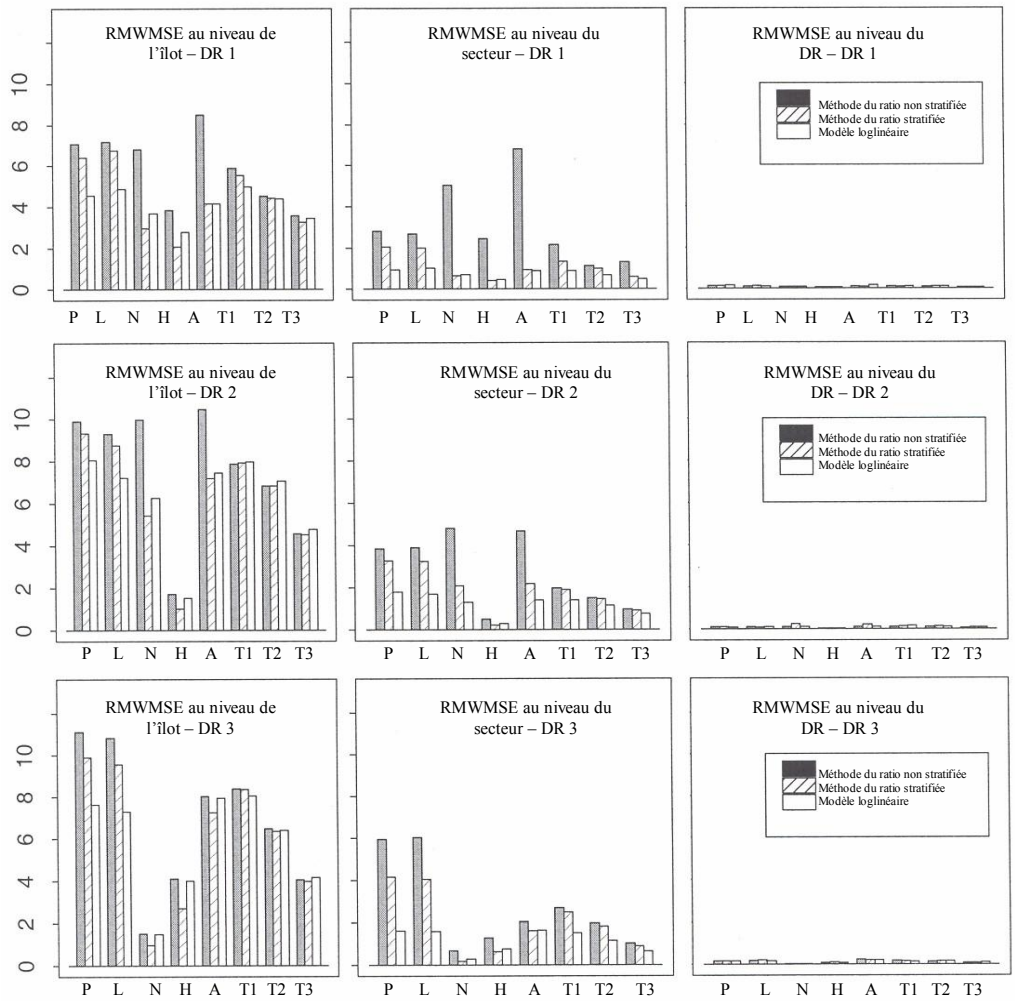


Figure 1. Estimations de la RMWMSE aux niveaux de l'îlot, du secteur et du DR pour chaque caractéristique des ménages, en utilisant le plan d'échantillonnage des logements pour les DR 1, 2 et 3, avec 30 échantillons simulés. (P = propriétaire, L = locataire, N = Noir, H = Hispanique, A = Autre race, T1 = Groupe de taille 1 (1 à 2 personnes), T2 = Groupe de taille 2 (3 à 4 personnes), T3 = Groupe de taille 3 (5 personnes et plus).

La RMWMSE est nettement plus faible pour la méthode du ratio stratifiée et le modèle loglinéaire que pour la méthode du ratio non stratifiée pour la plupart des caractéristiques des ménages au niveau de l'îlot et du secteur. Par conséquent, nous limitons la suite de la discussion à la comparaison des deux premières méthodes.

Les différences les plus importantes se dégagent pour les catégories de mode d'occupation du logement aux niveaux de l'îlot et du secteur. Dans chaque DR, les estimations des catégories de mode d'occupation aux niveaux de l'îlot et du secteur produites par le modèle loglinéaire ont une RMWMSE beaucoup plus faible que celles obtenues par la méthode du ratio stratifiée, principalement parce que le modèle loglinéaire donne lieu à un biais nettement plus petit (RMWSB). Les écarts-types (RMWV) sont un peu plus grands pour le modèle loglinéaire sous échantillonnage des logements, mais à peu près égaux pour les deux méthodes sous l'échantillonnage des îlots. Le modèle loglinéaire produit un biais plus faible pour les catégories de mode d'occupation au niveau du secteur, parce que le mode d'occupation du logement est inclus dans le modèle à titre d'effet au niveau du secteur, x_4. La stratification en fonction de la race dans la méthode du ratio réduit la RMWMSE pour les catégories de race au niveau de l'îlot, mais les deux méthodes donnent une RMWMSE comparable pour les catégories de race aux niveaux du secteur et du DR. La méthode du ratio stratifiée perd son avantage par rapport au modèle loglinéaire au niveau du secteur, parce qu'elle n'utilise aucune information du niveau du secteur. Dans l'ensemble, les deux méthodes produisent des estimations dont la RMWMSE est comparable à tous les niveaux de détail géographique pour les catégories de taille.

La signification statistique (dans les conditions des simulations) des écarts entre les RMWMSE des diverses méthodes a été évaluée au moyen de tests *t*. Presque tous les écarts observés aux niveaux de l'îlot et du secteur, à l'exclusion de la catégorie des logements inoccupés, ont une valeur $p \leq 0{,}001$ pour le test bilatéral et, par conséquent, ne peuvent être attribués à une erreur de simulation.

5. Évaluation et prédiction de l'erreur de modélisation

Faute d'espace, nous nous limitons ici à résumer brièvement les méthodes d'estimation de l'erreur quadratique moyenne des estimations ajustées en utilisant les données d'échantillon. Les méthodes et les résultats peuvent être obtenus en s'adressant au premier auteur.

Pour commencer, nous avons élaboré des approximations analytiques qui prédisent l'effet de la variation du taux d'échantillonnage sur l'exactitude de nos estimations sans que nous soyons obligés d'exécuter des simulations supplémentaires pour chaque taux. Ces approximations peuvent être utiles pour l'établissement du plan d'échantillonnage. Nous produisons une approximation de la RMWMSE des estimations aux niveaux de l'îlot, du secteur et du DR pour un nouveau taux d'échantillonnage, selon le plan d'échantillonnage des îlots ainsi que celui des logements, sachant que des résultats de simulation pour un premier taux d'échantillonnage existent déjà, en combinant les estimations du biais et de la variance au taux d'échantillonnage courant au moyen de deux facteurs de rééchelonnement. Le premier reflète la nouvelle proportion de logements sur lesquels doivent porter l'estimation sous le nouveau taux d'échantillonnage, qui a une incidence sur le biais et la variance des estimations combinées. L'autre facteur reflète l'effet du taux d'échantillonnage sur la variance des estimations pour les non-répondants. Les simulations ont démontré l'exactitude des prédictions de la RMWMSE lorsqu'on utilise ces approximations, sauf pour certaines extrapolations extrêmes.

À l'aide de ces résultats, nous avons mis au point une procédure de contrevalidation pour faciliter l'obtention des estimations intra-échantillon de la RMWMSE à utiliser dans des conditions de production où les caractéristiques réelles des ménages non-répondants non échantillonnés sont inconnues. Pour chaque secteur, l'échantillon de suivi est subdivisé aléatoirement en C groupes de contrevalidation (d'îlots pour l'échantillonnage des îlots et de ménages pour l'échantillonnage des logements). Les groupes de contrevalidation sont éliminés chacun à leur tour et le modèle est ajusté aux non-répondants dans les $C-1$ groupes de contrevalidation restants et aux répondants dans l'ensemble des C groupes. Nous pouvons alors estimer la RMWMSE dans les conditions du plan d'échantillonnage simulées par les contrevalidations et projeter cette estimation au taux d'échantillonnage courant, ou tout autre taux d'intérêt, en utilisant les approximations décrites au paragraphe précédent. Les simulations indiquent que cette procédure donne des estimations exactes de la RMWMSE aux niveaux de l'îlot et du DR, et une certaine surestimation au niveau du secteur. Cette méthode fournit aussi des estimations distinctes du biais et de la variance qui, d'après les simulations, sont très précises. Ces résultats sont très utiles pour évaluer l'adéquation du modèle, car un mauvais ajustement sera trahi par une composante d'erreur quadratique moyenne importante due au biais.

6. Conclusion

Aux sections précédentes, nous avons présenté une approche basée sur un modèle pour imputer les caractéristiques des ménages non-répondants à un recensement qui n'ont pas été échantillonnés pour le suivi des cas de non-réponse. Dans les simulations, notre modèle loglinéaire

produit des estimations dont l'erreur est beaucoup plus faible que les deux autres méthodes examinées pour certaines variables étudiées, et à peu près équivalentes pour d'autres. Ces conclusions tiennent pour le plan d'échantillonnage des îlots ainsi que celui des logements. L'un des avantages de notre approche est que les modèles peuvent être spécifiés de façon à n'imposer des contraintes que sur quelques tableaux de marge ou interactions des caractéristiques aux niveaux de détail géographique les plus fins, où les données sont peu nombreuses, tout en ajustant des distributions plus détaillées des caractéristiques à des niveaux d'agrégation géographique plus élevés auxquels un plus grand nombre de données sont disponibles. Cette approche est en harmonie avec les pratiques habituelles concernant la diffusion des données du recensement, qui contiennent un nombre minimal de caractéristiques au niveau de l'îlot, mais des caractéristiques de plus en plus détaillées pour les plus grandes unités.

De nombreuses applications importantes des données du recensement comportent l'estimation de la population et de ses caractéristiques pour de petits domaines tels que les districts législatifs et les secteurs de planification des services sociaux (comme les écoles et les cliniques) et du développement commercial. Bien que ces domaines ne coïncident pas toujours avec les secteurs utilisés pour le calcul des estimations de recensement, le fait de contrôler les estimations du recensement de façon à ce qu'elles concordent avec des estimations sans biais à plusieurs niveaux de détail géographique accroît la probabilité que les estimations calculées pour des domaines pertinents pour l'élaboration des politiques créés en regroupant le tout ou certaines parties de ces secteurs seront également presque sans biais. Notre méthode donne des propriétés au niveau agrégé plus prévisible que des alternatives complexes comme la modélisation spatiale hiérarchique. Bien que cette dernière puisse produire des estimations dont l'erreur quadratique moyenne est plus faible aux niveaux les plus fins de détail géographique, l'ajustement de ce genre de modèle et la vérification de leur biais à divers niveaux d'agrégation géographique nécessiteraient des mises au point locales de grande portée qui seraient vraisemblablement irréalistes dans les conditions de production d'un recensement.

Notre méthodologie est illustrée ici dans le contexte d'un échantillonnage pour le suivi des cas de non-réponse pour le recensement décennal des États-Unis, mais notre stratégie d'estimation et d'imputation peut être utilisée pour l'estimation et l'imputation pour de petits domaines dans le cadre de tout recensement ou enquête utilisant un échantillonnage SCNR où les populations présentent une structure hiérarchique. Nous pouvons aussi intégrer des enregistrements administratifs comme covariables afin de prédire les caractéristiques des ménages non-répondants correspondants (Zanutto et Zaslavsky 2002). Dans un tel scénario, les données sur les ménages inclus dans l'échantillon SCNR pour lesquelles nous possédons des renseignements provenant à la fois du recensement et des enregistrements administratifs sont utilisées pour estimer les écarts systématiques entre les deux sources de renseignements. Sous les mêmes modèles, nous imputons les caractéristiques des ménages non-répondants non échantillonnés. L'utilisation des enregistrements administratifs dans cette approche de modélisation peut améliorer l'exactitude des estimations sur petits domaines (niveau de l'îlot).

La discussion de l'échantillonnage dans le contexte du recensement des États-Unis s'avère politiquement litigieuse, mais il n'en reste pas moins qu'à long terme, il est probable qu'une forme ou l'autre d'estimation sera utilisée pour les non-répondants. Les possibilités pourraient être encore plus grandes dans les pays où les estimations démographiques reposent déjà sur une utilisation importante des dossiers administratifs (Redfern 1989). Des méthodes telles que celles décrites ici permettant de combiner l'information provenant de plusieurs sources de données tout en reflétant la diversité locale seront des éléments essentiels de ce genre d'efforts.

Annexe

Ajustement proportionnel itératif avec données partiellement croisées

Une approche type de l'ajustement de modèles loglinéaires à des données partiellement croisées consiste à utiliser un algorithme EM (Dempster, Laird et Rubin 1977; Little et Rubin 2002, chapitre 8) dans lequel, par étapes alternées, 1) les dénombrements prévus sont imputés sous les conditions du modèle et 2) le modèle est rajusté aux données observées et imputées par la technique de l'ajustement proportionnel itératif (API) (Darroch et Ratcliff 1972) pour des modèles sans solutions analytiques. Dans la modification ECM plus efficace de cet algorithme, un seul cycle de l'algorithme API est réalisé à chaque étape (Meng et Rubin 1993).

Dans le cas de notre application, nous avons développé un algorithme API plus rapide que les algorithmes EM et ECM pour nos modèles, qui comprend toujours une interaction îlot × réponse et ne comprend aucune interaction îlot × type × réponse. Nous avons constaté que notre algorithme API modifié converge après environ la moitié ou les deux tiers du nombre de cycles qu'exige l'algorithme EMC, en demandant moins de calculs à chaque étape (Zanutto 1998, partie 1, annexe A). (Nous déclarons qu'il y a convergence quand les valeurs prévue et observée des

statistiques minimales suffisantes du modèle sont suffisamment proches.)

Notre algorithme tire parti du fait que les observations partiellement classifiées ne contribuent à la vraisemblance que par la voie du nombre total de ménages non-répondants dans chaque îlots. Par conséquent, pour maximiser cette part de la vraisemblance, nous devons uniquement nous assurer que le nombre ajusté de non-répondants dans chaque îlot est égal au nombre observé, ce qui est automatique, parce que l'interaction îlot × réponse est toujours incluse dans notre modèle.

L'algorithme API modifié ajute le modèle aux observations entièrement classifiées au moyen d'un algorithme API ordinaire, en ne tenant pas compte des observations partiellement croisées. Dans le cas du plan d'échantillonnage des îlots, cela signifie que le modèle est ajusté en utilisant la partie entièrement observée du tableau îlot × type × réponse en utilisant un algorithme API ordinaire, en ne tenant pas compte de la partie partiellement classifiée du tableau. Nous obtenons les prédictions pour les cellules partiellement croisées en appliquant à ces cellules les mêmes proportions d'ajustement qu'à la partie entièrement observée du tableau. Enfin, nous rééchelonnons les cellules partiellement croisées de sorte que le nombre ajusté de non-répondants dans chaque îlot soit égal au nombre observé. Dans le cas du plan d'échantillonnage des logements, nous utilisons le même algorithme, en considérant l'ensemble de ménages répondants et de ménages non-répondants dans l'échantillon de suivi comme étant analogue à la partie entièrement observée du tableau dans le cas de l'échantillonnage des îlots et en considérant les îlots sans ménages non-répondants dans l'échantillon de suivi comme étant analogues aux îlots hors de l'échantillon dans le cas de l'échantillonnage des îlots. Ceci donne des prédictions pour les ménages non-répondants dans les îlots sans ménages non-répondants dans l'échantillon de suivi. Nous obtenons les prédictions pour les ménages non-répondants dans les îlots comptant un ou plusieurs ménages non-répondants dans l'échantillon de suivi en appliquant la répartition prévue des types de ménage entre les ménages non-répondants échantillonnés dans chacun de ces îlots aux ménages non-répondants non échantillonnés correspondants dans ces îlots. Pour plus de détails sur le cas de l'échantillonnage des logements, voir Zanutto et Zaslavsky (2002).

Nous illustrons maintenant l'algorithme API pour l'échantillonnage des îlots sous un modèle de Poisson tel que (1) avec $\log(m_{ijr}) = z_{ijr}^T \beta$, où m_{ijr} représente le nombre attendu, dans l'îlot i, de ménages de type j et de situation de réponse r, et Z est la matrice de plan d'expérience correspondant à l'expression du modèle $i * x + i * r + r * x$. Il s'agit d'une version simplifiée du modèle donné par (2) ne comportant qu'un seul niveau de détail géographique et une seule expression du modèle « x » représentant la classification croisée complète définissant les types de ménage. Nous observons n_{ijr} si $r = 1$ ou si $r = 0$ et $i \in S$, mais uniquement n_{i+0} si $i \notin S$, où S représente l'ensemble d'îlots sélectionnés dans l'échantillon SCNR.

L'algorithme API pour ajuster ce modèle débute par les estimations initiales $\hat{m}_{ijr}^0 = 1$ pour tout i, j, r et contient les trois étapes qui suivent dans le cycle t :

$$\text{Étape 1 :} \quad \hat{m}_{ijr}^{t+\frac{1}{3}} = \begin{cases} \hat{m}_{ijr}^t \left(\dfrac{n_{i+r}}{\hat{m}_{i+r}^t} \right) & \text{si } i \in S \text{ ou si } i \notin S,\ r = 1 \\ \hat{m}_{ijr}^t & \text{si } i \notin S,\ r = 0 \end{cases}$$

$$\text{Étape 2 :} \quad \hat{m}_{ijr}^{t+\frac{2}{3}} = \begin{cases} \hat{m}_{ijr}^{t+\frac{1}{3}} \left(\dfrac{n_{ij+}}{\hat{m}_{ij+}^{t+\frac{1}{3}}} \right) & \text{si } i \in S \\ \hat{m}_{ijr}^{t+\frac{1}{3}} \left(\dfrac{n_{ij1}}{\hat{m}_{ij1}^{t+\frac{1}{3}}} \right) & \text{si } i \notin S \end{cases}$$

$$\text{Étape 3 :} \quad \hat{m}_{ij1}^{t+1} = \hat{m}_{ij1}^{t+\frac{2}{3}} \left(\frac{n_{+j1}}{\hat{m}_{+j1}^{t+\frac{2}{3}}} \right)$$

$$\hat{m}_{ij0}^{t+1} = \hat{m}_{ij0}^{t+\frac{2}{3}} \left(\frac{\sum_{i \in S} n_{ij0}}{\sum_{i \in S} \hat{m}_{ij0}^{t+\frac{2}{3}}} \right).$$

À chaque étape, les facteurs d'échelle sont fondés uniquement sur les dénombrements observés.

Ces étapes sont répétées jusqu'à ce que les estimations des statistiques minimales suffisantes pour le modèle, à l'exclusion de $\hat{m}_{i+r}$ pour $i \notin S$, $r = 0$ (c'est-à-dire, $\hat{m}_{i+r}$ pour $i \in S$ et $i \notin S, r = 1$, $\hat{m}_{ij+}$ pour $i \in S$, $\hat{m}_{ij1}$ pour $i \notin S$, $\hat{m}_{+j1}$, et $\sum_{i \in S} \hat{m}_{ij0}$) soient suffisamment approchantes de leurs valeurs observées. Si nous dénotons l'étape à laquelle ceci a lieu par t^*, l'étape finale de cet algorithme consiste à fixer

$$\hat{m}_{ijr}^{t^*+1} = \begin{cases} \hat{m}_{ijr}^{t^*} \left(\dfrac{n_{i+r}}{\hat{m}_{i+r}^{t^*}} \right) & \text{si } i \notin S,\ r = 0 \\ \hat{m}_{ijr}^{t^*} & \text{autrement,} \end{cases}$$

pour assurer que la valeur de marge îlot × réponse estimée $(i * r)$ pour $i \notin S$, $r = 0$ soit égale à la valeur de marge observée.

Cette algorithme API produit des estimations qui convergent vers les estimations du maximum de vraisemblance

des paramètres du modèle (Zanutto 1998, partie 1, annexe A). À l'étape 2, le deuxième cas n'est pas nécessaire pour maximiser la vraisemblance, mais il est inclus pour obtenir des prédictions pour les cellules de non-répondants non échantillonnés (c'est-à-dire, $i \notin S$, $r = 0$).

Bibliographie

Bell, W.R., et Otto, M.C. (1994). Investigation of a model-based approach to estimation under sampling for nonresponse in the decennial census. Article non-publié présenté à la Joint Statistical Meetings, Toronto.

Birch, M.W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society*, Séries B, Methodological, 25, 220-233.

Brackstone, G.J., et Rao, J.N.K. (1976). Raking ratio estimators. *Techniques d'enquête*, 2, 63-69.

Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B., et Weidman, L. (1991). Multiple imputation of industry and occupation codes in census publicuse samples using Bayesian logistic regression. *Journal of the American Statistical Association*, 86, 68-78.

Cox, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.

Darroch, J.N., et Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43, 1470-1480.

Dempster, A.P., Laird, N.M., et Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Séries B, 39, 1-22.

Fuller, W.A., Isaki, C.T. et Tsay, J.H. (1994). Design and estimation for samples of census nonresponse. Dans *Proceedings of the Bureau of the Census Annual Research Conference*. Washington, DC: U.S. Bureau of the Census, 289-305.

Gelman, A., Carlin, J.B., Stern, H.S., et Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall Ltd.

George, J.A., et Penny, R.N. (1987). Initial experience in implementing controlled rounding for confidentiality control. Dans *Proceedings of the Bureau of the Census Annual Research Conference*, Volume 3. Washington, DC: U.S. Bureau of the Census, 253-262.

Ghosh, M., et Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-76.

Little, R.J.A., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, Deuxième édition. New York: John Wiley & Sons, Inc.

Meng, X.-L., et Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80, 267-278.

Oh, H.L., et Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse. Dans *Incomplete Data in Sample Surveys* (Éds. W.G. Madow, I. Olkin et D.B. Rubin). New York: Academic Press, 143-184.

Purcell, N.J., et Kish, L. (1980). Postcensal estimates for local areas (or domains). *Revue Internationale de Statistique*, 48, 3-18.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Redfern, P. (1989). L'expérience européenne relative à l'utilisation des données administratives pour recenser la population : Questions d'ordre politique. *Techniques d'enquête*, 15, 85-103.

Rubin, D.B., et Schenker, N. (1987). Interval estimation from multiply-imputed data: A case study using census agriculture industry codes. *Journal of Official Statistics*, 3, 375-387.

Schafer, J.L. (1995). Model-based imputation of census short-form items. Dans *Proceedings of the Bureau of the Census Annual Research Conference*. Washington, DC: Bureau of the Census, 267–299.

Schindler, E. (1993). Sampling for the count; sampling for non-mail returns. Rapport non-publié, U.S. Bureau of the Census.

U.S. Bureau of the Census (1997a). Census 2000 operational plan. Washington, DC.

U.S. Bureau of the Census (1997b). Report to Congress – the plan for Census 2000. Washington, DC.

Wilkinson, G.N. et Rogers, C.E. (1973). Symbolic description of factorial models for analysis of variance. *Applied Statistics*, 22, 392-399.

Zanutto, E. (1998). *Imputation for Unit Nonresponse*: *Modeling Sampled Nonresponse Follow-up*, *Administrative Records*, *and Matched Substitutes*. Thèse de maîtrise, Harvard University, Cambridge, Massachusetts.

Zanutto, E., et Zaslavsky, A.M. (1995a). A model for imputing nonsample households with sampled nonresponse follow-up. Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Zanutto, E., et Zaslavsky, A. M. (1995b). Models for imputing nonsample households with sampled nonresponse followup. Dans *Proceedings of the Bureau of the Census Annual Research Conference*. Washington, DC: U.S. Bureau of the Census, 673-686.

Zanutto, E., et Zaslavsky, A.M. (2002). Using administrative records to improve small area estimation: An example from the U.S. Decennial Census. *Journal of Official Statistics*, 18, 559-576.

Zaslavsky, A.M. (1988). Redressement des estimations régionales par une repondération des ménages. *Techniques d'enquête*, 14, 281-305.

Zaslavsky, A.M. (1993). Combining census, dual-system, and evaluation study data to estimate population shares. *Journal of the American Statistical Association*, 88, 1092-1105.

Zaslavsky, A.M. (2004). Representing the Census undercount by multiple imputation of households. Dans *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (Éds. A. Gelman et X.-L. Meng). West Sussex, England: John Wiley & Sons, Inc. 129-140.

Zhang, L.-C., et Chambers, R.L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society*, Séries B, 66, 479-496.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, juin 2006
Vol. 32, N° 1, pp. 87-96
Statistique Canada, N° 12-001-XIF au catalogue

Répartition de l'échantillon de la contre-vérification des dossiers de 2006

Alain Théberge [1]

Résumé

La répartition d'un échantillon peut être optimisée en fonction de divers objectifs. Lorsqu'il y a plus d'un objectif, on doit choisir une répartition qui équilibre ces objectifs. Traditionnellement, la Contre-vérification des dossiers a établi cet équilibre en consacrant une fraction de l'échantillon à chacun des objectifs (par exemple, les deux tiers de l'échantillon sont répartis de manière à obtenir de bonnes estimations provinciales, tandis qu'un tiers est réparti de manière à obtenir une bonne estimation nationale). Cet article suggère une méthode qui consiste à choisir le maximum de deux ou plusieurs répartitions. En étudiant l'impact de la précision des estimations démographiques sur les paiements de péréquation du gouvernement fédéral canadien aux provinces, on peut donner quatre objectifs à la répartition provinciale de l'échantillon de la Contre-vérification des dossiers. La répartition infraprovinciale de l'échantillon de la Contre-vérification des dossiers exige un lissage de paramètres définis au niveau des strates. Cet article montre comment le calage peut servir à ce lissage. Le problème de calage et sa solution n'exigent pas l'existence d'une solution aux contraintes de calage. Ceci évite des problèmes de convergence rencontrés par des méthodes connexes telles l'ajustement proportionnel itératif (raking).

Mots clés : Ajustement proportionnel itératif; calage; contre-vérification des dossiers; lissage; répartition de l'échantillon.

1. Introduction

Le Recensement de la population du Canada est réalisé tous les cinq ans; il l'a été la dernière fois en 2001. La Contre-vérification des dossiers (CVD) vise à mesurer le sous-dénombrement du recensement et une partie du surdénombrement. Pour la prochaine CVD, qui sera menée en 2006, on espère que la plus grande partie du surdénombrement du recensement sera mesurée au moyen d'une autre enquête, soit l'Étude par appariement automatisé, qui est plus efficace pour cette tâche. Cette façon de procéder devrait permettre d'optimiser la répartition de l'échantillon de la CVD aux fins de la mesure du sous-dénombrement. Les estimations de la CVD de la couverture sont utilisées conjointement avec les chiffres du recensement pour produire des estimations démographiques. Les estimations démographiques servent notamment au calcul, par le ministère fédéral des finances, des paiements de péréquation du gouvernement canadien aux gouvernements provinciaux.

Habituellement, on procède à la répartition de l'échantillon de la CVD aux provinces en tâchant de trouver un juste milieu entre la nécessité de produire une estimation nationale de qualité du taux de personnes omises au recensement et la nécessité de produire des estimations provinciales de bonne qualité de ces taux, afin de produire les estimations démographiques de Statistique Canada.

On espérait que cette approche répondrait également aux besoins de produire des estimations de bonne qualité des paiements de péréquation (il s'agit d'estimations dans la mesure où ils sont fondés sur les estimations démographiques), mais cela n'a jamais été vérifié. Les paiements de péréquations sont effectués par le gouvernement fédéral canadien aux provinces moins prospères. Dans le présent article, nous examinons l'effet de la répartition provinciale de l'échantillon sur la qualité des estimations des paiements de péréquation.

Si la variance d'une variable d'intérêt est la même dans chaque province, alors on obtient une répartition optimale pour une estimation nationale dont la variance est minimale si la taille de l'échantillon est proportionnelle à la taille de la base de sondage pour la province p, N_p. Une répartition qui donne des estimations provinciales de variance égale est une répartition où la taille de l'échantillon est constante (proportionnelle à N_p^0). Une façon de mettre en équilibre les deux besoins souvent utilisée consiste à rendre la taille de l'échantillon proportionnelle à $N_p^{1/2}$. Aux fins de la CVD, on a utilisé dans le passé une méthode différente pour réaliser cet équilibre, selon laquelle une partie de l'échantillon est répartie de manière à produire des estimations provinciales de variance égale et l'autre partie est répartie de manière à produire une estimation nationale dont la variance est minimale. Habituellement, environ les deux tiers de l'échantillon sont répartis de manière à produire des estimations provinciales de variance égale.

Nous proposons dans le présent article une nouvelle méthode de répartition provinciale qui établit un juste équilibre entre deux objectifs ou plus. Elle consiste à calculer une répartition distincte pour chaque objectif, chaque répartition portant peut-être sur une taille totale d'échantillon différente; on obtient la répartition finale, qui devrait répondre à tous les objectifs, en prenant pour chaque province la taille d'échantillon maximale sur chacune des répartitions.

1. Alain Théberge, Division des méthodes d'enquêtes sociales, Statistique Canada, 15e étage, Immeuble R.-H.-Coats, Ottawa, (Ontario), Canada, K1A 0T6.

La répartition infraprovinciale optimale est simplement donnée par la répartition de Neyman. La difficulté est de prévoir la variance dans des strates relativement petites ou, plus précisément, de prévoir les totaux (nombre de personnes omises au recensement, nombre de non-répondants dans l'échantillon de la CVD) dont dépend la variance. Pour chaque province, l'approche utilisée dans cet article consiste à prendre d'abord les valeurs nationales plus stables au niveau de la cellule (âge × sexe × état matrimonial) et à les mettre à l'échelle de sorte que les totaux correspondent aux valeurs provinciales pour chaque groupe d'âge, pour chaque sexe et pour chaque état matrimonial. Cet objectif rappel celui de la méthode d'ajustement proportionnel itératif proposée par Deming et Stephan (1940) et utilisée également par Brackstone et Rao (1976). Deville et Särndal (1992) ont montré comment on peut utiliser le calage pour obtenir le même résultat. Dans le cas de la CVD, on aura recours au calage même s'il est impossible d'aligner les cellules dans une matrice à trois dimensions, étant donné que les groupes d'âge diffèrent pour chaque état matrimonial. La méthode itérative du quotient parfois ne converge pas en raison de l'impossibilité de respecter les contraintes. En énonçant le problème de calage comme dans Théberge (1999), on tient compte de la possibilité que les contraintes soient incohérentes et ceci ne cause pas de problèmes de convergence. En outre, l'utilisation de l'inverse de Moore-Penrose dans la solution permet aussi aux contraintes d'être linéairement dépendantes.

Dans la section qui suit, nous examinerons la relation entre les estimations démographiques et les paiements de péréquation. Comme nous le constaterons, le problème de répartition de l'échantillon exige de trouver un juste équilibre entre quatre objectifs. À la section 3, nous utilisons une formule de variance approximative qui s'appuie sur un effet du plan pour déterminer la répartition optimale qui découle de chacun des quatre objectifs. Nous déterminons empiriquement la valeur de l'effet du plan à la section 4. À la section 5, nous expliquons comment une répartition finale peut établir un juste équilibre entre les répartitions individuelles pour des objectifs distincts. Enfin, la section 6 porte sur la répartition infraprovinciale. Nous n'examinons pas dans cet article la répartition de l'échantillon pour les trois territoires.

2. Incidence des estimations démographiques sur les paiements de péréquation

Statistique Canada est chargé de produire des estimations démographiques. Ces estimations démographiques trouvent une utilisation importante dans le calcul des paiements de péréquation effectué par le ministère fédéral des finances. Bien que Statistique Canada ne soit pas directement concerné par la formule servant à calculer les paiements de péréquation, il est utile de déterminer l'effet de la précision des estimations démographiques sur la précision des paiements de péréquation. L'incidence de la répartition de l'échantillon sur la précision des estimations démographiques retient l'attention depuis longtemps; dans cet article, nous examinerons également l'incidence de la répartition de l'échantillon sur la précision des paiements de péréquation.

On utilise la CVD pour mesurer le taux de personnes omises au recensement. Dans le passé, la répartition de l'échantillon visait à la fois à réduire au minimum la variance pour le taux national estimé de sous-dénombrement (objectif I) et à produire des variances égales pour les taux estimés de sous-dénombrement dans chaque province (objectif II). Deux autres objectifs seront ajoutés en examinant l'impact de la répartition de l'échantillon sur la précision des paiements de péréquation.

La formule servant à calculer les paiements de péréquation, avant tout lissage basé sur des moyennes mobiles, est :

$$E_p = \sum_{j=1}^{33} \frac{R_{tj}}{T_{tj}} \left(\frac{T_{\text{norme } j}}{P_{\text{norme}}} - \frac{T_{pj}}{P_p} \right) P_p \,, \qquad (2.1)$$

où E_p est le paiement de péréquation pour la province bénéficiaire p (au moment d'écrire ces lignes, toutes les provinces sauf l'Ontario et l'Alberta), R_{tj} représente les recettes totales (toutes les provinces) provenant de la source de recettes j, T_{tj} est l'assiette fiscale totale pour la source de recettes j, $T_{\text{norme } j}$ est l'assiette fiscale des provinces de référence servant à établir la norme (toutes les provinces sauf les provinces de l'Atlantique et l'Alberta) pour la source de recettes j, P_{norme} est la population des provinces de référence, T_{pj} est l'assiette fiscale de la province bénéficiaire p pour la source de recettes j, et P_p est la population de la province bénéficiaire p.

Pour mesurer l'incidence des estimations démographiques sur les paiements de péréquation, nous réécrirons l'équation (2.1) comme suit :

$$E_p = \left(\frac{P_p}{P_{\text{norme}}} \right) C_{\text{norme}} - K_p \,, \qquad (2.2)$$

où

$$C_{\text{norme}} = \sum_{j=1}^{33} \frac{R_{tj} T_{\text{norme } j}}{T_{tj}}$$

et

$$K_p = \sum_{j=1}^{33} \frac{R_{tj} T_{pj}}{T_{tj}} \,.$$

Nous constatons que la population de l'Alberta n'a pas d'incidence sur le paiement de péréquation d'une province bénéficiaire. La population de l'Ontario a un effet sur le paiement de péréquation seulement par P_{norme}. Dans le cas des provinces de l'Atlantique, leur paiement de péréquation varie linéairement en fonction de leur population, puisque celle-ci n'a pas d'effet sur P_{norme}. Si nous supposons que P_{norme} est connu, alors nous pouvons dire que, pour toute province bénéficiaire, une erreur d'une personne dans sa population a un effet de $C_{\text{norme}} / P_{\text{norme}}$ dollars sur son paiement de péréquation. Cela ne veut pas dire que le paiement de péréquation d'une province bénéficiaire dépend seulement de sa population et non de la population des provinces de référence. Toutefois, comme nous pourrons le constater, la plus grande partie de l'erreur d'échantillonnage dans le paiement de péréquation est attribuable à l'erreur d'échantillonnage dans l'estimation de la population de la province bénéficiaire et une partie relativement petite, à l'erreur d'échantillonnage dans l'estimation de la population des provinces de référence.

Si les symboles avec un chapeau représentent des estimateurs, alors d'après (2.2),

$$V(\hat{E}_p) \simeq C_{\text{norme}}^2 \frac{1}{P_{\text{norme}}^2} \left(\begin{array}{l} V(\hat{P}_p) \\ + \left(\dfrac{P_p}{P_{\text{norme}}} \right)^2 V(\hat{P}_{\text{norme}}) \\ - 2 \dfrac{P_p}{P_{\text{norme}}} \text{Cov}(\hat{P}_p, \hat{P}_{\text{norme}}) \end{array} \right). \quad (2.3)$$

Comme nous stratifions séparément pour chaque province, pour une province bénéficiaire *p*, qui n'est pas l'une des provinces de référence, nous avons, en faisant abstraction de la migration interprovinciale, $\text{Cov}(\hat{P}_p, \hat{P}_{\text{norme}}) = 0$, tandis que $\text{Cov}(\hat{P}_p, \hat{P}_{\text{norme}}) = V(\hat{P}_p)$ pour l'une quelconque des provinces de référence. Une approximation est obtenue en négligeant les deux derniers termes de (2.3) :

$$V(\hat{E}_p) \simeq \left(\frac{C_{\text{norme}}}{P_{\text{norme}}} \right)^2 V(\hat{P}_p). \quad (2.4)$$

En utilisant les données de la CVD de 2001, on peut vérifier que l'écart-type du paiement de péréquation dérivé de (2.4) diffère de celui dérivé de (2.3) par au plus 7 %, sauf pour deux provinces bénéficiaires : Terre-Neuve-et-Labrador où l'approximation sous-estime l'écart-type de 11 %, et le Québec où l'approximation surestime l'écart-type de 12 %.

Comme nous pouvons le constater d'après l'équation (2.4), une répartition de l'échantillon qui produit des variances égales pour les estimations de la population des provinces bénéficiaires produit des variances égales pour les estimations des paiements de péréquation des provinces bénéficiaires. Toutefois, le fait d'avoir des CV égaux pour les estimations de la population des provinces bénéficiaires ne garantit pas des CV égaux pour l'estimation des paiements de péréquation des provinces bénéficiaires, puisque, d'après l'équation (2.2), E_p n'est pas directement proportionnel à P_p parce que K_p n'est pas nul. Le fait d'avoir des CV égaux pour les estimations de la population des provinces bénéficiaires demeure un objectif valable, puisqu'il assure des intervalles de confiance de longueur égale pour le paiement de péréquation par personne. En fait, dû à l'utilisation de l'approximation (2.4), si la situation observée en 2001 se reproduit en 2006, l'intervalle de confiance pour Terre-Neuve-et-Labrador sera 11 % trop court (c'est-à-dire que la précision pour le paiement de péréquation par personne sera inférieure à celle des autres provinces bénéficiaires), tandis que l'intervalle de confiance pour le Québec sera 12 % trop long (c'est-à-dire que la précision pour le paiement de péréquation par personne sera supérieure à celle des autres provinces bénéficiaires). En outre, si nous faisons abstraction de la migration interprovinciale, alors les estimations démographiques provinciales sont indépendantes et la variance du total des paiements de péréquation est réduite au minimum si et seulement si la variance de la population totale des provinces bénéficiaires est réduite au minimum.

Nous cherchons à obtenir une répartition de l'échantillon au niveau provincial qui réduit au minimum la variance du paiement de péréquation total ou, de façon équivalente, celle de la population totale des provinces bénéficiaires (objectif III). Nous tâchons aussi d'obtenir une répartition de l'échantillon au niveau provincial qui produit des CV égaux pour l'estimation de la population de chaque province bénéficiaire (objectif IV), de manière à obtenir une précision égale pour le paiement de péréquation par personne.

La plus grande partie de la variation des estimations démographiques est attribuable à la variation des estimations du sous-dénombrement. Si l'on fait abstraction de la contribution du surdénombrement à la variation de l'estimation démographique, alors il est facile de vérifier que l'erreur type du taux estimé de sous-dénombrement est égale au CV de l'estimation démographique. Les objectifs I et II peuvent alors être reformulés comme étant de réduire au minimum le CV de l'estimation nationale de la population et de produire des estimations démographiques provinciales dont le CV est égal, respectivement. La différence entre les objectifs III et I, et entre les objectifs IV et II, c'est que ceux-ci s'appliquent, dans le premier cas, aux provinces bénéficiaires et, dans le deuxième cas, à toutes les provinces. Dans ce qui suit, nous supposons effectivement que la variance des estimations démographiques est égale à la variance des estimations du sous-dénombrement.

Les objectifs de la répartition provinciale de l'échantillon sont résumés dans le tableau 2.1.

Tableau 2.1
Les quatre objectifs de la répartition provinciale de l'échantillon

Objectif	Description (description équivalente)
I	Réduire au minimum la variance pour le taux national estimé de sous-dénombrement. (Réduire au minimum le CV de l'estimation nationale de la population.)
II	Produire des variances égales pour les taux estimés de sous-dénombrement dans chaque province. (Produire des estimations démographiques provinciales dont le CV est égal.)
III	Réduire au minimum la variance du paiement de péréquation totale. (Réduire au minimum la variance de l'estimation de la population totale des provinces bénéficiaires.)
IV	Produire des variances égales pour le paiement de péréquation par personne dans chaque province bénéficiaire. (Produire des CV égaux pour l'estimation de la population de chaque province bénéficiaire, ou encore, produire des variances égales pour les taux estimés de sous-dénombrement des provinces bénéficiaires.)

3. Répartition optimale de l'échantillon au niveau provincial

Dans la présente section, nous fournissons d'abord quelques précisions sur la notation adoptée, puis nous présentons des formules de variance approximatives pour les estimations démographiques et les taux de sous-dénombrement estimés. Nous examinons la question de l'optimalité relativement aux quatre objectifs susmentionnés.

Cinq bases de sondage sont utilisées aux fins de la CVD dans les provinces : la base du recensement (personnes dénombrées au recensement précédent), la base des naissances (naissances intercensitaires), la base des immigrants (immigrants intercensitaires), la base des résidents non permanents et la « base des personnes omises ». La « base des personnes omises » se compose des personnes échantillonnées de la CVD précédente qui ont été omises au recensement précédent. Avec leurs poids, ils représentent la sous population de personnes dénombrables qui ne sont pas couvertes par l'une des quatre autres bases. Chaque base dans chaque province est stratifiée séparément. Un échantillon aléatoire stratifié est sélectionné dans chaque base. Toutes les personnes de la base des personnes omises sont incluses dans l'échantillon.

Soit U_{hp} le nombre de personnes omises dans la strate h qui sont classées dans la province (de classification) p. De la même façon, soient E_{hp} et O_{hp}, respectivement, le nombre de personnes dénombrées et le nombre de personnes surdénombrées dans la strate h qui sont classées dans la province p, et $P_{hp} = U_{hp} + E_{hp} - O_{hp}$. Le taux de sous-dénombrement pour la province p peut alors s'écrire :

$$R_{.p} = U_{.p} / P_{.p}, \tag{3.1}$$

où $U_{.p} = \sum_h U_{hp}$ et $P_{.p} = \sum_h P_{hp}$. Nous voyons que $P_{.p}$ est égal à P_p tel que défini à la section précédente.

Un estimateur du taux de sous-dénombrement pour la province p est

$$\hat{R}_{.p} = \hat{U}_{.p} / \hat{P}_{.p}, \tag{3.2}$$

où $\hat{U}_{.p}$ et $\hat{P}_{.p}$ sont des estimateurs de $U_{.p}$ et $P_{.p}$, respectivement. Une linéarisation donne

$$V(\hat{R}_{.p}) \cong \frac{1}{P_{.p}^2}\left[V(\hat{U}_{.p}) + \frac{U_{.p}^2}{P_{.p}^2} V(\hat{P}_{.p})\right]. \tag{3.3}$$

Le deuxième terme entre crochets étant négligeable par rapport au premier, nous avons

$$V(\hat{R}_{.p}) \cong \frac{1}{P_{.p}^2} \sum_h \frac{U_{hp}(N_h - U_{hp})}{n_h}, \tag{3.4}$$

où N_h est la taille de la strate h, et n_h est la taille de l'échantillon dans la strate h. Cette expression ne tient pas compte du facteur de correction pour population finie. Pour ce qui suit, nous supposerons qu'il n'y a pas de non-réponse et nous supposerons qu'il y a une seule strate par province de sélection (pas de stratification selon la base, l'âge, le sexe, *etc.*). Cette dernière supposition sera bien sûr abandonnée à la section 6 qui traite de la répartition de l'échantillon aux strates infraprovinciales. Pour compenser les effets de la stratification infraprovinciale et de la non-réponse, nous introduisons un effet de plan, D_h. Nous supposons que cet effet de plan ne varie qu'avec la strate h; en particulier, le même effet de plan est utilisé pour exprimer la variance de l'estimateur du nombre de personnes sélectionnées dans la strate h qui sont omises dans la province p, quel que soit p. Une approximation de la variance (3.4) peut être donnée par

$$V(\hat{R}_{.p}) \cong \frac{1}{P_{.p}^2} \sum_{h=1}^{10} \frac{D_h U_{hp}(N_h - U_{hp})}{n_h}, \tag{3.5}$$

et

$$V(\hat{U}_{.p}) \cong \sum_{h=1}^{10} \frac{D_h U_{hp}(N_h - U_{hp})}{n_h}, \tag{3.6}$$

où, cette fois, les sommations sont faites sur les provinces de sélection.

Objectif I:

À partir de l'équation (3.5), nous obtenons :

$$V(\hat{R}_{..}) \cong \frac{1}{P_{..}^2} \sum_{h=1}^{10} \frac{D_h U_{h.}(N_h - U_{h.})}{n_h}, \tag{3.7}$$

où $P_{..} = \sum_{p=1}^{10} P_{.p}$, $\hat{R}_{..} = \hat{U}_{..} / \hat{P}_{..}$, $U_{..} = \sum_{p=1}^{10} U_{.p}$ et $U_{h.} = \sum_{p=1}^{10} U_{hp}$. Cette variance du taux national estimé de sous dénombrement sera réduite au minimum si n_h est proportionnel à $\sqrt{D_h U_{h.}(N_h - U_{h.})} = N_h \sqrt{D_h R_{h.}(1 - R_{h.})}$ où $R_{h.} = U_{h.} / N_h$. Par conséquent, la répartition optimale pour l'objectif I d'un échantillon de taille totale n_I est :

$$n_{p\mathrm{I}} = n_{\mathrm{I}} \left[\frac{N_p \sqrt{D_p R_{p.}(1 - R_{p.})}}{\sum_{p=1}^{10} N_p \sqrt{D_p R_{p.}(1 - R_{p.})}} \right] \quad p = 1, \ldots, 10. \quad (3.8)$$

Cette formule représente une amélioration par rapport à celle utilisée pour la CVD de 2001 (voir Clark (2000), où aucun effet du plan de sondage n'était appliqué à la partie de l'échantillon répartie de manière à produire la meilleure estimation au niveau du Canada. En outre, pour la CVD de 2001, n_p était proportionnel à la population prévue dans la province p. Il est correct que n_p dépende plutôt de la taille des bases de sondage provinciales; il est aussi correct qu'il dépende de la répartition provinciale du sous-dénombrement.

Objectif II:

Nous pouvons utiliser l'équation (3.5) pour déterminer quelles valeurs de n_h donnent la même variance pour les taux provinciaux estimés de sous-dénombrement. Ce problème a dix équations à dix inconnues. Une deuxième difficulté consiste à obtenir des estimations suffisamment précises des U_{hp} pour $p \neq h$, spécialement si p est une petite province. Bien qu'il soit souvent raisonnable de croire que le taux de personnes omises dans une petite province p, $R_{.p} = U_{.p} / P_{.p}$, qui est observé lors d'un recensement soit une bonne prédiction du taux qui sera observé au prochain recensement, les valeurs individuelles des U_{hp} pour $p \neq h$ sont plus difficiles à estimer et encore plus à prédire. Nous supposerons plutôt que $U_{hp} = 0$ pour $p \neq h$ et que $U_{pp} = U_{.p}$, ce qui aura pour effet de mitiger l'effet de données aberrantes sur les variances attendues. Les estimations provinciales du taux de sous-dénombrement seront alors de variance égale, si n_h, pour $h = p$, est proportionnel à $(1/P_{.p}^2) D_p U_{.p}(N_p - U_{.p}) = D_p R_{.p}(N_p / P_{.p} - R_{.p})$. Par conséquent, la répartition optimale pour l'objectif II de l'échantillon de taille totale n_{II} est :

$$n_{p\mathrm{II}} = n_{\mathrm{II}} \left[\frac{D_p R_{.p}(N_p / P_{.p} - R_{.p})}{\sum_{p=1}^{10} D_p R_{.p}(N_p / P_{.p} - R_{.p})} \right] \quad p = 1, \ldots, 10. \quad (3.9)$$

Il convient de souligner que pour la CVD de 2001, pour la partie de l'échantillon répartie de manière à assurer l'égalité sur le plan de la précision des estimateurs provinciaux, les tailles des échantillons ont été fixées proportionnellement à $D_p \hat{R}_{.p}(1 - \hat{R}_{.p})$ (voir Clark (2000)). Utiliser $N_p / P_{.p}$ au lieu de 1, permet de tenir compte des unités figurant dans la base de sondage de la province qui quittent la population de la province, ainsi que des unités de la population de la province qui ne font pas partie de la base de sondage pour cette province, de sorte que l'effet de plan n'est là que pour expliquer la non-réponse et le plan d'échantillonnage. En 2001, une correction pour les unités sortant de la population était apportée par le biais de l'effet de plan, et aucune correction n'a été apportée pour les unités de la population ne faisant pas partie de la base.

Objectif III:

L'estimation de la population totale des provinces bénéficiaires a une variance égale à

$$V(\hat{P}_{.\mathrm{ben}}) = V(\hat{U}_{.\mathrm{ben}}) \cong \sum_{h=1}^{10} \frac{D_h U_{h\mathrm{ben}}(N_h - U_{h\mathrm{ben}})}{n_h}, \quad (3.10)$$

où $P_{.\mathrm{ben}} = \sum_{p=1}^{8} P_{.p}$, $U_{h\mathrm{ben}} = \sum_{p=1}^{8} U_{hp}$ et $U_{.\mathrm{ben}} = \sum_{p=1}^{8} U_{.p}$ sont des sommes sur les huit provinces bénéficiaires (on suppose que les provinces bénéficiaires sont numérotées avec $p = 1,\ldots, 8$, et que les provinces non bénéficiaires sont numérotées avec $p = 9, 10$). L'équation (3.10) est réduite au minimum si n_h, pour $h = 1, \ldots, 10$, est proportionnel à $\sqrt{D_h U_{h\mathrm{ben}}(N_h - U_{h\mathrm{ben}})} = N_h \sqrt{D_h R_{h\mathrm{ben}}(1 - R_{h\mathrm{ben}})}$ où $R_{h\mathrm{ben}} = U_{h\mathrm{ben}} / N_h$. Par conséquent, la répartition optimale pour l'objectif III d'un échantillon de taille totale n_{III} est :

$$n_{p\mathrm{III}} = n_{\mathrm{III}} \left[\frac{N_p \sqrt{D_p R_{p\mathrm{ben}}(1 - R_{p\mathrm{ben}})}}{\sum_{p=1}^{10} N_p \sqrt{D_p R_{p\mathrm{ben}}(1 - R_{p\mathrm{ben}})}} \right]$$

$$p = 1, \ldots, 10. \quad (3.11)$$

Signalons qu'étant donné que les unités sélectionnées dans une province peuvent être classées dans une autre province, $R_{p\mathrm{ben}}$ et $n_{p\mathrm{III}}$ ne sont pas nécessairement nuls lorsque p est une province non bénéficiaire.

Objectif IV:

À partir de l'équation (3.6), nous obtenons :

$$CV(\hat{P}_{.p}) \cong \frac{1}{P_{.p}} \sqrt{\sum_{h=1}^{10} \frac{D_h U_{hp}(N_h - U_{hp})}{n_h}}. \quad (3.12)$$

Nous pouvons utiliser cette formule pour déterminer quelles valeurs de n_h donnent le même coefficient de variation pour les estimations de la population de chaque

province bénéficiaire. Ce problème a huit équations à huit inconnues. Ici aussi, une deuxième difficulté consiste à obtenir des estimations suffisamment précises des U_{hp} pour $p \neq h$, spécialement si p est une petite province. Comme nous l'avons fait pour l'objectif II, nous supposerons plutôt que $U_{hp} = 0$ pour $p \neq h$ et que $U_{pp} = U_{.p}$. Les coefficients de variation des estimations de la population des provinces bénéficiaires seront alors égaux si n_h, pour $h = p$, est proportionnel à $(1/P_{.p}^2)\, D_p U_{.p}(N_p - U_{.p}) = D_p R_{.p}(N_p / P_{.p} - R_{.p})$. Par conséquent, la répartition optimale pour l'objectif IV de l'échantillon de taille totale n_{IV} est :

$$n_{p\mathrm{IV}} = n_{\mathrm{IV}} \left[\frac{D_p R_{.p}(N_p / P_{.p} - R_{.p})}{\sum_{p=1}^{8} D_p R_{.p}(N_p / P_{.p} - R_{.p})} \right] \quad p = 1, \ldots, 8 \quad (3.13)$$

avec les deux provinces non bénéficiaires ayant $n_{p\mathrm{IV}} = 0$, $p = 9,10$.

Il convient de signaler que $n_{p\mathrm{II}} / n_{p\mathrm{IV}}$ est constant pour les huit provinces bénéficiaires. Cela montre un important chevauchement de l'objectif II (précision égale des taux provinciaux estimés de sous-dénombrement), qui est un objectif traditionnel de la répartition de l'échantillon de la CVD, et de l'objectif IV (précision égale des paiements de péréquation par personne des provinces bénéficiaires). Comme nous le constaterons à la section 5, $n_{p\mathrm{I}} / n_{p\mathrm{III}}$, pour les huit provinces bénéficiaires, est pratiquement constant également. Il en ressort un important chevauchement de l'objectif I (précision maximale du taux national estimé de sous-dénombrement), un objectif traditionnel de la répartition de l'échantillon de la CVD, et de l'objectif III (précision maximale du total des paiements de péréquation).

4. Effet de plan

L'erreur-type des estimations de la CVD de 2001 a été calculée par le système généralisé d'estimation (SGE). Cette erreur-type prend en compte le plan de sondage de la CVD et la non-réponse à l'enquête en supposant que les répondants sont sélectionnés au moyen d'un plan de sondage à plusieurs degrés. Le tableau 4.1 présente une comparaison de l'erreur-type obtenue de (3.6) avec celle calculée par le SGE. Pour cette comparaison, un effet de plan de sondage égal à l'inverse du cube du taux de réponse de la province de sélection a été utilisé.

Nous voyons que l'erreur-type pour l'Île-du-Prince-Édouard dérivée de (3.6) est supérieure de 39 % à celle calculée par le SGE; ceci est attribuable à une observation aberrante qui a un effet plus important sur l'estimation à partir de (3.6) que sur l'estimation du SGE. Pour la plupart des provinces, l'erreur-type (3.6) est proche de celle calculée par le SGE. Ces résultats empiriques montrent donc que l'effet de plan de sondage des équations (3.5) et (3.6) est approximativement égal à l'inverse du cube du taux de réponse. Ceci semble indiquer qu'un échantillon de « n » unités avec un taux de réponse « r » correspond à un échantillon de $n \times r^3$ unités plutôt qu'à la taille attendue de $n \times r$ unités, à cause de la concentration des non-répondants parmi les personnes omises par le recensement. Le SGE tient compte que les personnes omises sont moins susceptibles de répondre. Cette érosion de la précision due à la non-réponse se produit même si le plan d'échantillonnage est stratifié de façon plus efficace que le plan d'échantillonnage d'une strate par province que nous avons supposé.

Tableau 4.1
Comparaison de l'erreur-type

Province	Taux de réponse	D = (taux de réponse)$^{-3}$	Erreur-type de l'estimation des omis de (3.6)	Erreur-type de l'estimation des omis du SGE	E.-T. (3.6) / E.-T. GES
T.-N.-L.	0,97	1,08	1 783	1 689	1,06
Î.-P.-É.	0,97	1,09	1 021	734	1,39
N.-É.	0,95	1,15	3 903	3 955	0,99
N.-B.	0,96	1,13	3 272	3 229	1,01
Qc	0,95	1,17	19 915	19 664	1,01
Ont.	0,92	1,28	31 502	31 602	1,00
Man.	0,95	1,15	4 762	5 115	0,93
Sask.	0,96	1,12	3 921	3 840	1,02
Alb.	0,93	1,25	10 493	10 505	1,00
C.-B.	0,91	1,34	14 619	14 763	0,99
Can.	0,94	1,20	42 074	42 041	1,00

Aucune étude semblable n'a été faite pour comparer l'effet de plan et le taux de non-réponse lors des CVD précédentes. La méthode du rajustement des poids pour compenser la non-réponse est différente, et la nature même de la non-réponse est significativement différente de ce qu'elle était avant 2001.

5. Répartition finale de l'échantillon au niveau des provinces et exemple

Le tableau 5.1 montre les valeurs des paramètres qui seront utilisées dans l'exemple. Les valeurs de $\hat{N}_p$ sont des projections de la taille de la base de sondage de la CVD pour 2006; les autres paramètres sont fondés sur les données de la CVD de 2001.

Comme nous pourrions nous y attendre, les valeurs de $\hat{R}_{p\,\mathrm{ben}}$ en Alberta et en Ontario montrent que seulement un petit nombre des unités sélectionnées dans ces deux provinces sont classées comme omises par le recensement dans les provinces bénéficiaires.

La taille de l'échantillon final attribué à la province p est simplement

$$n_p = \max(n_{p\mathrm{I}}, n_{p\mathrm{II}}, n_{p\mathrm{III}}, n_{p\mathrm{IV}}) \quad p = 1, \ldots, 10. \quad (5.1)$$

Qu'on utilise le maximum des 4 tailles comme dans (5.1), une moyenne arithmétique pondérée, ou une moyenne géométrique pondérée, chaque méthode fait usage de 4 paramètres arbitraires (3 si la taille totale d'échantillon est fixe). Pour la méthode du maximum, des valeurs relatives plus élevées de n_{I} (respectivement n_{II}, n_{III} et n_{IV}), donnent une plus grande importance relative à l'objectif I (respectivement II, III et IV).

Le tableau 5.2 donne un exemple avec $n_{\mathrm{I}} = 30,000$, $n_{\mathrm{II}} = 64,000$, $n_{\mathrm{III}} = 25,000$ et $n_{\mathrm{IV}} = 48,078$.

La taille totale de l'échantillon qui résulte est 70 028. Les chiffres en caractères gras sont égaux au maximum sur les quatre répartitions, n_p. De petits changements apportés à n_{III} n'auraient un effet que sur la répartition finale pour le Québec. Cela indique qu'avec les tailles d'échantillon n_{I}, n_{II}, n_{III} et n_{IV} choisies ci-dessus, la taille finale de l'échantillon attribué au Québec est dictée par l'objectif III : une estimation précise du paiement de péréquation totale. De même, la taille finale de l'échantillon attribué à l'Ontario est dictée par l'objectif I : une estimation précise du taux national de sous-dénombrement. La taille finale de l'échantillon attribué à l'Alberta est dictée par l'objectif II : variances égales pour le taux estimé de sous-dénombrement de chaque province. Les tailles finales des échantillons des autres provinces sont aussi bien dictées par l'objectif II que par l'objectif IV (précision égale du paiement de péréquation estimé par personne). Comme nous l'avons signalé à la section 3, $n_{p\mathrm{II}} / n_{p\mathrm{IV}}$ est constant pour les huit provinces bénéficiaires. Dans l'exemple ci-dessus, à cause du choix « judicieux » de n_{IV}, la valeur de la constante est un. Une diminution apportée à n_{IV} entraînerait une diminution de la taille finale de l'échantillon de l'Alberta, mais non de ceux des autres provinces. Nous observons également que $n_{p\mathrm{I}} / n_{p\mathrm{III}}$ ne varie pas beaucoup pour les huit provinces bénéficiaires. L'ajout des objectifs III et IV (se rapportant aux paiements de péréquation) permet de contrôler séparément la taille de l'échantillon du Québec et de celui de l'Alberta. Lorsque seuls les objectifs I et II étaient utilisés, la taille de l'échantillon du Québec avait tendance à être étroitement liée à celle de l'Ontario tandis que la taille de l'échantillon de l'Alberta était étroitement liée à celle des échantillons des autres provinces.

Tableau 5.1
Valeurs des paramètres

Province	$\hat{N}_p$	D_p	$\hat{P}_{.p}$	$\hat{R}_{.p}$	$\hat{R}_{p.}$	$\hat{R}_{p\mathrm{ben}}$
T.-N.-L.	551 987	1,0804	524 722	0,0339	0,0464	0,0368
Î.-P.-É.	145 173	1,0882	132 473	0,0334	0,0334	0,0307
N.-É.	995 651	1,1527	947 099	0,0492	0,0464	0,0440
N.-B.	797 488	1,1345	736 129	0,0493	0,0466	0,0440
Qc	8 079 167	1,1740	7 381 352	0,0510	0,0471	0,0460
Ont.	13 423 132	1,2752	11 702 797	0,0653	0,0565	0,0017
Man.	1 262 547	1,1558	1 136 146	0,0466	0,0437	0,0392
Sask.	1 082 238	1,1223	996 562	0,0437	0,0430	0,0402
Alb.	3 373 128	1,2478	3 010 105	0,0490	0,0403	0,0028
C.-B.	4 570 444	1,3369	4 014 502	0,0761	0,0669	0,0620
Can.	34 280 955	1,2039	30 581 887	0,0587	0,0524	0,0258

Tableau 5.2
Répartition de l'échantillon au niveau des provinces avec
$n_{\mathrm{I}} = 30\,000$, $n_{\mathrm{II}} = 64\,000$, $n_{\mathrm{III}} = 25\,000$, et $n_{\mathrm{IV}} = 48\,078$

Province	$n_{p\mathrm{I}}$	$n_{p\mathrm{II}}$	$n_{p\mathrm{III}}$	$n_{p\mathrm{IV}}$	n_p	$n_{p\mathrm{I}} / n_{p\mathrm{III}}$
T.-N.-L.	427	**3 816**	546	3 816	3 816	0,78
Î.-P.-É.	96	**3 956**	132	3 956	3 956	0,73
N.-É.	796	**5 822**	1 107	5 822	5 822	0,72
N.-B.	634	**5 921**	881	5 921	5 921	0,72
Qc	6 562	6 399	**9 262**	6 399	9 262	0,71
Ont.	**12 385**	9 220	3 148	0	12 385	3,93
Man.	982	**5 867**	1 331	5 867	5 867	0,74
Sask.	823	**5 234**	1 139	5 234	5 234	0,72
Alb.	2 622	**6 702**	1 015	0	6 702	2,58
C.-B.	4 673	**11 063**	6 440	11 063	11 063	0,73
Total	30 000	64 000	25 000	48 078	70 028	

Grâce à une méthode de répartition qui utilise (5.1) et un tableau comme le tableau 5.2, on peut voir clairement la raison d'être de la taille de l'échantillon d'une province. Par exemple, en examinant la répartition finale de l'échantillon au tableau 5.2, si on estime que 5 867 observations au Manitoba ne sont pas suffisantes, alors il faut préciser l'objectif pour lequel elles ne sont pas suffisantes. S'il s'agit d'apporter une amélioration pour l'objectif II (ou l'objectif IV), alors il faut également accroître la taille de l'échantillon dans toutes les provinces de l'Atlantique et dans toutes les provinces de l'Ouest (ou dans toutes les provinces de l'Atlantique et dans toutes les provinces de l'Ouest sauf l'Alberta).

6. Répartition infraprovinciale de l'échantillon

Bien que l'équation (3.5) permette de voir que la répartition infraprovinciale de l'échantillon dans une province de sélection affecte la variance des estimations pour les autres provinces, nous n'essaierons d'optimiser la répartition dans une province que pour l'estimation dans cette même province. Autrement dit, notre problème pour chaque province p, consiste à minimiser

$$\sum_{h \in \left\{\begin{array}{l}\text{strates de la province}\\ \text{de sélection } p\end{array}\right\}} \frac{D_h U_{hp}(N_h - U_{hp})}{n_h} \qquad (6.1)$$

sous la contrainte

$$\sum_{h \in \left\{\begin{array}{l}\text{strates de la province}\\ \text{de sélection } p\end{array}\right\}} n_h = n_p ,$$

où n_p est une taille totale d'échantillon déterminée antérieurement pour la province p. Notons que la taille d'échantillon allouée à la base des personnes omises est fixe, si bien que nous ne tiendrons pas compte dans la suite, des strates de cette base de sondage et que n_p n'inclut pas la taille de l'échantillon tiré de la base des personnes omises. La solution de ce problème de minimisation est

$$n_{h^*} = n_p \frac{\sqrt{D_{h^*} U_{h^* p}(N_{h^*} - U_{h^* p})}}{\sum_{h \in \left\{\begin{array}{l}\text{strates de la province}\\ \text{de sélection } p\end{array}\right\}} \sqrt{D_h U_{hp}(N_h - U_{hp})}} \qquad (6.2)$$

pour chaque strate h^* de la province de sélection p.

Il a été vu à la section 4 que des données empiriques au niveau provincial montrent que le facteur D_h est inversement proportionnel au cube du taux de réponse à la CVD. Pour la répartition de l'échantillon de 2001, on a supposé que D_h variait comme l'inverse du taux de réponse. Pour limiter le déplacement de l'échantillon comparativement à celui de 2001, c'est-à-dire passer de strates à taux de réponse élevé, comme celles de la base du recensement et de la base des naissances, à des strates à taux de réponse faible, comme celles de la base des immigrants ou de la base des résidents non permanents, nous rendons D_h inversement proportionnel au carré du taux de réponse dans la strate h. Notons qu'à la différence de ce qu'on a dû supposer à la section 3, ici le facteur D_h sert à compenser la non-réponse seulement; il n'a pas à compenser pour la stratification puisqu'il est défini au niveau de la strate. Ceci milite aussi en faveur d'un facteur inférieur à l'inverse du cube du taux de réponse.

Comme pour la répartition de l'échantillon de 2001, nous devons résoudre le problème de la projection fiable des valeurs de 2006 de U_{hp} et D_h pour chaque strate h. Puisque les bases des naissances, des immigrants et des résidents non permanents ne contiennent chacune qu'une seule strate par province, nous proposons d'utiliser pour ces strates leurs tailles de 2006, ainsi que les taux de réponse et de personnes omises de 2001, en apportant, si nécessaire, des corrections spéciales pour les provinces les moins peuplées. Une procédure comparable peut être utilisée pour les strates des réserves indiennes de la base du recensement. Les autres strates de la base du recensement sont formées à partir du sexe, de l'état matrimonial (marié(e), non marié(e)), et du groupe d'âge des unités. Pour ces strates, en utilisant les mêmes groupes d'âge selon le sexe et l'état matrimonial, il serait possible, pour chaque province, de caler par ajustement proportionnel itératif (raking), les projections *nationales* sur des valeurs de marge données par des projections *provinciales*, et d'utiliser les valeurs ainsi calées dans l'équation (6.2). Plus précisément, pour calculer les projections de U_{hp} pour toutes les strates h de la province de sélection p, nous commençons par calculer, en nous fondant sur les taux estimés de 2001 et les tailles des strates de 2006, une projection du nombre de personnes omises, classées dans la province où elles ont été sélectionnées, pour chaque cellule (sexe × état matrimonial × groupe d'âge). Ces chiffres *nationaux* pourraient occuper les cellules d'une matrice tridimensionnelle. Toujours en nous fondant sur les taux estimés de 2001 et les tailles des strates de 2006, nous calculons ensuite une projection du nombre de personnes omises, classées dans la *province p*, pour toutes les strates de la province selon le sexe, puis pour toutes les strates de la province selon l'état matrimonial, et enfin pour toutes les strates de la province selon le groupe d'âge. Ces chiffres nous donneraient les totaux de marge souhaités de la matrice tridimensionnelle. Par ajustement proportionnel itératif, nous pourrions obtenir les projections pour U_{hp} dont la somme correspond aux totaux provinciaux souhaités selon le sexe, selon l'état matrimonial et selon le groupe d'âge. Afin d'éviter les problèmes de convergence, de simplifier la programmation et de rendre le processus plus flexible, l'ajustement proportionnel itératif (raking) sera remplacé par la résolution d'un problème de calage. En fait, la plus

grande flexibilité est nécessaire dans notre cas, car les groupes d'âge ne sont pas les mêmes pour les personnes mariées et non mariées.

Voici maintenant un exemple d'utilisation du calage. La méthode s'inspire du résultat suivant donné dans Théberge (1999) :

Soient $\mathbf{U}$ et $\mathbf{T}$ des matrices diagonales positives de dimension n et q respectivement, $\mathbf{w}_0$ un vecteur de dimension n, $\mathbf{A}$ une matrice de dimensions $q \times n$ et $\mathbf{b}$ un vecteur de dimension q; alors, parmi les vecteurs de poids $\mathbf{w}$ de dimension n qui minimisent $\| \mathbf{Aw} - \mathbf{b} \|_{\mathbf{T}}^2$, l'unique vecteur de poids qui minimise $\| \mathbf{w} - \mathbf{w}_0 \|_{\mathbf{U}}^2$ est donné par

$$\mathbf{w} = \mathbf{w}_0 + \mathbf{U}^{-1}\mathbf{A}'\,\mathbf{T}^{1/2}(\mathbf{T}^{1/2}\mathbf{A}\mathbf{U}^{-1}\mathbf{A}'\mathbf{T}^{1/2})^{\dagger}\mathbf{T}^{1/2}(\mathbf{b} - \mathbf{A}\mathbf{w}_0), \quad (6.3)$$

où $\| \mathbf{z} - \mathbf{z}_0 \|_{\mathbf{F}}^2 = (\mathbf{z} - \mathbf{z}_0)'\mathbf{F}(\mathbf{z} - \mathbf{z}_0)$ est une mesure de distance pondérée entre $\mathbf{z}$ et $\mathbf{z}_0$, et $\mathbf{G}^{\dagger}$ est l'inverse de Moore-Penrose de $\mathbf{G}$.

L'équation $\mathbf{Aw} = \mathbf{b}$ donne l'ensemble de q contraintes de calage. Nous fixons la valeur de $\mathbf{T}$ à celle de la matrice identité dans l'équation (6.3). Si les contraintes peuvent être satisfaites, la matrice $\mathbf{T}$ n'est pas pertinente; sinon, attribuer à $\mathbf{T}$ la valeur de la matrice identité revient à accorder la même importance à chacune des q contraintes lorsqu'on minimise la distance entre $\mathbf{Aw}$ et $\mathbf{b}$.

Dans le cas qui nous préoccupe, pour projeter le nombre de personnes omises dans chaque strate d'une province donnée, nous avons $\mathbf{A} = \mathbf{MX}$ avec

$$\mathbf{M} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix},$$

$$\mathbf{X} = \operatorname{diag}\begin{pmatrix} x_{FN0-14} \\ x_{FN15-24} \\ x_{FN25-44} \\ x_{FN45+} \\ x_{FM25-34} \\ x_{FM35+} \\ x_{MN0-14} \\ x_{MN15-24} \\ x_{MN25-44} \\ x_{MN45+} \\ x_{MM25-34} \\ x_{MM35+} \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_{FN0-14} \\ w_{FN15-24} \\ w_{FN25-44} \\ w_{FN45+} \\ w_{FM25-34} \\ w_{FM35+} \\ w_{MN0-14} \\ w_{MN15-24} \\ w_{MN25-44} \\ w_{MN45+} \\ w_{MM25-34} \\ w_{MM35+} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_{F..} \\ b_{M..} \\ b_{.N.} \\ b_{.M.} \\ b_{..0-14} \\ b_{..15-24} \\ b_{..25+} \end{pmatrix},$$

où, par exemple, $x_{MN25-44}$ est le nombre de personnes omises, classées dans la province où elles ont été sélectionnées, dans les strates d'hommes non mariés de 25 à 44 ans, $w_{MN25-44}$ est le poids recherché pour cette strate et, par exemple, $b_{.N.}$ est le nombre de personnes omises, sélectionnées et classées dans la province, qui appartiennent aux strates de personnes « non mariées ». Toutes les personnes de 0 à 24 ans sont classées dans les strates de personnes « non mariées », quel que soit leur état matrimonial réel. Notons que tant dans le calcul des chiffres nationaux, $\mathbf{X}$, que dans le calcul des chiffres provinciaux, $\mathbf{b}$, on ne comptent que les personnes n'ayant pas changé de province, afin d'être consistant avec l'objectif mentionné au début de cette section.

Poursuivons le parallèle avec l'ajustement proportionnel itératif. La matrice $\mathbf{X}$ donne les valeurs de la matrice tridimensionnelle pour laquelle l'ajustement itératif doit être fait, sauf que les éléments sont présentés dans une matrice diagonale; le vecteur $\mathbf{w}$ donne les « facteurs d'ajustement proportionnel itératif » finaux appliqués aux éléments de $\mathbf{X}$ pour obtenir les valeurs calées, $\mathbf{Xw}$; la contrainte est que des sommes de ces éléments calés, $\mathbf{MXw}$, doivent s'approcher autant que possible des « valeurs de marge » voulues données par le vecteur $\mathbf{b}$; et $\mathbf{w}$ doit s'approcher autant que possible du vecteur $\mathbf{w}_0$ décrit ci-dessous.

Nous pouvons choisir le vecteur $\mathbf{w}_0$ de sorte que chaque élément soit égal à un facteur constant qui réduira les chiffres nationaux à des chiffres plus appropriés pour la province. Cela si nous voulons que la somme des chiffres nationaux pondérés corresponde aux totaux de marge provinciaux, en utilisant des poids aussi proches que possible d'une constante, afin de maintenir la répartition nationale qui est plus fiable. La répartition nationale des personnes omises peut ne pas convenir, si la répartition des tailles de strate n'est pas la même pour le Canada dans son ensemble que pour la province. Par conséquent, une meilleure solution consiste à fixer l'élément de $\mathbf{w}_0$ correspondant à la strate h^* à

$$w_{0h^*} = N_{h^*} \Big/ \sum_{h \in S_{h^*}} N_h , \quad (6.4)$$

où S_{h^*} est l'ensemble des 10 strates (une par province) similaires à la strate h^* (par exemple, les 10 strates d'hommes non mariés de 15 à 24 ans).

Nous pourrions supprimer deux contraintes parce que les rangées correspondantes de $\mathbf{M}$ sont des combinaisons linéaires des autres (par exemple, quatrième et dernière rangées), mais la solution (6.3) est suffisamment générale pour qu'il ne soit pas nécessaire de le faire. Avec $\mathbf{A} = \mathbf{MX}$, $\mathbf{U} = \mathbf{X}$ et $\mathbf{T}$ égal à la matrice identité, (6.3) se simplifie en

$$\mathbf{w} = \mathbf{w}_0 + \mathbf{M}'(\mathbf{MXM}')^{\dagger}(\mathbf{b} - \mathbf{MXw}_0). \quad (6.5)$$

Les valeurs lissées pour chaque strate sont les éléments du vecteur $\mathbf{Xw}$.

Un problème similaire peut être posé pour les non-répondants, lorsqu'on veut lisser les effets du plan de sondage.

7. Conclusion

Il y a beaucoup de chevauchement entre les deux objectifs traditionnels de la répartition de l'échantillon de la CVD, qui sont d'obtenir une variance minimale pour le taux national estimé de sous-dénombrement (objectif I) et d'obtenir des variances égales pour les taux de sous-dénombrement estimés de chaque province (objectif II), et les deux objectifs additionnels examinés dans cet article, qui sont de réduire au minimum la variance du paiement de péréquation total (objectif III) et de produire des CV égaux pour l'estimation de la population de chaque province bénéficiaire (objectif IV). Néanmoins, la prise en compte explicite de ces deux objectifs additionnels peut permettre à la taille de l'échantillon pour le Québec et pour l'Alberta de varier indépendamment de ceux des autres provinces. La méthode proposée dans cet article pour réaliser un compromis entre différentes répartitions, optimal en ce qui a trait aux différents objectifs, consiste à prendre, pour chaque province, la taille maximale de l'échantillon sur chacune des répartitions. Cette méthode fournit une justification plus directe de la répartition.

Une comparaison des erreurs types du SGE et des erreurs-types résultant de la formule approximative (3.6) montre que, pour la CVD de 2001, n unités échantillonnées avec un taux de réponse de r sont équivalentes à seulement $n \times r^3$ unités avec réponse complète.

La répartition infraprovinciale optimale exige le lissage des paramètres provinciaux au niveau âge×sexe×état matrimonial. Le calage peut être une méthode commode d'ajuster des valeurs nationales par âge × sexe × état matrimonial qui sont plus stables, de sorte que leur somme soit égale à celle des valeurs provinciales par âge, par sexe et par état matrimonial. Le principal objectif de la méthode rappelle le principal objectif de la méthode itérative du quotient, mais une solution comme celle appliquée dans Théberge (1999), qui traite de la possibilité que les contraintes ne soient pas respectées, permet d'éviter les problèmes de convergence. En outre, l'utilisation de l'inverse de Moore-Penrose permet d'éviter les problèmes de colinéarité.

Bibliographie

Brackstone, G.J., et Rao, J.N.K. (1976). Raking ratio estimators. *Survey Methodology*, 2, 63-69.

Clark, C. (septembre 2000). 2001 Reverse Record Check: Provincial and Territorial sample allocation. Document non publié. Ottawa. Statistique Canada.

Deming, W.E., et Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11, 427, 444.

Deville, J.-C., et Särndal, C.-E. (1992), Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Théberge, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.

Techniques d'enquête, juin 2006
Vol. 32, N° 1, pp. 97-106
Statistique Canada, N° 12-001-XIF au catalogue

Calcul de la taille de l'échantillon pour l'estimation pour petits domaines

Nicholas Tibor Longford [1]

Résumé

Nous décrivons une approche générale de détermination du plan d'échantillonnage des enquêtes planifiées en vue de faire des inférences pour de petits domaines (sous-domaines). Cette approche nécessite la spécification des priorités d'inférence pour les petits domaines. Nous établissons d'abord des scénarios de répartition de la taille de l'échantillon pour l'estimateur direct, puis pour les estimateurs composite et bayésien empirique. Nous illustrons les méthodes à l'aide d'un exemple de planification d'un sondage de la population suisse et d'estimation de la moyenne ou de la proportion d'une variable pour chacun des 26 cantons.

Mots clés : Efficacité; estimation pour petits domaines; priorité d'inférence; répartition de la taille de l'échantillon.

1. Introduction

Le plan d'échantillonnage est un instrument essentiel à la production d'estimations efficaces et d'autres formes d'inférence au sujet d'une grande population, lorsque les ressources disponibles ne permettent pas de recueillir l'information pertinente pour chaque membre de la population. Dans ce contexte, nous interprétons l'efficacité comme étant la combinaison optimale d'un plan d'échantillonnage et d'un estimateur d'un paramètre de population θ. Par optimale, nous entendons que l'erreur quadratique moyenne est minimale, quoique le développement présenté dans l'article puisse être adapté à d'autres critères. Le groupe de plans de sondage possibles est délimité par les ressources et celles-ci sont habituellement exprimées en fonction d'une taille fixe d'échantillon. Cette approche n'est pas toujours appropriée, parce que les coûts moyens par sujet ne sont pas nécessairement les mêmes pour tous les plans d'échantillonnage. Toutefois, si nous considérons une gamme limitée de plans, nous pouvons ignorer ce point.

Le problème de l'établissement du plan d'échantillonnage afin d'estimer efficacement une grandeur unique est bien compris et des solutions existent pour bon nombre de spécifications utilisées fréquemment. La plupart comportent un problème d'optimisation univarié sous contraintes. L'établissement du plan d'échantillonnage pour l'estimation de plusieurs paramètres est considérablement plus complexe, parce que le problème comprend plusieurs facteurs, habituellement un pour chaque paramètre. Il est essentiel d'optimiser le plan simultanément pour tous les facteurs, parce que les objectifs d'inférence efficace au sujet des paramètres cibles peuvent être conflictuels. Par exemple, dans l'estimation pour petits domaines, l'allocation d'une part plus généreuse de la taille de l'échantillon à un petit domaine doit être compensée par une allocation moins généreuse à un ou à plusieurs autres.

Au cours des dernières décennies, la production de statistiques pour des petits domaines est devenue un important sujet de recherche en méthodologie d'enquête (Fay et Herriot 1979; Platek, Rao, Särndal et Singh 1987; Ghosh et Rao 1994; Longford 1999; Rao 2003), étant donné l'intérêt grandissant des organismes gouvernementaux, du secteur de la publicité et du marketing et de celui de la finance et des assurances pour ce genre d'information. À l'heure actuelle, de nombreuses enquêtes à grande échelle sont conçues en vue de produire des estimations de niveau national, mais sont parfois utilisées après coup pour faire des inférences au sujet de petits domaines. Cela n'aurait pas d'inconvénient si les plans d'échantillonnage optimaux pour l'inférence sur petits domaines et l'inférence nationale étaient les mêmes. Nous montrons dans le présent article qu'il n'en est pas ainsi et que le plan d'échantillonnage peut effectivement être ciblé pour l'estimation pour petits domaines, en tenant compte de l'objectif de production d'estimations efficaces de paramètres de niveau national. Pour éviter le cas banal, supposons que les populations des petits domaines soient de taille inégale. Nous appliquons les méthodes au problème de la planification d'inférences au sujet des 26 cantons de la Suisse; la taille de la population de ces cantons varie de 15 000 (Appenzell-Innerrhoden) à 1,23 million (Zürich). La population de la Suisse se chiffre à 7,26 millions d'habitants.

La littérature traitant de la planification des enquêtes pour l'estimation pour petits domaines est peu abondante. L'une des contributions importantes est celle de Singh, Gambino et Mantel (1994). Dans l'une des approches dont discutent ces auteurs, la taille prévue de l'échantillon de l'Enquête sur la population active du Canada est divisée en deux. Une partie est répartie optimalement en vue de la production d'estimations de niveau national (domaine) et l'autre est répartie optimalement en vue de l'estimation pour petits domaines (sous-domaines). Pour ce dernier objectif, des

1. Nicholas Tibor Longford, Departament d'Económia i Empresa, Universitat Pompeu Fabra, Ramón Trias Fargas 25-27, 08005 Barcelone, Espagne. Courriel : NTL@SNTL.co.uk.

sous-échantillons de même taille sont attribués à chaque petit domaine, lorsque les variances dans les sous-domaines sont égales, que la correction pour population finie peut être ignorée et que les coûts d'enquête par sujet sont les mêmes pour tous les sous-domaines, mais aussi quand les paramètres visées par l'inférence sont les moyennes de petit domaine. Si l'on veut estimer des totaux de population, l'équirépartition de l'échantillon entre les sous-domaines n'est pas efficace, parce qu'elle pénalise l'estimation pour les petits domaines les plus peuplés. Même si l'on estime des proportions ou des taux (pourcentages), les variances intradomaine dépendent de la proportion de population, quoique la dépendance soit faible lorsque toutes les proportions sont loin de zéro et de l'unité. Pour des travaux plus récents sur les plans d'échantillonnage pour l'estimation pour petits domaines, voir Marker (2001).

La section suivante décrit l'approche proposée, fondée sur la minimisation de la somme pondérée des variances d'échantillonnage (erreurs quadratiques moyennes) des estimateurs prévus, avec les pondérations spécifiées de façon à refléter les priorités d'inférence. Nous l'appliquons pour commencer à l'estimation directe de paramètres au niveau du petit domaine. Puis, nous l'étendons afin d'intégrer l'objectif de production d'estimations nationales et, enfin, l'estimation composite à la section 3. La section 4, qui conclut l'article, contient une discussion.

La présente section se termine par une description de la notation utilisée dans la suite de l'article. Nous supposons que les paramètres de population au niveau du petit domaine θ_d, $d = 1, ..., D$, sont estimées par $\hat{\theta}_d$ avec des erreurs quadratiques moyennes (EQM) v_d respectives qui sont des fonctions des tailles des sous-échantillons dans les petits domaines n_d; $v_d = v_d(n_d)$. La taille globale de l'échantillon est dénotée par n et nous supposons qu'elle est fixe. Les tailles de population sont dénotées par N (globale) et N_d (pour le petit domaine d). Par souci de concision, nous dénotons $\mathbf{n} = (n_1, ..., n_D)^\mathsf{T}$. La plupart des paramètres de population θ sont des fonctions d'une seule variable, comme la moyenne, le total et ainsi de suite. La variable peut être enregistrée directement durant le sondage ou construite d'après une ou plusieurs variables directes. Bien que notre développement ne soit pas limité à ce genre de paramètres, la justification est plus simple en ce qui les concerne. Nous disons qu'un estimateur de θ_d est *direct* s'il s'agit d'une fonction de la variable étudiée sur les sujets du petit domaine d seulement.

Nous supposons que chaque estimateur direct envisagé est sans biais. Cette hypothèse n'est pas particulièrement restrictive, car la plupart des estimateurs directs sont des estimateurs naïfs ou étroitement reliés à ces derniers. Nous supposons que les tailles d'échantillon pour les petits domaines sont sous le contrôle du concepteur de l'enquête. Il en est ainsi pour les plans d'échantillonnage stratifiés dans lesquels les strates coïncident avec les petits domaines. À la section 4, nous discutons des plans d'échantillonnage pour lesquels ce genre de contrôle ne peut être exercé; ces plans sont particulièrement indiqués pour la subdivision du pays en un grand nombre (centaines) de petits domaines.

2. Plan optimal pour l'estimation directe

Nous résolvons le conflit entre les objectifs d'estimation efficace de paramètres au niveau du petit domaine θ_d en choisissant le plan d'échantillonnage à ce niveau qui minimise la somme pondérée des variances d'échantillonnage (EQM),

$$\min_{\mathbf{n}} \sum_{d=1}^{D} P_d v_d, \qquad (1)$$

sachant que la taille globale d'échantillon $n = \mathbf{n}^\mathsf{T} \mathbf{1}_D$ est fixe; $\mathbf{1}_D$ est le vecteur des unités de longueur D. Le coefficient P_d est nommé *priorité d'inférence*. Une valeur plus grande de P_d (par rapport aux valeurs $P_{d'}$, $d' \neq d$) implique qu'il est plus important de réduire v_d, parce que l'augmentation de la contribution du petit domaine d à la somme (1) est plus importante que pour les autres petits domaines.

Le problème d'optimisation (1) est résolu par la méthode des multiplicateurs de Lagrange, ou simplement par substitution de $n_1 = n - n_2 - ... - n_D$, si bien qu'il comporte alors $D-1$ variables fonctionnellement non corrélées. La solution satisfait la condition

$$P_d \frac{\partial v_d}{\partial n_d} = \text{const.}$$

En général, il n'est pas possible d'obtenir une expression analytique des tailles optimales des sous-échantillons n_d, mais si $v_d = \sigma_d^2 / n_d$, comme dans le cas de l'échantillonnage aléatoire simple à l'intérieur des petits domaines, la solution est proportionnelle à $\sigma_d \sqrt{P_d}$, c'est-à-dire

$$n_d^\dagger = n \frac{\sigma_d \sqrt{P_d}}{\sigma_1 \sqrt{P_1} + ... + \sigma_D \sqrt{P_D}}.$$

Lorsque les variances intra domaine σ_d^2 sont égales, $\sigma_1^2 = ... = \sigma_D^2 = \sigma^2$, la solution se simplifie encore davantage; les tailles optimales d'échantillon sont proportionnelles à $\sqrt{P_d}$ et ne dépendent pas de σ^2.

Dans la plupart des contextes, il est difficile d'exprimer un ensemble approprié de priorités P_d et il est donc plus constructif de proposer une classe paramétrique commode de priorités $\mathbf{P} = (P_1, ..., P_D)^\mathsf{T}$ et d'illustrer son effet sur la répartition de la taille de l'échantillon. Nous proposons les priorités $P_d = N_d^q$ pour $0 \leq q \leq 2$. Si $q = 0$, l'inférence est de même importance pour chaque petit domaine. À mesure

que q augmente, une importance relativement plus grande est accordée aux petits domaines les plus peuplés. Lorsque $v_d = \sigma^2 / n_d$, la répartition optimale de la taille de l'échantillon pour $q = 2$, $n_d^{\dagger} = n N_d / N$ est proportionnelle aux tailles de population dans les petits domaines et le même plan d'échantillonnage est donc optimal pour les inférences calculées au niveau national et du petit domaine. Pour $q > 2$, la répartition de la taille de l'échantillon est encore plus généreuse à l'égard des petits domaines les plus peuplés, aux dépens de ceux qui le sont moins. Comme cette situation est contre-intuitive dans le contexte de l'estimation pour petits domaines, le choix d'un exposant $q > 2$ n'est probablement jamais approprié. Un exposant de priorité q négatif conviendrait pour une enquête dont le but est de se concentrer sur les petits domaines les moins peuplés. Naturellement, ce genre de plan est très inefficace pour l'estimation du paramètre θ de niveau national, surtout si les tailles de population des petits domaines sont très dispersées.

Les priorités d'inférence P_d peuvent être des fonctions d'autres paramètres que N_d. Par exemple, les tailles de certaines sous-population présentant un intérêt particulier, comme une minorité ethnique dans le petit domaine, peuvent être utilisées au lieu de N_d, P_d peut être défini différemment dans les diverses régions du pays, ou bien la formule pour le calculer peut-être outrepassée pour un petit domaine ou quelques-uns d'entre eux.

Dans certains rapports d'analyse de données d'enquête, une estimation n'est publiée que si elle est fondée sur un échantillon de taille suffisamment grande ou que son coefficient de variation (le ratio de l'erreur-type estimée à l'estimation) est inférieur à un seuil spécifié. Si une « pénalité » associée au fait de ne pas publier un paramètre est précisée, elle peut être intégrée dans la définition des priorités d'inférence. La difficulté qui risque de se poser est que la fonction objectif (1) soit discontinue et que l'on ne puisse plus appliquer les approches standard d'optimisation. La pénalité doit être déterminée minutieusement. Si elle est trop faible, elle est inefficace; si elle est trop élevée, la solution accordera la préférence à la publication d'estimations pour un aussi grand nombre de petits domaines que possible, mais avec, pour chacun, une taille d'échantillon ou une précision qui n'excède que de justesse le seuil fixé. Voir Marker (2001) pour une autre approche de ce problème.

La figure 1 illustre l'effet de l'exposant de priorité q sur la répartition de la taille de l'échantillon d'une enquête planifiée en Suisse dans le but d'estimer les moyennes de population d'une variable dans les 26 cantons, en supposant qu'ils ont tous la même variance intracanton σ^2. La taille globale prévue de l'échantillon est $n = 10\,000$. Dans n'importe quel volet, les courbes relient les tailles d'échantillon optimales pour chaque exposant q; elles sont tracées sur l'échelle linéaire (à gauche) et sur l'échelle logarithmique (à droite). Les tailles de population sont inscrites sur la barre horizontale au bas de chaque graphique. Sur l'échelle logarithmique, les courbes sont linéaires. Cette échelle produit aussi une répartition plus uniforme des tailles de population des cantons.

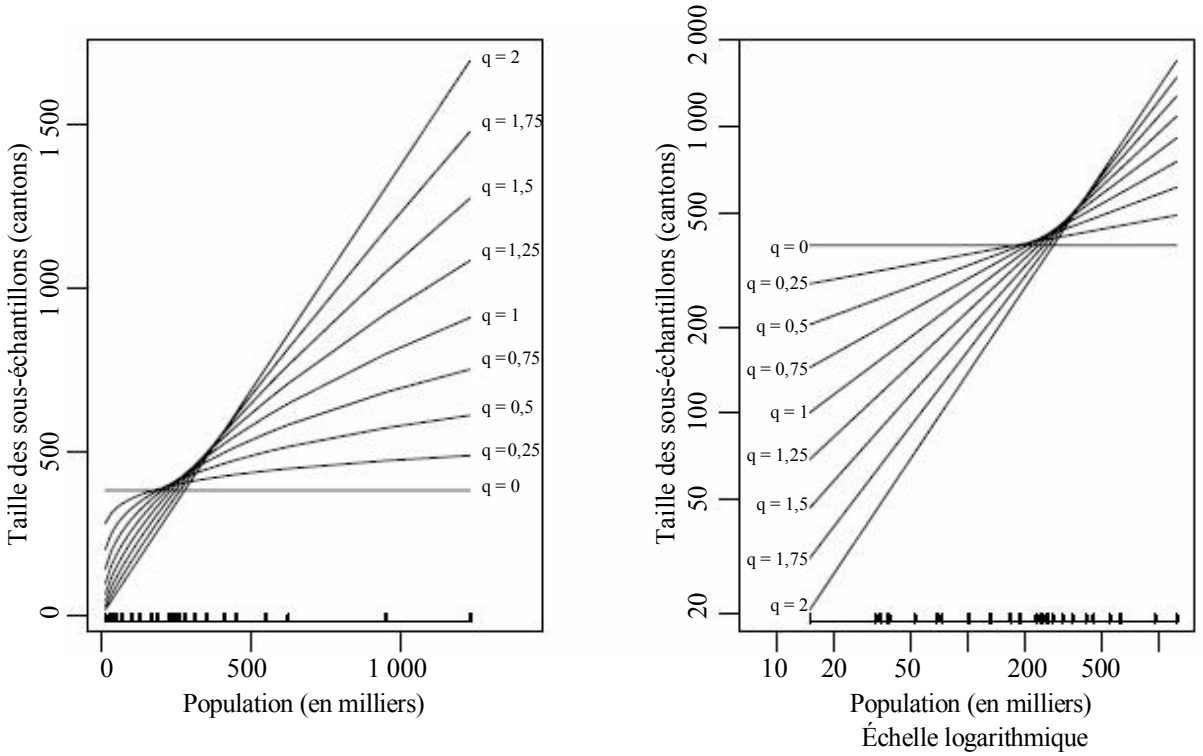


Figure 1. Répartition de la taille de l'échantillon entre les cantons suisses pour une gamme d'exposants de priorité q. Les tailles de population des cantons sont inscrites sur la barre horizontale au bas de chaque graphique.

Pour $q = 0$, une même taille d'échantillon est attribuée à chaque canton, soit $10\,000/26 = 385$, et pour $q = 2$, la répartition est proportionnelle à la taille de population du canton. Pour les valeurs intermédiaires de q, les tailles d'échantillon des cantons les moins peuplés sont augmentées par rapport à la répartition proportionnelle $(q = 2)$, au prix de l'attribution d'une taille réduite aux cantons les plus peuplés. Pour les cantons dont la population est supérieure à 250 000, environ 3 % du chiffre national de population, la taille des sous-échantillons dépend fort peu de la valeur de q.

2.1 Priorité accordée à l'estimation nationale

Comme les tailles de sous-échantillon au niveau du canton diffèrent de la répartition proportionnelle pour l'exposant de priorité $q < 2$, l'estimation optimale au niveau du canton est assortie d'une perte d'efficacité de l'estimateur national. Considérons l'estimateur stratifié

$$\hat{\theta} = \frac{1}{N}\sum_{d=1}^{D} N_d \hat{\theta}_d$$

de la moyenne nationale θ d'une variable, où $\hat{\theta}_d$ représente les estimateurs sans biais des moyennes intracanton de la même variable. En supposant que l'échantillonnage est stratifié avec échantillonnage aléatoire simple dans les strates (cantons) et que la valeur de $\hat{\theta}_d$ est fixée à la moyenne d'échantillon intrastrate,

$$\mathrm{var}(\hat{\theta}) = \frac{1}{N^2}\sum_{d=1}^{D} \frac{N_d^2}{n_d}(1 - f_d)\sigma_d^2,$$

où $f_d = n_d / N_d$ est la correction pour population finie.

La figure 2 représente la fonction qui relie l'erreur-type $\sqrt{\mathrm{var}(\hat{\theta})}$ à l'exposant de priorité q, calculée en supposant que $\sigma^2 = 100$. L'erreur-type est une fonction décroissante de q; elle diminue plus rapidement à $q = 0$ qu'à $q = 2$, où elle est relativement constante. Pour $q = 2$, les objectifs d'estimation au niveau du canton et au niveau national concordent, et $\sqrt{\mathrm{var}(\hat{\theta})} = 0{,}100$. Pour $q = 0$, $\sqrt{\mathrm{var}(\hat{\theta})} = 0{,}143$; dans ces conditions, l'optimalité de l'estimation pour petits domaines a sur l'estimation nationale un effet défavorable important, équivalant à la réduction de moitié de la taille de l'échantillon $(0{,}143/0{,}100 \doteq \sqrt{2})$. Pour une valeur négative de q, cet effet est encore plus prononcé.

Donc, nous pouvons répondre au besoin d'efficacité de l'estimateur national en augmentant la valeur de l'exposant de priorité. Par exemple, les parties ayant des intérêts concurrents en matière d'inférence pourraient négocier la perte d'efficacité de $\hat{\theta}$ qu'elles jugent acceptables et fixer ensuite l'exposant de priorité de façon à égaler cette perte. Ou bien, la perte pourrait être prise en considération lors de l'application du plan d'échantillonnage optimal pour l'estimation au niveau du petit domaine. Si elle est jugée excessive, q pourrait être augmenté jusqu'à l'obtention d'un équilibre entre les pertes d'efficacité de l'estimation nationale et de celle sur petits domaines.

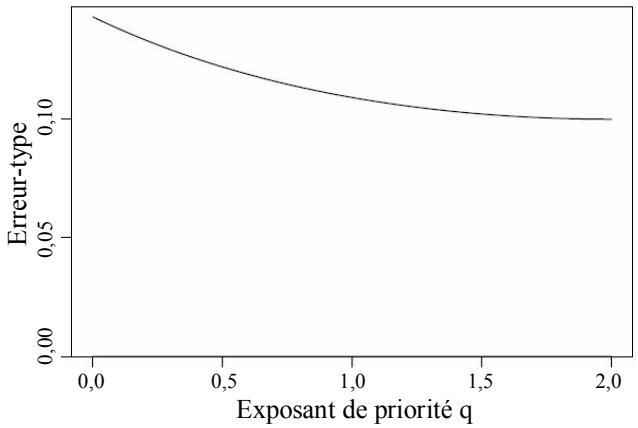


Figure 2. Erreur-type de l'estimateur national $\hat{\theta}$ de la moyenne d'une variable, sous forme de fonction de l'exposant q pour les priorités de l'estimation au niveau du canton.

Un aspect insatisfaisant de ces approches est qu'elles compromettent l'objectif premier des priorités **P**, c'est-à-dire refléter l'importance relative des inférences au sujet de petits domaines distincts. Pour contourner cet inconvénient, nous associons $\hat{\theta}$ à une priorité, dénotée G, relative à une estimation pour petits domaines, et nous considérons l'estimation optimale de l'ensemble de D paramètres cibles au niveau du petit domaine θ_d en même temps que le paramètre cible nationale θ. Donc, nous minimisons la fonction objectif

$$\sum_{d=1}^{D} P_d v_d(n_d) + GP_+ v(\mathbf{n}),$$

où $v = \mathrm{var}(\hat{\theta})$ et $P_+ = \mathbf{P}^{\mathsf{T}}\mathbf{1}_D$. Le facteur P_+ est introduit pour améliorer l'effet des tailles absolues de P_d et du nombre de petits domaines sur la priorité relative G. Les priorités P_d peuvent être interprétées uniquement d'après leurs tailles relatives, car, pour toute constante $c > 0$, P_d et cP_d correspondent à des ensembles identiques de priorités pour l'estimation pour petits domaines dans (1).

Lorsque le plan d'échantillonnage dans chaque petit domaine est aléatoire simple et que $\hat{\theta}$ est l'estimateur stratifié standard, le minimum est atteint quand

$$\sigma_d^2 \frac{P_d'}{n_d^2} = \mathrm{const},$$

où $P_d' = P_d + GP_+ N_d^2 / N^2$. Les tailles optimales d'échantillon pour les petits domaines sont

$$n_d^* = n\frac{\sigma_d\sqrt{P_d'}}{\sigma_1\sqrt{P_1'} + \ldots + \sigma_D\sqrt{P_D'}}.$$

Cette solution correspond à un ajustement des priorités P_d par $GP_+N_d^2/N^2$. Notons que cet ajustement n'est ni additif, ni multiplicatif. L'accroissement de la priorité est plus important pour les petits domaines plus peuplés. Par conséquent, les tailles des sous-échantillons de petit domaine sont réduites davantage quand la priorité relative de l'estimation nationale est intégrée et que les priorités au niveau des petits domaines ne changent pas. La correction pour population finie n'a aucun effet sur n_d^*, parce qu'elle réduit chaque variance d'échantillonnage v_d et v d'une quantité qui ne dépend pas de **n**.

La priorité G peut être fixée en insistant sur le fait que la perte d'efficacité lors de l'estimation de la grandeur nationale θ n'excède pas un pourcentage donné ou qu'au plus, quelques-uns seulement des écarts absolus $|P_d' - P_d|$ ou des logarithmes des ratios $|\log(P_d'/P_d)|$ (voire aucun) ne soient très grands. Cependant, le problème analytique est facile à résoudre, de sorte que la gestion de l'enquête peut être présentée au moyen des plans d'échantillonnage qui sont optimaux pour une gamme de valeur G.

La variation de la taille des sous-échantillons en fonction de l'exposant q et de la priorité relative G est représentée graphiquement à la figure 3 pour les cantons le moins et le plus peuplés, Appenzell-Innerrhoden et Zürich, dans les volets A et C. Les volets B et D donnent la représentation des mêmes courbes qu'A et C, respectivement, sur l'échelle logarithmique. Ne pas tenir compte de l'objectif de production d'une estimation nationale correspond au cas où $G = 0$ et ne pas tenir compte de l'objectif de production d'une estimation pour petits domaines correspond au cas des valeurs très grandes de G. Tout au long de l'article, nous supposons que $n = 10\,000$ et que $\sigma^2 = 100$ pour tous les cantons.

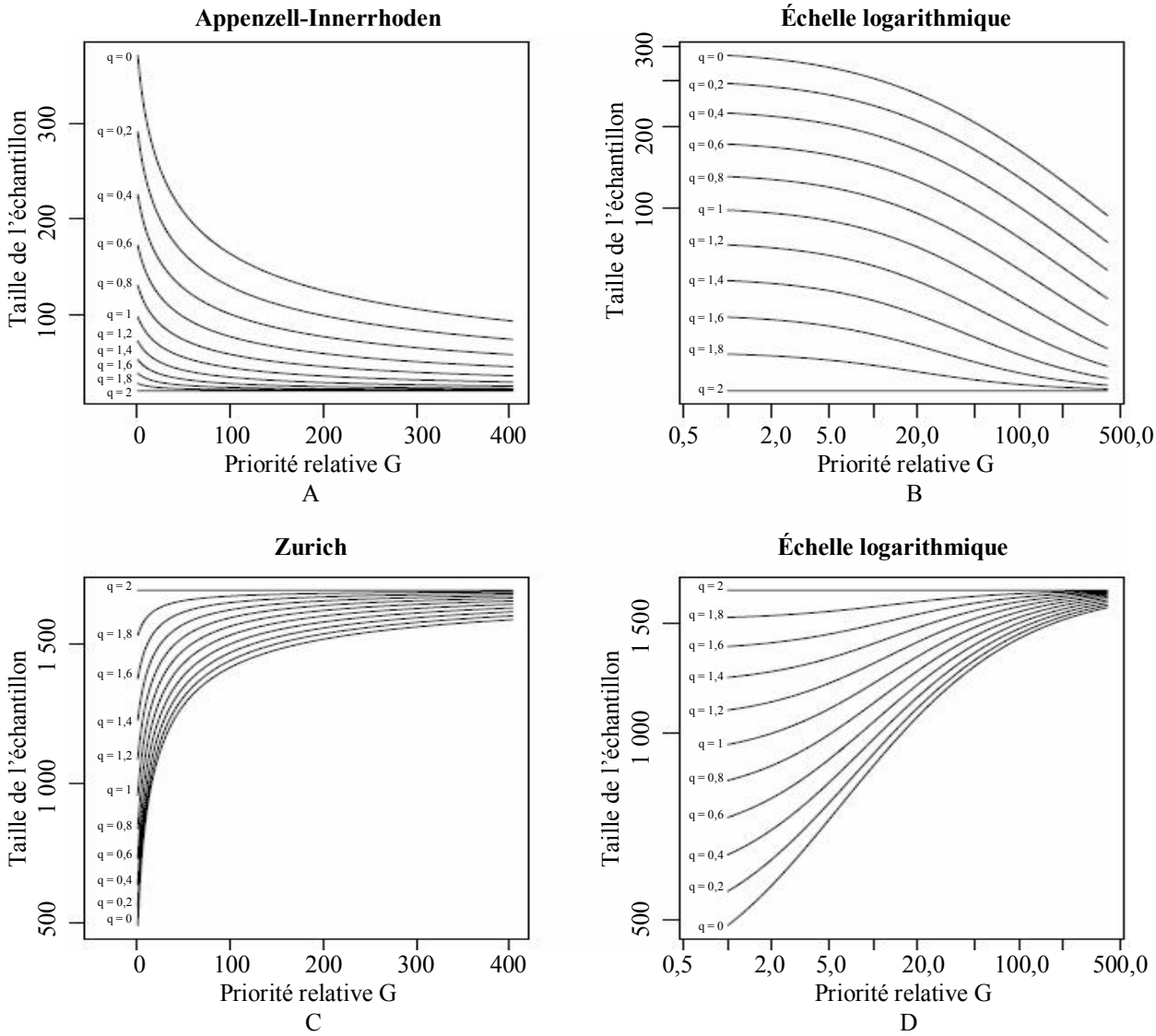


Figure 3. Tailles d'échantillon optimales pour l'estimateur direct $\hat{\theta}_d$ pour les combinaisons d'exposants de priorité q et de priorités relatives G pour les cantons le moins et le plus peuplés.

Dans le cas de chaque exposant $q < 2$, la courbe de répartition de la taille de l'échantillon $n_d(G)$ montre une diminution pour les cantons les moins peuplés et une augmentation pour les plus peuplés en direction de la représentation proportionnelle, $n_d = nN_d / N$, qui correspond à $q = 2$. Sur l'échelle linéaire, l'augmentation est assez rapide pour Zürich pour les faibles valeurs de q et de G, tandis que la réduction pour Appenzell-Innerrhoden est plus progressive. À mesure que la priorité relative G est réduite, la taille d'échantillon excédentaire est réaffectée de Zürich (et de quelques autres cantons peuplés) à plusieurs cantons moins peuplés.

La figure 4 représente graphiquement l'erreur-type « nationale » $\sqrt{\text{var}(\hat{\theta})}$ sous la répartition optimale de l'échantillon pour une matrice de valeurs de q et de G. Le graphique montre qu'une légère augmentation de G aux alentours de $G = 0$ réduit spectaculairement l'erreur-type de $\hat{\theta}$, tandis que pour les valeurs plus grandes de G, l'erreur-type ne varie que légèrement. Pour chaque G, un exposant de priorité plus élevé q est associé à une précision plus élevée de $\hat{\theta}$.

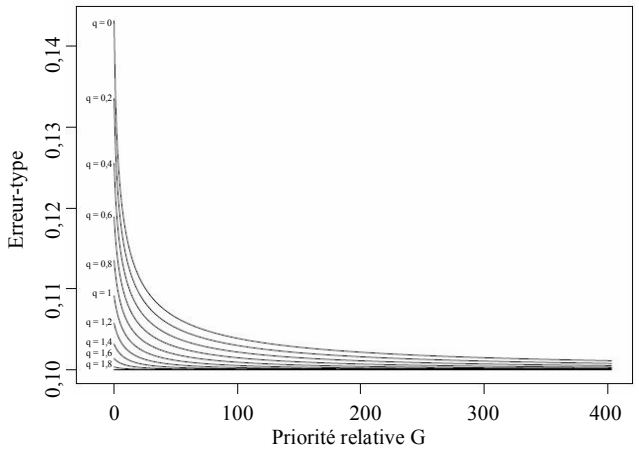


Figure 4. Erreur-type de l'estimateur national pour la répartition optimale sous une matrice de priorités données par q et G.

3. Estimation composite

L'utilisation la plus efficace des ressources disponibles pour réaliser une enquête s'obtient par combinaison optimale d'un plan d'échantillonnage et d'un ou de plusieurs estimateurs, si bien que le plan d'échantillonnage et (le choix de) l'estimateur devraient, dans des circonstances idéales, être optimisés simultanément. Ce problème est difficile à résoudre formellement dans la plupart des conditions, quoique certains estimateurs soient plus efficaces que leurs concurrents et que l'on considère une grande gamme de plans d'échantillonnage. Les estimateurs composites (Longford 1999, 2004) représentent l'une de ces classes d'estimateurs. Il s'agit de combinaisons convexes des estimateurs directs sur petits domaines et au niveau national,

$$\tilde{\theta}_d = (1 - b_d)\,\hat{\theta}_d + b_d\,\hat{\theta}, \qquad (2)$$

avec des coefficients particuliers aux petits domaines b_d qui sont des estimations de l'optimum. La composition $\tilde{\theta}_d$ tire parti de la similarité des petits domaines et est particulièrement efficace lorsqu'ils présentent une faible variance interdomaines $\sigma_B^2 = D^{-1}\sum_d (\theta_d - \bar{\theta})^2$, où $\bar{\theta} = D^{-1}\sum_d \theta_d$. Cette variance est définie sur les D paramètres de population θ_d et n'est pas affectée par le plan d'échantillonnage. En pratique, il faut estimer σ_B^2. Lors de la planification d'une enquête, il est nécessaire d'utiliser des estimations provenant d'autres enquêtes auprès de la même population ou de populations apparentées, et de tenir compte de l'incertitude au sujet de σ_B^2, ce qui peut se faire par analyse de sensibilité, en recherchant les plans d'échantillonnage optimaux pour une gamme de valeurs plausibles de σ_B^2.

Si les écarts $\Delta_d = \theta_d - \bar{\theta}$ étaient connus, le coefficient optimal b_d dans (2) serait, approximativement, $b_d^* = \sigma_d^2 / (\sigma_d^2 + n_d \Delta_d^2)$. Puisque nous ne connaissons pas Δ_d (sinon, θ_d serait estimé avec une grande précision par $\bar{\theta} + \Delta_d$), nous remplaçons Δ_d^2 par sa moyenne sur les petits domaines, égale à σ_B^2, ce qui donne le coefficient $b_d = 1/(1 + n_d \omega_d)$, où $\omega_d = \sigma_B^2 / \sigma_d^2$ est le ratio de variance. La variance σ_B^2 doit aussi être estimée, mais, si le nombre de petits domaines est élevé, l'estimation est beaucoup plus précise que ne le sont la plupart des Δ_d^2.

Si les coefficients b_d sont estimés avec suffisamment de précision, l'estimateur composite $\tilde{\theta}_d$ est plus efficace que les deux estimateurs qui le constituent, $\hat{\theta}_d$ et $\hat{\theta}$. Si nous ne tenons pas compte de l'incertitude au sujet des variances intra et interdomaines, ni au sujet de la moyenne nationale $\bar{\theta}$ et de la corrélation entre les estimateurs (direct) au niveau national et sur petits domaines, l'EQM moyenne de $\tilde{\theta}_d$ est

$$\text{aEQM}(\tilde{\theta}_d) = \frac{\sigma_B^2}{1 + n_d\,\omega_d}, \qquad (3)$$

où « aEQM » dénote l'EQM dans laquelle Δ_d^2 est remplacé par σ_B^2, sa moyenne sur l'ensemble des petits domaines. Dans (3), aEQM est aussi une approximation de la variance conditionnelle de l'estimateur EBLUP de la moyenne au niveau du petit domaine fondée sur le modèle (empirique bayésien) à deux niveaux (Longford 1993, Goldstein 1995, Marker 1999 et Rao 2003). Voir Ghosh et Rao (1994) pour une revue reconnue de l'application de ces modèles à l'estimation pour petits domaines.

Pour les estimateurs composites des moyennes de petit domaine, nous recherchons la répartition de l'échantillon qui minimise la fonction objectif

$$\sum_{d=1}^{D} P_d \, \text{aEQM}(\tilde{\theta}_d) + GP_+ v.$$

La solution satisfait la contrainte

$$\frac{N_d^q \sigma_B^2 \omega_d}{(1+n_d\omega_d)^2} + GP_+ \frac{N_d^2}{N^2}\frac{\sigma_d^2}{n_d^2} = \text{const.} \qquad (4)$$

Cette équation ne possède pas de solution analytique commode, mais elle peut être résolue par application de scénarios itératifs. La valeur de n_1 détermine les autres tailles d'échantillon n_d, de sorte que l'optimisation correspond à une recherche unidimensionnelle. Si les tailles d'échantillon provisoires $\mathbf{n}$ fondées sur un ensemble de valeur de n_1 sont trop grandes, on réduit $\mathbf{n}^\top \mathbf{1}_D > n$, n_1 et on calcule les autres tailles d'échantillon n_d en résolvant (4). Notons que la solution dépend des variances σ_d^2 et σ_B^2. Le problème se simplifie quelque peu lorsque la variance est la même pour tous les petits domaines $\sigma^2 = \sigma_1^2 = ... = \sigma_D^2$. Alors, la solution de (4) dépend des variances uniquement par la voie du ratio $\omega = \sigma_B^2/\sigma^2$, parce que σ^2 est un facteur multiplicatif qui n'a aucun effet sur l'optimisation.

À titre d'exemple, supposons que $q = 1$ et $G = 10$ lors de la planification d'une enquête auprès de la population suisse, avec $n = 10\,000$, et en supposant que $\omega = 0{,}10$. Comme solution initiale, nous utilisons la répartition optimale pour l'estimation directe avec les mêmes valeurs de q et de G. Une itération met à jour la taille de l'échantillon de chaque canton et, dans les cantons, la mise à jour pour tous, sauf celui de référence sélectionné arbitrairement $d = 1$, est également itérative. La taille provisoire du sous-échantillon pour le canton de référence détermine la valeur courante de la constante dans le deuxième membre de (4). L'équation (4) est résolue, itérativement, pour chaque canton $d = 2, ..., D$, en utilisant la méthode de Newton. Dans l'application, le nombre d'itérations était inférieur à dix pour chaque canton. Enfin, la taille du sous-échantillon pour le canton de référence est ajustée par le facteur $1/D$ un multiple de la différence entre le total courant des tailles des sous-échantillons et le total cible n. La mise à jour des tailles d'échantillon des cantons est elle-même itérée, mais quelques itérations seulement sont nécessaires pour atteindre la convergence; par exemple, toutes les variations des tailles des sous-échantillons étaient inférieures à 1,0 après trois itérations et inférieures à 0,01 après huit itérations. La convergence est rapide, parce que la solution de départ est proche de la solution optimale; l'écart le plus important entre les deux tailles de sous-échantillon est celui observé pour Zurich, soit 20,0 (de 1199,5 au départ à 1219,5 après huit itérations). Pour Appenzell-Innerrhoden, la taille d'échantillon est réduite de 81,6 à 73,4. Des changements de moins d'une unité ont lieu pour cinq cantons dont la taille de population varie de 228 000 à 278 000. Notons qu'en pratique, les tailles des sous-échantillons seraient arrondies et éventuellement ajustées davantage afin de satisfaire aux diverses contraintes de gestion de l'enquête.

Pas de priorité accordée à l'estimation nationale

Si l'estimation nationale n'a aucune priorité, $G = 0$, l'équation (4) possède la solution explicite

$$n_d^* = \frac{n\omega + D}{\omega}\frac{N_d^{q/2}}{U^{(q)}} - \frac{1}{\omega},$$

où $U^{(q)} = N_1^{q/2} + ... + N_D^{q/2}$. Cette répartition est reliée à la répartition $n_d^\dagger$, $d = 1, ..., D$, qui est optimale pour l'estimation directe de θ_d, par l'identité

$$n_d^* = n_d^\dagger + \frac{1}{\omega}\left(\frac{DN_d^{q/2}}{U^{(q)}} - 1\right).$$

Donc, quand $q > 0$, la répartition optimale est plus dispersée dans le cas de l'estimation composite que dans celui de l'estimation directe. La taille de population au point d'équilibre est $N_T = (U^{(q)}/D)^{2/q}$; la taille du sous-échantillon pour les petits domaines ayant une taille de population $N_d < N_T$ est plus petite dans le cas de l'estimation composite que dans celui de l'estimation directe, et elle est plus grande pour les petits domaines dont la population est plus grande. (Pour $q = 0$, $n_d^* \equiv n/D$). Le degré de dispersion supplémentaire est inversement proportionnel à ω.

Si $\omega = 0$, les équations pour le plan d'échantillonnage optimal donnent lieu à une singularité. Dans ce cas, chaque θ_d est estimé efficacement par l'estimateur national $\hat{\theta}$, si bien que le plan optimal pour l'estimation composite coïncide avec le plan optimal pour l'estimateur national ($n_d^* = nN_d/N$). Pour $q > 0$, la répartition optimale donne des tailles d'échantillon négatives n_d^* quand

$$N_d < \left\{\frac{U^{(q)}}{n\omega + D}\right\}^{2/q}. \qquad (5)$$

Cette solution (formelle) n'a pas de sens. Une solution négative ne devrait pas être étonnante, car l'aEQM de (3) est une fonction analytique pour $n_d > -1/\omega_d$. Si les valeurs de $\omega > 0$ sont faibles, l'aEQM est une fonction décroissante à pente faible de la taille d'échantillon n_d. Une valeur négative de n_d^* indique qu'un « petit » canton ne vaut pas la peine d'être échantillonné, à cause de la faible priorité d'inférence P_d. Bien que l'accroissement de la taille de l'échantillon d'un canton plus peuplé d' puisse donner lieu à une réduction plus faible de l'aEQM que cela ne serait le cas pour un petit canton d, la priorité plus grande $P_{d'}$ augmente l'effet.

Priorité positive pour la moyenne nationale

Dans (3), l'aEQM ne tient pas compte de l'incertitude au sujet de la moyenne nationale θ, situation qui devient

critique lorsque l'un des cantons n'est pas représenté dans l'échantillon. Cette déficience de (3) peut être compensée en fixant la priorité relative G à une valeur positive.

La figure 5 résume l'effet de la priorité relative G et de l'exposant de priorité q sur les tailles d'échantillon optimales pour les cantons le moins et le plus peuplés, ainsi que le canton de Thurgau qui possède la 13e taille de population par ordre décroissant (médiane), soit 228 000. Chaque valeur de q, indiquée dans le titre, et de G, indiquée en utilisant différents types de lignes, est représentée pour un canton par un graphique de la taille d'échantillon optimale en fonction du ratio de variance ω. La limite de cette fonction lorsque $\omega \to +\infty$, égale à la taille d'échantillon optimale pour l'estimation directe, est marquée par une barre dans la marge de droite du volet en question. Pour $\omega = 0$, on obtient le plan d'échantillonnage optimal pour l'estimation de la moyenne nationale θ. Les volets A et B au haut de la figure correspondent à la taille d'échantillon globale $n = 10\,000$ et les volets C et D, à $n = 1\,000$.

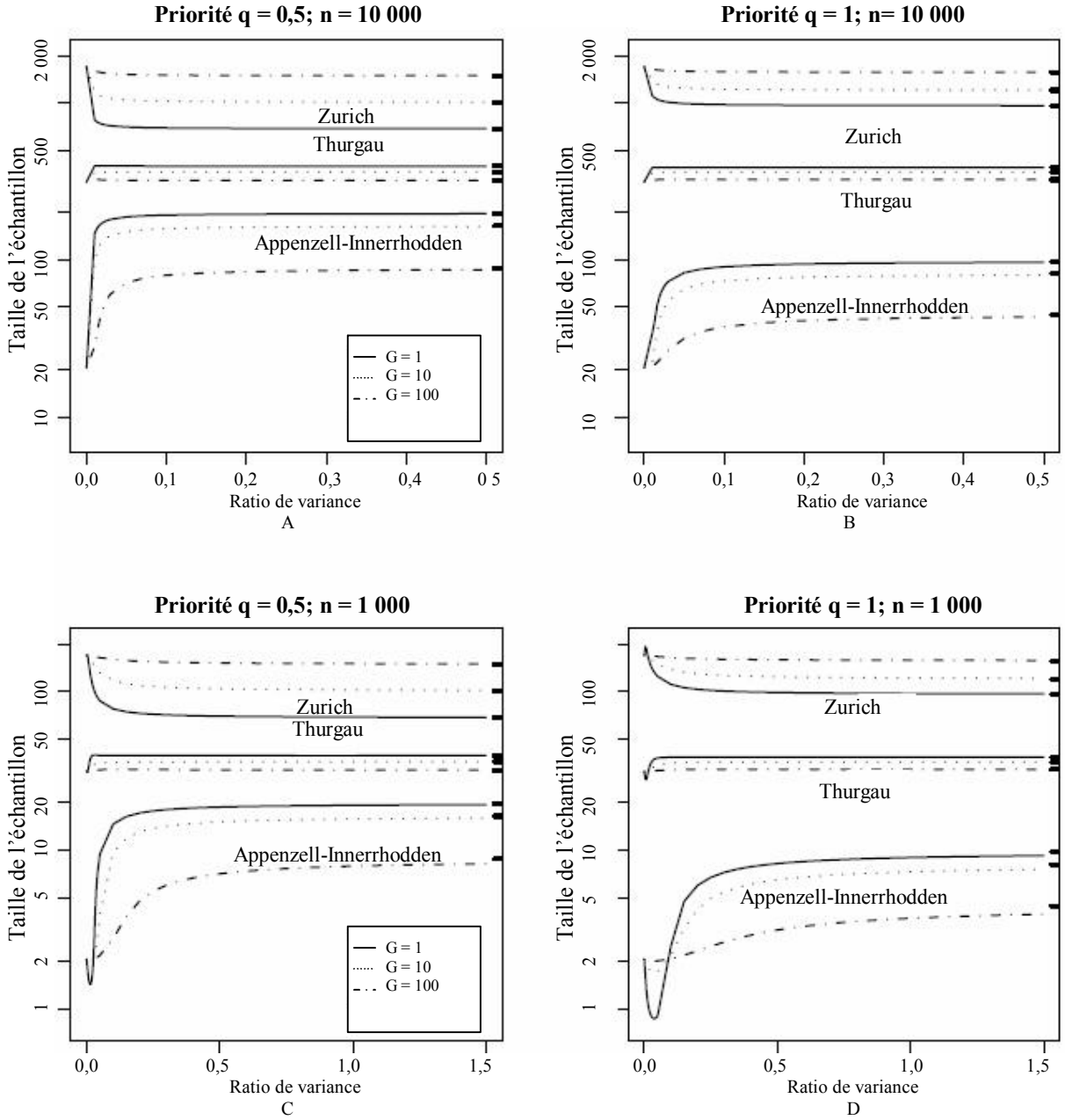


Figure 5. Tailles d'échantillon optimales pour l'estimation composite des moyennes de population pour trois cantons pour une gamme de rapports de variance ω, les exposants de priorité $q = 0,5$ et $q = 1,0$ et les priorités relatives $G = 1$, 10 et 100. Les tailles globales d'échantillon sont 10 000 (volets A et B) et 1 000 (volets C et D).

Le graphique montre que les tailles d'échantillon optimales sont presque constantes dans la fourchette $\omega \in (\omega^*, +\infty)$; ω^* augmente avec q, G et $1/n$. Il s'agit d'une conséquence de la taille d'échantillon relativement grande n, qui assure que les sous-échantillons de la plupart des cantons soient trop grands pour qu'un emprunt important d'information entre les cantons aient lieu, à moins que les cantons soient fort semblables ($\omega < \omega^*$). La plupart des coefficients de rétrécissement $b_d = 1/(1+n_d\,\omega)$ sont très petits. Lorsqu'une taille $n = 10\,000$ est prévue, pour les valeurs faibles de ω, la taille d'échantillon optimale augmente fortement pour les cantons les moins peuplés et chute brusquement pour les plus peuplés. La dispersion des tailles d'échantillon optimales augmente avec q et G, convergeant vers la répartition optimale pour l'estimation de la moyenne nationale θ, qui correspond à $\omega = 0$. Par contre, les tailles d'échantillon optimales sont discontinues à $\omega = 0$ quand $G = 0$; les solutions divergent vers $-\infty$ pour les cantons les moins peuplés.

Dans les volets C et D, pour $n = 1\,000$, la variation des tailles d'échantillon en fonction de ω persiste pour une plus grande fourchette de valeur de ω, parce que la portée de l'emprunt d'information entre les cantons est plus grande pour les tailles d'échantillon plus petites. Les tailles d'échantillon optimales ne sont pas des fonctions monotones de ω; pour les cantons les moins peuplés, on observe un creux pour les faibles valeurs de ω.. Le creux est plus prononcé pour les faibles valeurs de G et pour les grandes valeurs de q, c'est-à-dire lorsque les disparités entre les priorités des cantons sont grandes et que l'importance relative de l'inférence au sujet de la moyenne nationale est plus faible. Ce phénomène, quelque peu exagéré par l'échelle logarithmique de l'axe des ordonnées, est semblable au cas discuté pour $G = 0$. À cause des différences de priorité P_d, une faible réduction de l'aEQM pour un canton plus peuplé est préférable à une réduction plus importante pour un canton moins peuplé. Le creux existe aussi quand $n = 10\,000$, mais il est si peu profond et si étroit qu'il n'est pas visible dans les conditions de résolution du graphique. Notons que, dans les volets C et D, l'axe des abscisses possède une fourchette de valeurs de ω trois fois plus grande que dans les volets A et B.

Dans le contexte de l'enquête planifiée, il a été convenu qu'il était peu probable que la valeur de ω soit inférieure à 0,05. Par conséquent, le calcul des tailles d'échantillon a pu être fondé sur l'estimateur direct.

4. Discussion

La méthode décrite dans le présent article permet de déterminer le plan d'échantillonnage optimal pour les conditions artificielles d'échantillonnage stratifié avec échantillonnage aléatoire simple dans des strates homoscédastiques. La spécification des priorités en ce qui concerne l'estimation pour petits domaines et l'estimation nationale est un élément essentiel de la méthode. En pratique, il peut être difficile de se mettre d'accord sur les priorités et certaines hypothèses peuvent être problématiques, en particulier celles de l'égalité des variances intrastrate et de l'échantillonnage aléatoire simple. La méthode peut être étendue à des estimateurs plus complexes, mais les valeurs de paramètres supplémentaires sont alors nécessaires. Une approche plus constructive consiste à considérer le plan d'échantillonnage optimal pour les conditions simplifiées en tant qu'approximation du plan d'échantillonnage qui est optimal pour les conditions plus réalistes. Même si le plan d'échantillonnage optimal était déterminé, il ne pourrait être appliqué littéralement, à cause des imperfections de la base de sondage et (éventuellement) de la non-réponse informative et non uniformément distribuée. Cependant, l'approche est applicable, en principe, à tout estimateur sur petits domaines pour lequel il existe une expression analytique exacte ou approximative de l'EQM. Cela inclut tous les estimateurs fondés sur les modèles bayésiens empiriques, auxquels l'estimateur composite est étroitement associé. Les poids de sondage peuvent être intégrés dans le calcul de la taille de l'échantillon s'ils sont connus ou que leurs distributions dans les petits domaines sont connues a priori, sous réserve de certaines approximations. Le calcul de la taille d'échantillon pour une grandeur (nationale) unique pose le même problème.

Bien que la solution numérique du problème pour l'estimation composite avec une priorité positive G soit simple et ne présente aucun problème de convergence, il est avantageux de disposer d'une solution analytique, afin de pouvoir étudier une gamme de scénarios. La proximité des solutions obtenues pour les estimations directe et composite donne à penser que la répartition optimale pour l'estimation directe pourrait également s'approcher de la situation optimale pour l'estimation composite avec des valeurs raisonnables de ω, disons, $\omega > 0,05$.

Diverses contraintes de gestion et d'organisation constituent un autre obstacle à l'application littérale d'un plan d'échantillonnage établi analytiquement. Dans les enquêtes-ménages, il est souvent préférable d'attribuer un quota (presque) complet d'adresses à chaque intervieweur, si bien que l'on accorde la préférence aux tailles d'échantillon qui sont des multiples du quota. Ces considérations et de nombreuses autres contraintes peuvent être intégrées dans le problème d'optimisation, quoiqu'elles soient souvent difficile à quantifier ou que le concepteur de l'enquête ne soit pas conscient de leur existence à cause d'une communication imparfaite. L'improvisation, après l'obtention d'un plan d'échantillonnage optimal pour des conditions plus

simples, pourrait être plus pratique. En outre, les priorités, ou l'opinion d'experts à leur sujet, peuvent évoluer au cours du temps, même pendant la réalisation de l'enquête et l'analyse des données. Les estimations associées à une erreur-type ou à un coefficient de variation supérieur à un seuil précisé sont souvent exclues des rapports analytiques. L'intention de les exclure peut être reflétée dans le calcul de la taille d'échantillon en considérant $\hat{\theta}$ comme étant l'estimateur de θ_d, c'est-à-dire en fixant l'EQM connexe à l'aEQM $\sigma_B^2 + \mathrm{var}(\hat{\theta})$ correspondante ou à une autre (grande) valeur constante.

Bien que nous proposions une classe particulière de priorités pour les petits domaines, aucune difficulté conceptuelle ne se pose lorsque l'on utilise une autre classe. Elle pourrait dépendre de plusieurs grandeurs de population plutôt que de la taille de population uniquement. En principe, les priorités peuvent aussi être fixées individuellement pour les petits domaines, bien que cela ne soit pratique que si leur nombre est faible. Les priorités fondées sur la formule ou établies individuellement peuvent être combinées en ajustant les priorités, telles que $P_d = N_d^q$, pour quelques petits domaines afin de refléter leur rôle exceptionnel dans l'analyse.

Une analyse de sensibilité, en vue d'étudier les modifications du plan d'échantillonnage en fonction de diverses données d'entrée est essentielle à la compréhension de l'incertitude au sujet des paramètres estimés (le ratio de variance ω en particulier) et le caractère arbitraire, aussi limité qu'il soit, de l'établissement des priorités. Pour cela, il est préférable de disposer d'une solution analytiquement simple qui peut être exécutée de nombreuses fois, pour une gamme de conditions, plutôt qu'une solution plus complexe, dont les propriétés sont difficiles à étudier.

Les estimateurs composites multivariés exploitent la similarité non seulement entre les petits domaines, mais aussi entre les variables (auxiliaires), les périodes, les sous-populations, et ainsi de suite (Longford 1999 et 2005). L'aEQM de ces estimateurs dépend de la matrice de variances mise à l'échelle $\mathbf{\Omega}$, qui est le pendant multivarié de ω. Le calcul de la taille d'échantillon par cette méthode est difficile à appliquer directement, parce que, dans $\mathbf{\Omega}$, les variances et les covariances sont les unes et les autres essentielles à l'efficacité des estimateurs. Une approche plus constructive consiste à faire concorder la matrice $\mathbf{\Omega}$ avec un ratio ω qui peut être interprété comme étant la similarité des petits domaines après correction pour l'information auxiliaire, comme dans les méthodes bayésiennes empiriques.

Lorsque il est impossible d'exercer un contrôle sur les tailles d'échantillon affectées aux petits domaines, leur calcul demeure utile comme indication de la façon dont elles devraient être réparties *en moyenne*. En général, une réduction unitaire de la taille d'échantillon est associée à une perte plus importante de précision qu'un accroissement unitaire. Par conséquent, les plans dans lesquels la variance d'échantillonnage (estimée par rééchantillonnage) des tailles des sous-échantillons n_d (d fixé) est plus faible sont mieux adaptés à l'estimation pour petits domaines. Dans les plans d'échantillonnage où les grappes sont importantes, ces variances sont grandes parce que, dans le cas extrême, un petit domaine pourrait ne pas être représenté dans l'échantillon lors de certaines répliques et pourrait être surreprésenté plusieurs fois dans d'autres. En général, il est préférable d'utiliser de plus petites grappes pour l'estimation pour petits domaines, si cela n'augmente pas les coûts d'enquête et qu'il est possible de maintenir une taille globale d'échantillon fixe.

Remerciements

Je remercie le rédacteur en chef délégué et les examinateurs d'avoir proposé plusieurs améliorations, mais surtout de m'avoir fait découvrir une erreur dans une version antérieure du manuscrit. Je tiens aussi à mentionner mes discussions avec l'équipe polonaise du projet EURAREA.

Bibliographie

Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Ghosh, M., et Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

Goldstein, H. (1995). *Multilevel Statistical Models.* Deuxième Édition. Edward Arnold, London, UK.

Longford, N.T. (1993). *Random Coefficient Models*. Oxford University Press, Oxford.

Longford, N.T. (1999). Multivariate shrinkage estimation of small-area means and proportions. *Journal of the Royal Statistical Society*, Séries A, 162, 227-245.

Longford, N.T. (2004). Missing data and small area estimation in the UK Labour Force Survey. *Journal of the Royal Statistical Society*, Séries A, 167, 341-373.

Longford, N.T. (2005). *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician*. Springer-Verlag, New York.

Marker, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.

Marker, D.A. (2001). Production d'estimations régionales d'après les données d'enquêtes nationales : Méthodes visant à réduire au minimum l'emploi d'estimateurs indirects. *Techniques d'enquête*, 27, 201-207.

Platek, R., Rao, J.N.K., Särndal, C.-E. et Singh, M.P. (Éds.) (1987). S*mall Area Statistics*. New York: John Wiley & Sons.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Singh, M.P., Gambino, J. et Mantel, H.J. (1994). Les petites régions : Problèmes et solutions. *Techniques d'enquête*, 20, 3-23.

Techniques d'enquête, juin 2006
Vol. 32, N° 1, pp. 107-114
Statistique Canada, N° 12-001-XIF au catalogue

Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances d'échantillonnage

Yong You et Beatrice Chapman [1]

Résumé

Dans le contexte de l'estimation pour petits domaines, des modèles régionaux, comme le modèle de Fay-Herriot (Fay et Herriot, 1979), sont très souvent utilisés en vue d'obtenir de bons estimateurs fondés sur un modèle pour les petits domaines ou petites régions. Il est généralement supposé que les variances d'erreur d'échantillonnage incluses dans le modèle sont connues. Dans le présent article, nous considérons la situation où les variances d'erreur d'échantillonnage sont estimées individuellement au moyen d'estimateurs directs. Nous construisons un modèle hiérarchique bayésien (HB) complet pour les estimateurs par sondage directs et pour les estimateurs de variance de l'erreur d'échantillonnage. Nous employons la méthode d'échantillonnage de Gibbs pour obtenir les estimateurs HB pour les petites régions. L'approche HB proposée tient compte automatiquement de l'incertitude supplémentaire associée à l'estimation des variances d'erreur d'échantillonnage, particulièrement quand la taille des échantillons régionaux est très faible. Nous comparons le modèle HB proposé au modèle de Fay-Herriot grâce à l'analyse de deux ensembles de données d'enquête. Nos résultats montrent que les estimateurs HB proposés donnent d'assez bons résultats comparativement aux estimations directes. Nous discutons également du problème des lois a priori sur les composantes de la variance.

Mots clés : Échantillonnage de Gibbs; hiérarchique bayésien; sensibilité aux lois a priori; taille d'échantillon; composantes de la variance.

1. Introduction

Dans la plupart des applications, les enquêtes par sondage sont conçues afin de fournir des estimations directes fiables pour l'ensemble de la population, de même que pour les grandes régions au moyen de données d'échantillon propres à la région. Toutefois, fréquemment, cette méthode d'estimation directe ne produit pas d'estimations fiables pour les petites régions, à cause de la très petite taille des échantillons obtenus pour ces dernières. Puisque les estimations directes pour les petites régions sont souvent assorties d'une erreur-type trop grande, si l'on veut augmenter la précision et la fiabilité, il est nécessaire d'« emprunter de l'information » aux régions apparentées, donc d'accroître la taille efficace de l'échantillon, en vue de produire des estimations indirectes pour les petites régions (Rao 1999). Les méthodes fondées sur un modèle explicite, qui s'appuient sur des données supplémentaires, telles que des données de recensement ou des données administratives, associées aux petites régions dans des modèles explicites en vue de relier ces régions, ont été utilisées à grande échelle en pratique pour obtenir des estimateurs fondés sur un modèle fiables. Ces modèles se répartissent en deux grandes catégories, à savoir les modèles au niveau de la région et les modèles au niveau de l'unité. Les modèles de niveau régional sont fondés sur des estimateurs par sondage régionaux directs et les modèles de niveau unitaire sont fondés sur les observations individuelles recueillies dans les régions. Pour une vue d'ensemble et une évaluation des modèles appliqués à l'estimation pour petits domaines ou petites régions, voir Rao (1999, 2003). Dans le présent article, nous étudions les modèles de niveau régional.

Pour obtenir un modèle régional de base, nous supposons que le paramètre d'intérêt de la petite région θ_i est relié à des données auxiliaires propres à la région $x_i = (x_{i1}, \ldots, x_{ip})'$ grâce à un modèle linéaire

$$\theta_i = x_i'\beta + v_i, i = 1, \ldots, m, \qquad (1)$$

où m est le nombre de petites régions, $\beta = (\beta_1, \ldots, \beta_p)'$ est le vecteur de dimensions $p \times 1$ de coefficients de régression, et les v_i sont les effets aléatoires propres à la région que nous supposons être indépendants et identiquement distribués (iid) avec $\mathrm{E}(v_i) = 0$ et $\mathrm{var}(v_i) = \sigma_v^2$. L'hypothèse de normalité peut également être incluse. Ce modèle est appelé modèle de liaison pour θ_i.

Le modèle régional de base repose aussi sur l'hypothèse qu'étant donné la taille d'échantillon propre à la région $n_i > 1$, il existe un estimateur par sondage direct y_i (habituellement sans biais par rapport au plan de sondage) pour le paramètre de petite région θ_i tel que

$$y_i = \theta_i + e_i,\ i = 1, \ldots, m, \qquad (2)$$

où e_i est l'erreur d'échantillonnage associée à l'estimateur direct y_i. Nous supposons aussi que les e_i sont des variables aléatoires normales indépendantes de moyenne $E(e_i|\theta_i) = 0$ et de variance d'échantillonnage $\mathrm{var}(e_i|\theta_i) = \sigma_i^2$. La combinaison des modèles (1) et (2) donne un modèle linéaire mixte régional

$$y_i = x_i'\beta + v_i + e_i,\ i = 1, \ldots, m. \qquad (3)$$

1. Yong You et Beatrice Chapman, Division des méthdodes d'enquêtes auprès des ménages, Statistique Canada, K1A 0T6. Courriel : yongyou@statcan.ca.

Le modèle bien connu de Fay–Herriot (Fay et Herriot 1979) appliqué à l'estimation pour petites régions a la forme du modèle (3) sous l'hypothèse que la variance d'échantillonnage σ_i^2 est connue dans le modèle, c'est-à-dire une hypothèse très forte. Habituellement, on utilise dans le modèle un estimateur lissé de σ_i^2 que l'on traite alors comme étant connue. Dans le présent article, nous considérons la situation où les variances d'échantillonnage σ_i^2 sont inconnues et sont estimées au moyen d'estimateurs sans biais s_i^2. À l'instar de Rivest et Vandal (2002) et de Wang et Fuller (2003), nous supposons que les estimateurs s_i^2 sont indépendants des estimateurs par sondage direct y_i et que leur distribution d'échantillonnage est $d_i\, s_i^2 \sim \sigma_i^2\, \chi_{d_i}^2$, où $d_i = n_i - 1$ et n_i est la taille d'échantillon pour la i^{e} région. Par exemple, supposons que nous ayons n_i observations provenant de la petite région i et que ces observations soient iid $N(\mu_i, \sigma^2)$. Soit y_i la moyenne d'échantillon des n_i observations. Alors, $y_i \sim N(\mu_i, \sigma_i^2)$ et $\sigma_i^2 = \sigma^2 / n_i$. Nous pouvons alors obtenir un estimateur direct de σ_i^2 sous la forme $s_i^2 = \tau_i^2 / n_i$, où τ_i^2 est la variance d'échantillon des n_i observations. En outre, y_i et s_i^2 sont indépendants et $(n_i - 1)\, s_i^2 \sim \sigma_i^2\, \chi_{n_i-1}^2$.

Nous voulons estimer les paramètres de petite région θ_i. Rivest et Vandal (2002), ainsi que Wang et Fuller (2003) ont obtenu les estimateurs par la méthode empirique du meilleur prédicteur linéaire sans biais (EBLUP) de θ_i et les approximations des erreurs quadratiques moyenne (EQM) associées en supposant que m et n_i sont relativement grands. Dans le présent article, nous considérons une approche hiérarchique bayésienne (HB) s'appuyant sur la méthode d'échantillonnage de Gibbs. L'un des avantages de l'approche HB est qu'elle est simple et que les inférences pour les paramètres θ_i sont « exactes », contrairement à celles obtenues par l'approche EBLUP. Le paramètre de petite région θ_i est estimé par sa moyenne a posteriori et sa précision est mesurée par sa variance a posteriori. L'approche HB tient compte automatiquement des incertitudes associées aux paramètres inconnus dans le modèle. À la section 2, nous présentons les modèles régionaux HB et les inférences basées sur l'échantillonnage de Gibbs connexes. À la section 3, nous décrivons l'analyse de deux ensembles de données d'enquête et une analyse de sensibilité. Enfin, à la section 4, nous offrons certaines conclusions et proposons certaines orientations pour de futurs travaux.

2. Approche hiérarchique bayésienne

Nous allons maintenant présenter le modèle régional (3) et les variances d'échantillonnage estimées s_i^2 dans un cadre hiérarchique bayésien (HB) comme il suit :

Modèle 1

- $y_i \mid \theta_i, \sigma_i^2 \sim \text{ind}\ N(\theta_i, \sigma_i^2),\ i = 1, \ldots, m;$
- $d_i\, s_i^2 \mid \sigma_i^2 \sim \text{ind}\ \sigma_i^2\, \chi_{d_i}^2,\ d_i = n_i - 1,\ i = 1, \ldots, m;$
- $\theta_i \mid \beta, \sigma_v^2 \sim \text{ind}\ N(x_i'\beta, \sigma_v^2),\ i = 1, \ldots, m;$
- Lois a priori sur les paramètres : $\pi(\beta) \propto 1,\ \pi(\sigma_i^2) \sim \text{GI}(a_i, b_i), i = 1, \ldots, m,\ \pi(\sigma_v^2) \sim \text{GI}(a_0, b_0)$, où les $a_i, b_i\ (0 \le i \le m)$ sont des constantes connues fixées à une valeur très petite afin de refléter les connaissances vagues au sujet de σ_i^2 et σ_v^2. GI dénote la loi gamma inverse.

Dans le modèle 1, les variances d'échantillonnage σ_i^2 sont inconnues. Cependant, en pratique, nous pourrions avoir un modèle plus simple en remplaçant σ_i^2 par son estimation s_i^2 (ici s_i^2 est traitée comme si elle était constante) et obtenir le modèle suivant :

Modèle 2

- $y_i \mid \theta_i \sim \text{ind}\ N(\theta_i, \sigma_i^2 = s_i^2),\ i = 1, \ldots, m;$
- $\theta_i \mid \beta, \sigma_v^2 \sim \text{ind}\ N(x_i'\beta, \sigma_v^2),\ i = 1, \ldots, m;$
- Lois a priori : $\pi(\beta) \propto 1,\ \pi(\sigma_v^2) \sim \text{GI}(a_0, b_0).$

Le modèle 2 est, en fait, le modèle de Fay-Herriot avec variances d'échantillonnage connues, s_i^2. Si les tailles des échantillons régionaux n_i sont faibles, l'utilisation de s_i^2 dans le modèle 2 peut donner lieu à une sous-estimation de l'EQM sous l'approche EBLUP ou de la variance a posteriori sous l'approche HB. Nous souhaitons évaluer les effets de l'utilisation de s_i^2 pour σ_i^2 dans le modèle. Nous obtiendrons les estimations HB de θ_i sous le modèle 1 ainsi que le modèle 2 et les comparerons en procédant à l'analyse de données d'enquête réelles.

Sous l'approche HB, nous utilisons la moyenne a posteriori $E(\theta_i \mid y)$ en tant qu'estimation ponctuelle de θ_i et la variance a posteriori $V(\theta_i \mid y)$ en tant que mesure de la variabilité, où $y = (y_1, \ldots, y_m)'$. Pour estimer $E(\theta_i \mid y)$ et $V(\theta_i \mid y)$, nous employons la méthode d'échantillonnage de Gibbs (Gelfand et Smith 1990). Partant du modèle 1, nous obtenons les lois conditionnelles complètes suivantes pour l'échantillonneur de Gibbs :

- $[\theta_i \mid y, \beta, \sigma_i^2, \sigma_v^2] \sim N(\gamma_i\, y_i + (1 - \gamma_i) x_i'\beta,\ \gamma_i\, \sigma_i^2),$ où
  $\gamma_i = \dfrac{\sigma_v^2}{\sigma_v^2 + \sigma_i^2},\ i = 1, \ldots, m;$
- $[\beta \mid y, \theta, \sigma_i^2, \sigma_v^2] \sim N_p\left(\left(\sum_{i=1}^m x_i x_i'\right)^{-1}\left(\sum_{i=1}^m x_i \theta_i\right),\ \sigma_v^2\left(\sum_{i=1}^m x_i x_i'\right)^{-1}\right);$

- $[\sigma_i^2 \mid y, \theta, \beta, \sigma_v^2] \sim \text{GI}\left(\begin{array}{l} a_i + \dfrac{d_i + 1}{2}, b_i \\ + \dfrac{(y_i - \theta_i)^2 + d_i s_i^2}{2} \end{array}\right)$,

  où $d_i = n_i - 1,\ i = 1, \ldots, m$;

- $[\sigma_i^2 \mid y, \theta, \beta, \sigma_i^2] \sim \text{GI}\left(\begin{array}{l} a_0 + \dfrac{m}{2}, b_0 \\ + \dfrac{1}{2}\displaystyle\sum_{i=1}^{m} (\theta_i - x_i' \beta)^2 \end{array}\right)$.

Il est facile de tirer des échantillons à partir de ces lois conditionnelles complètes. Pour les applications, nous utilisons $L = 5$ exécutions parallèles, chacune avec une durée de « rodage » de $B = 1\,000$ et une taille d'échantillon de Gibbs de $G = 5\,000$. Les paramètres a priori a_i, b_i et a_0, b_0 sont fixés à 0,0001. Nous obtenons donc l'estimateur HB de θ_i sous le modèle 1 suivant

$$\hat{\theta}_i^{\text{HB}} = (LG)^{-1} \sum_{l=1}^{L} \sum_{g=1}^{G} (\gamma_i^{(\text{lg})} y_i + (1 - \gamma_i^{(\text{lg})})\, x_i' \beta^{(\text{lg})}), \qquad (4)$$

où $\gamma_i^{(\text{lg})} = \sigma_v^{2(\text{lg})} / (\sigma_v^{2(\text{lg})} + \sigma_i^{2(\text{lg})})$, et la variance a posteriori de θ_i peut être estimée par

$$\begin{aligned} \hat{V}(\theta_i) = {} & (LG)^{-1} \sum_{l=1}^{L} \sum_{g=1}^{G} (\gamma_i^{(\text{lg})} \sigma_i^{2(\text{lg})}) \\ & + (LG)^{-1} \sum_{l=1}^{L} \sum_{g=1}^{G} (\gamma_i^{(\text{lg})} y_i + (1 - \gamma_i^{(\text{lg})}) x_i' \beta^{(\text{lg})})^2 \\ & - \left\{ (LG)^{-1} \sum_{l=1}^{L} \sum_{g=1}^{G} (\gamma_i^{(\text{lg})} y_i + (1 - \gamma_i^{(\text{lg})}) x_i' \beta^{(\text{lg})}) \right\}^2, \qquad (5) \end{aligned}$$

où $\{\beta^{(\text{lg})}, \sigma_i^{2(\text{lg})}, \sigma_v^{2(\text{lg})};\ g = 1, \ldots, G;\ l = 1, \ldots, L\}$ est l'échantillon généré au moyen de l'échantillonneur de Gibbs. Les estimateurs (4) et (5) sont les estimateurs HB dits rao-blackwellisés. Les estimateurs rao-blackwellisés sont plus stables pour ce qui est des erreurs de simulation, comme l'ont montré, par exemple, Gelfand et Smith (1991), ainsi que You et Rao (2000).

Maintenant, considérons le modèle 2. Les lois conditionnelles complètes pour l'échantillonneur de Gibbs sous le modèle 2 sont

- $[\theta_i \mid y, \beta, \sigma_v^2] \sim N(\gamma_i\, y_i + (1 - \gamma_i) x_i' \beta, \gamma_i\, s_i^2)$, où $\gamma_i = \dfrac{\sigma_v^2}{\sigma_v^2 + s_i^2},\ i = 1, \ldots, m$;

- $[\beta \mid y, \theta, \sigma_v^2] \sim N_p\left(\begin{array}{l} \left(\displaystyle\sum_{i=1}^{m} x_i x_i'\right)^{-1} \left(\displaystyle\sum_{i=1}^{m} x_i \theta_i\right), \\ \sigma_v^2 \left(\displaystyle\sum_{i=1}^{m} x_i x_i'\right)^{-1} \end{array}\right)$;

- $[\sigma_v^2 \mid y, \theta, \beta] \sim \text{GI}\left(\begin{array}{l} a_0 + \dfrac{m}{2}, b_0 \\ + \dfrac{1}{2}\displaystyle\sum_{i=1}^{m} (\theta_i - x_i' \beta)^2 \end{array}\right)$.

Sous le modèle 2, l'estimateur HB de θ_i et l'estimateur de la variance a posteriori correspondant sont donnés par (4) et (5), respectivement, avec $\sigma_i^{2(\text{lg})}$ remplacé par s_i^2. Soulignons que l'utilisation de s_i^2 au lieu de $\sigma_i^{2(\text{lg})}$ peut donner lieu à une sous-estimation importante de la variance a posteriori de θ_i pour certaines régions pour lesquelles la taille d'échantillon n_i est petite. Nous comparerons les estimateurs HB et évaluerons les effets de l'utilisation de s_i^2 dans le modèle 2 grâce à une analyse de données à la section suivante.

3. Analyse de données

3.1 Les ensembles de données

Nous considérons deux ensembles de données intéressants pour nos analyses. Le premier contient des données sur les cultures de maïs et de soja pour huit régions seulement pour lesquelles la taille d'échantillon est petite. Le deuxième contient des données sur le lait pour 43 régions pour lesquelles la taille d'échantillon est relativement grande. Nous comparerons les modèles HB et les estimations basées sur ces deux ensembles de données.

Données sur le maïs et le soja : Ces données, qui proviennent du U.S. Department of Agriculture, ont été étudiées pour la première fois par Battese, Harter et Fuller (1988). Elles contiennent les nombres d'hectares cultivés déclarés et des données recueillies par le satellite LANDSAT pour les cultures de maïs et de soja dans des segments échantillonnés de 12 comtés de l'Iowa. Les nombres déclarés d'hectares pour chaque culture constituent les estimations directes par sondage. Les données auxiliaires sont les moyennes de population du nombre de pixels d'une culture donnée par segment. Les tailles d'échantillon sont petites pour ces régions, variant de un à cinq. Pour l'étude, nous utilisons uniquement les comtés pour lesquels la taille d'échantillon est égale ou supérieure à trois (huit régions répondent à ce critère). Par conséquent, la taille d'échantillon des comtés varie de trois à cinq. Les données originales sont des données au niveau de l'unité. Afin d'obtenir des données au niveau de la région, nous avons calculé la moyenne d'échantillon et l'erreur-type d'échantillon pour chaque comté. Pour les données sur le maïs et le soja, les erreurs-types d'échantillon sont en général assez grandes (donnant certains c.v. dans la fourchette de 0,3 à 0,4 et un c.v. de 0,532), mais, par hasard, dans certains cas elles

sont faibles (pour les données sur le maïs, l'erreur-type est de 5,704 et le c.v., de 0,036 pour le comté de Franklin). Comme les tailles d'échantillon sont très faibles, ces erreurs-types d'échantillon ne peuvent être considérées comme des approximations fiables des erreurs-types réelles. Le tableau 1 présente les données de niveau régional modifiées pour le maïs et le soja produites d'après les données au niveau unitaire de Battese et coll. (1988).

Tableau 1
Données de niveau régional modifiées sur les cultures, d'après Battese, Harter et Fuller (1988)

		Maïs			Soja		
Pays	n_i	y_i	e.-t.	c.v.	y_i	e.-t.	c.v.
Franklin	3	158,623	5,704	0,036	52,473	16,425	0,313
Pocahontas	3	102,523	43,406	0,423	118,697	50,290	0,424
Winnebago	3	112,773	30,547	0,271	88,573	10,453	0,118
Wright	3	144,297	53,999	0,374	97,800	52,034	0,532
Webster	4	117,595	21,298	0,181	112,980	23,531	0,208
Hancock	5	109,382	15,661	0,143	117,478	17,209	0,146
Kossuth	5	110,252	12,112	0,110	117,844	20,954	0,178
Hardin	5	120,054	36,807	0,307	101,834	26,790	0,263

Données sur le lait : Les données sur le lait, utilisées dans un article publié par Arora et Lahiri (1997), proviennent du U.S. Bureau of Labor Statistics. Les valeurs estimées sont les dépenses moyennes en lait frais pour 1989. L'ensemble contient des données sur 43 régions dont la taille d'échantillon varie de 95 à 633. Les c.v. varient de 0,074 à 0,341 sur les 43 régions. Le lecteur trouvera une description plus détaillée des données dans Arora et Lahiri (1997). Par souci de complétude, nous présentons les données au tableau 2. À l'instar d'Arora et Lahiri (1997), nous utilisons $x_i'\beta = \beta_j$ si $i \in j^{\text{e}}$ grande région (série de régions semblables pour la publication). Arora et Lahiri (1997) ont utilisé huit grandes régions. Puisque cette division en huit grandes régions n'est pas décrite dans leur article, après avoir relevé les tendances dans les données, nous avons utilisé le modèle de Fay-Herriot pour tester deux nouvelles divisions en six et en quatre grandes régions obtenues en regroupant les estimations par sondage semblables. En général, l'utilisation de ces grandes régions produit une réduction importante des c.v. Alors que les six groupes ont produit une réduction moyenne des c.v. d'environ 20 %, les quatre groupes ont donné une réduction moyenne d'environ 25 % des c.v. comparativement aux estimations directes. La comparaison des estimations ponctuelles et des c.v. montre que l'utilisation des quatre grandes régions donne de meilleurs résultats que l'utilisation des six grandes régions. Les quatre grandes régions sont 1–7, 8–14, 15–25 et 26–43. Ici, nous utiliserons ces quatre groupes comme variables auxiliaires aux fins d'illustration.

Tableau 2
Donnée sur le lait, tirées de Arora et Lahiri (1997)

Petite région	n_i	y_i	e.-t.	c.v.
1	191	1,099	0,163	0,148
2	633	1,075	0,080	0,074
3	597	1,105	0,083	0,075
4	221	0,628	0,109	0,174
5	195	0,753	0,119	0,158
6	191	0,981	0,141	0,144
7	183	1,257	0,202	0,161
8	188	1,095	0,127	0,116
9	204	1,405	0,168	0,120
10	188	1,356	0,178	0,131
11	149	0,615	0,100	0,163
12	290	1,460	0,201	0,138
13	250	1,338	0,148	0,111
14	194	0,854	0,143	0,167
15	184	1,176	0,149	0,127
16	193	1,111	0,145	0,131
17	218	1,257	0,135	0,107
18	266	1,430	0,172	0,120
19	214	1,278	0,137	0,107
20	213	1,292	0,163	0,126
21	196	1,002	0,125	0,125
22	95	1,183	0,247	0,209
23	195	1,044	0,140	0,134
24	187	1,267	0,171	0,135
25	479	1,193	0,106	0,089
26	230	0,791	0,121	0,153
27	186	0,795	0,121	0,152
28	199	0,759	0,259	0,341
29	238	0,796	0,106	0,133
30	207	0,565	0,089	0,158
31	165	0,886	0,225	0,254
32	153	0,952	0,205	0,215
33	210	0,807	0,119	0,147
34	383	0,582	0,067	0,115
35	255	0,684	0,106	0,155
36	226	0,787	0,126	0,160
37	224	0,440	0,092	0,209
38	212	0,759	0,132	0,174
39	211	0,770	0,100	0,130
40	179	0,800	0,113	0,141
41	312	0,756	0,083	0,110
42	241	0,865	0,121	0,140
43	205	0,640	0,129	0,202

3.2 Analyse des résultats

Données sur le maïs et le soja : Pour commencer, nous examinons l'effet de notre traitement de σ_i^2 en utilisant l'approche HB. Le tableau 3 donne les estimations HB, $\hat{\theta}_i^{\text{HB}}$, et les erreurs-types (e.-t.) et les coefficients de variation (c.v.) connexes pour les ensembles de données de niveau régional pour le maïs et le soja. L'erreur-type est la racine carrée de la variance a posteriori. Sous le modèle 1 (σ_i^2 inconnue), les e.-t. et les c.v. sont systématiquement plus élevés que les valeurs correspondantes sous le modèle 2 ($\sigma_i^2 = s_i^2$ connue). L'accroissement des e.-t. et des c.v. sous le modèle 1 est prévisible, puisque ce modèle tient compte

de la variabilité supplémentaire due à l'estimation de σ_i^2. En moyenne, l'accroissement des e.-t. et des c.v. est de l'ordre de 20 % (ce calcul exclut le comté de Franklin pour les données sur le maïs). Les résultats confirment que si l'on suppose que $\sigma_i^2 = s_i^2$, l'estimation directe connue de σ_i^2, on obtient une sous-estimation de l'erreur-type et du coefficient de variation de $\hat{\theta}_i$. L'examen des comtés de Franklin et de Webster pour les données sur le maïs et du comté de Winnebago pour les donnée sur le soja établit que, dans certains cas où les erreurs d'échantillonnage sont, par hasard, assez faibles, cette sous-estimation est importante.

Tableau 3
Comparaison des estimations HB pour les données sur les cultures

	σ_i^2 connue ($\sigma_i^2 = s_i^2$)			σ_i^2 inconnue		
Comté	$\hat{\theta}_i^{HB}$	e.-t.	c.v.	$\hat{\theta}_i^{HB}$	e.-t.	c.v.
			Maïs			
Franklin	155,788	6,061	0,039	142,862	18,408	0,129
Pocahontas	100,813	28,297	0,281	91,560	32,420	0,356
Winnebago	115,337	28,406	0,246	113,130	35,207	0,311
Wright	131,630	28,345	0,215	123,547	30,764	0,250
Webster	109,030	20,634	0,189	97,856	29,834	0,307
Hancock	121,682	15,656	0,129	123,478	17,857	0,145
Kossuth	115,710	11,180	0,097	114,910	12,510	0,109
Hardin	135,626	23,228	0,171	135,178	23,804	0,176
			Soja			
Franklin	75,375	16,272	0,216	88,186	21,067	0,239
Pocahontas	116,943	27,031	0,231	109,052	30,098	0,276
Winnebago	87,525	10,304	0,118	88,053	18,854	0,214
Wright	104,184	23,671	0,227	105,825	24,497	0,232
Webster	115,510	20,789	0,180	109,455	25,801	0,236
Hancock	101,368	15,741	0,155	102,876	17,311	0,169
Kossuth	102,388	14,948	0,146	101,862	15,019	0,148
Hardin	87,455	17,774	0,203	93,397	20,251	0,217

La comparaison des estimations HB sous les modèles 1 et 2 aux estimations directes peut se faire en se servant des c.v. présentés aux tableaux 1 et 3. Sous le modèle 2, les estimations HB ont un c.v. plus petit que les estimations directes pour six des huit comtés pour les données sur le maïs et, de même, dans six des huit comtés pour le données sur le soja. Dans le cas des deux cultures, pour les deux comtés restants, les c.v. sous le modèle 2 sont les mêmes ou légèrement plus grands que les c.v. des estimations directes par sondage. Par conséquent, les estimateurs provenant du modèle 2 semblent être plus efficaces que les estimateurs directs par sondage. Par contre, l'examen des estimations HB sous le modèle 1 et des estimations directes par sondage produit des résultats mixtes pour les ensembles de donnée sur le maïs et le soja. Le modèle 1 tient compte de l'incertitude supplémentaire due à l'estimation des variances d'échantillonnage et, par conséquent, les estimations HB ne sont meilleures dans le cas des données sur le maïs que pour quatre des huit comtés. Dans le cas des données sur le soja, les estimations HB représentent une amélioration par rapport aux c.v. des estimations directes par sondage pour cinq des huit comtés. Pour les autres, les c.v. des estimations directes sont plus faibles, voire même considérablement plus faibles dans certains cas. Pour les données sur le maïs, les c.v. des estimations pour les comtés de Franklin et de Webster augmentent de plus de 0,09 et 0,12, respectivement, dans le cas du modèle 1. En outre, pour les données sur le soja, le c.v. pour le comté de Winnebago augmente de près de 0,10 par rapport à l'estimation directe par sondage lorsqu'on utilise le modèle 1. Les régions où les estimations directes ont un c.v. plus faible que les estimations HB correspondantes comprennent plusieurs régions où les c.v. sont, par hasard, anormalement petits. Donc, le c.v. plus élevé produit par le modèle reflète une valeur plus appropriée pour ces régions. Parmi les sept cas où le c.v. direct est plus petit que le c.v. HB sous le modèle 1, pour les trois cas susmentionnés, l'écart est important et pour les quatre autres, l'utilisation du modèle 1 ne cause qu'une légère réduction d'efficacité. Puisque les estimations directes par sondage produisent fréquemment des c.v. inacceptablement grands, mais peuvent néanmoins donner par hasard des c.v. anormalement et inexplicablement faibles, l'estimation HB sous le modèle 1 pourrait être plus fiable et raisonnable, parce qu'elle tient compte de l'incertitude due à l'estimation de σ_i^2.

Données sur le lait : Le tableau 4 contient les estimations HB pour les données sur le lait. Comme prévu, sur l'ensemble des 43 régions, le fait de supposer que la variance σ_i^2 est connue ou inconnue donne lieu à une variation négligeable des estimations ponctuelles, des erreurs-types et des coefficients de variation, étant donné la grande taille des échantillons pour les 43 régions. Par conséquent, la substitution de $\sigma_i^2 = s_i^2$ dans le modèle est raisonnable, lorsque les tailles des échantillons régionaux sont grandes, comme l'illustre clairement cet exemple. En outre, les e.-t. et les c.v. des estimations HB sont plus petits que ceux des estimations directes par sondage présentées au tableau 2. Comme il faut s'y attendre, l'approche HB représente donc une amélioration par rapport aux estimations directes par sondage.

3.3 Lois a priori et analyse de sensibilité

Dans le modèle 1, nous supposons que les variances d'échantillonnage σ_i^2 sont indépendantes et suivent une loi a priori gamma inverse $\mathrm{GI}(a_i, b_i)$, et que la variance sous le modèle σ_v^2 suit aussi une loi a priori gamma inverse $\mathrm{GI}(a_0, b_0)$, où a_i, b_i $(0 \le i \le m)$ sont des constantes connues fixées à une valeur très faible afin de refléter les connaissances vagues au sujet de σ_i^2 et σ_v^2. Donc, nous avons utilisé les lois a priori appropriées afin d'éviter que toute loi a posteriori soit inappropriée. Nous pourrions envisager d'utiliser des lois a priori uniformes pour σ_i^2 et σ_v^2, c'est-à-dire $\pi(\sigma_i^2) \propto 1$, et $\pi(\sigma_v^2) \propto 1$, semblables à la loi a priori uniforme sur β. Avec les lois a priori uniformes

sur σ_i^2 et σ_v^2, les lois conditionnelles complètes pour σ_i^2 et σ_v^2 sont données par

$$[\sigma_i^2 | y, \theta, \beta, \sigma_v^2] \sim \text{GI}\left(\frac{d_i - 1}{2}, \frac{(y_i - \theta_i)^2 + d_i s_i^2}{2}\right),$$

et

$$[\sigma_v^2 | y, \theta, \beta, \sigma_i^2] \sim \text{GI}\left(\frac{m-2}{2}, \frac{1}{2}\sum_{i=1}^{m}(\theta_i - x_i'\beta)^2\right).$$

Tableau 4
Comparaison des estimations HB pour les données sur le lait

Petite région	σ_i^2 connue ($\sigma_i^2 = s_i^2$)			σ_i^2 inconnue		
	$\hat{\theta}_i^{HB}$	e.-t.	c.v.	$\hat{\theta}_i^{HB}$	e.-t.	c.v.
1	1,020	0,113	0,111	1,021	0,111	0,109
2	1,045	0,072	0,069	1,045	0,071	0,068
3	1,065	0,073	0,069	1,065	0,074	0,069
4	0,767	0,095	0,124	0,770	0,096	0,125
5	0,849	0,096	0,113	0,852	0,096	0,113
6	0,975	0,103	0,106	0,975	0,102	0,105
7	1,058	0,125	0,118	1,055	0,125	0,118
8	1,097	0,099	0,090	1,096	0,099	0,090
9	1,219	0,121	0,099	1,215	0,121	0,100
10	1,192	0,122	0,102	1,190	0,122	0,102
11	0,793	0,094	0,119	0,799	0,097	0,122
12	1,213	0,131	0,108	1,209	0,130	0,107
13	1,206	0,112	0,093	1,203	0,112	0,093
14	0,984	0,107	0,109	0,987	0,107	0,109
15	1,187	0,105	0,088	1,187	0,104	0,087
16	1,156	0,104	0,090	1,156	0,102	0,089
17	1,225	0,101	0,083	1,225	0,100	0,081
18	1,284	0,115	0,089	1,281	0,113	0,088
19	1,234	0,101	0,082	1,235	0,100	0,081
20	1,233	0,110	0,089	1,233	0,110	0,089
21	1,092	0,097	0,089	1,095	0,098	0,089
22	1,192	0,128	0,107	1,193	0,127	0,106
23	1,122	0,103	0,092	1,125	0,103	0,091
24	1,221	0,113	0,092	1,220	0,111	0,091
25	1,193	0,086	0,072	1,193	0,086	0,072
26	0,761	0,091	0,120	0,762	0,091	0,120
27	0,763	0,092	0,120	0,762	0,091	0,119
28	0,734	0,125	0,170	0,732	0,123	0,169
29	0,768	0,085	0,110	0,767	0,085	0,110
30	0,615	0,076	0,124	0,618	0,076	0,123
31	0,769	0,122	0,158	0,767	0,120	0,156
32	0,795	0,119	0,150	0,792	0,118	0,148
33	0,771	0,091	0,118	0,770	0,090	0,117
34	0,612	0,060	0,099	0,613	0,062	0,100
35	0,701	0,085	0,121	0,701	0,084	0,120
36	0,757	0,094	0,123	0,759	0,093	0,123
37	0,534	0,080	0,150	0,538	0,081	0,151
38	0,744	0,096	0,129	0,743	0,095	0,128
39	0,754	0,082	0,108	0,753	0,082	0,108
40	0,768	0,088	0,115	0,768	0,088	0,115
41	0,747	0,071	0,095	0,747	0,070	0,094
42	0,801	0,093	0,116	0,800	0,092	0,116
43	0,682	0,094	0,139	0,682	0,094	0,138

L'application de l'échantillonneur de Gibbs sous les lois a priori uniformes est également simple. Cependant, les lois a priori uniformes sur σ_i^2 et σ_v^2 peuvent mener à des lois a posteriori, ou posteriors, inappropriées si les tailles d'échantillon et le nombre de petites régions sont faibles. Pour mieux visualiser le problème des lois a priori sur σ_i^2, nous pouvons étudier le modèle 1 en deux étapes. En premier lieu, nous pouvons obtenir la loi a posteriori de σ_i^2, connaissant l'estimation directe s_i^2 de cette dernière, sous la forme

$$\begin{aligned}\pi(\sigma_i^2 | s_i^2) &\propto f(s_i^2 | \sigma_i^2) \cdot \pi(\sigma_i^2) \\ &\propto (\sigma_i^2)^{-d_i/2} \cdot \exp\{-\sigma_i^{-2} d_i s_i^2 / 2\} \cdot \pi(\sigma_i^2).\end{aligned}$$

En postulant une loi a priori uniforme $\pi(\sigma_i^2) \propto 1$, nous obtenons

$$\pi(\sigma_i^2 | s_i^2) \sim \text{GI}\left(\frac{d_i}{2} - 1, \frac{d_i s_i^2}{2}\right),$$

à condition que $d_i > 2$, ou $n_i > 3$. Alors, nous pouvons utiliser cette loi a posteriori GI approprié $\pi(\sigma_i^2 | s_i^2)$ en tant que loi a priori informative sur σ_i^2 dans le modèle d'échantillonnage $y_i | \theta_i, \sigma_i^2 \sim \text{ind } N(\theta_i, \sigma_i^2)$. Cela assurera une inférence a posteriori correcte. Pour les données modifiées sur le maïs et le soja, l'utilisation des lois a priori uniformes sur σ_i^2 produira une loi a posteriori impropre, à cause de la petite taille de l'échantillon ($n_i = 3$) pour certaines régions. Donc, nous utilisons des lois a priori gamma inverses appropriées dans l'analyse des données pour nous assurer que toute les lois a posteriori soient corrects, comme cela est fait habituellement en pratique lors de l'estimation HB sur petites régions (par exemple, Arora et Lahiri 1997; Datta, Lahiri, Maiti et Lu 1999; You et Rao 2000; Rao 2003). Par conséquent, nous n'avons pas à craindre que certaines lois a posteriori soient inappropriés, puisque l'inférence HB correcte devrait être fondée sur des lois a posteriori appropriées. Sous le modèle 2 avec variance d'échantillonnage connue donnée par $\sigma_i^2 = s_i^2$, et l'utilisation d'une loi a priori uniforme $\pi(\sigma_v^2) \propto 1$ sur σ_v^2, la loi a posteriori de σ_v^2 sera approprié à condition que $m > p + 2$, où m est le nombre de petites régions et p est la taille des paramètres de régression β (Rao 2003, page 238). Puisque le nombre de petites régions est habituellement assez grand, cette conditions est en général satisfaite en pratique.

Pour l'analyse de sensibilité des lois a priori appropriées vagues, nous pouvons tester la sensibilité des estimations a posteriori au choix des paramètres a priori a_i, $b_i (0 \le i \le m)$. Sous le modèle 1, nous fixons $a_i = b_i$ à quatre valeurs différentes, c'est-à-dire 0,0001, 0,001, 0,01 et 0,1. Le tableau 5 donne les estimations des moyennes a posteriori pour les données sur le maïs et le soja, et le tableau 6, les c.v. correspondants.

Il est évident, si l'on examine les tableaux 5 et 6, que les estimations a posteriori et les c.v. correspondants sont à peu près les mêmes et stables, ce qui indique que les estimations HB ne sont pas sensibles au choix des lois a priori appropriées vagues. Dans le cas des données sur le lait, les estimations HB sont très stables au choix de ces lois a priori

appropriées vagues (résultats non présentés ici). Puisque les données sur le lait proviennent d'échantillons de grande taille, nous pouvons également utiliser des lois a priori uniformes sur les composantes de la variance pour les analyser sous le modèle 1. Nous obtenons donc les estimations HB fondées sur les lois a priori uniformes et les comparons aux estimations HB fondées sur les lois a priori GI vagues. Ces estimations HB sont presque identiques et stables, l'écart relatif variant de 0,07 % à 2,23 %, avec une valeur moyenne de 0,69 % sur 43 régions, ce qui indique que les estimations a posteriori des moyennes de petite région fondées sur le modèle 1 sont très stables et ne sont pas sensibles au choix des lois a priori uniformes ni des lois a priori GI vagues, à condition que les tailles d'échantillons et le nombre de petites régions soient relativement grands.

Tableau 5
Comparaison des estimations des moyennes a posteriori pour les données sur le maïs

	GI $(a_i, b_i), a_i = b_i$			
Comté	0,0001	0,001	0,01	0,1
		Maïs		
Franklin	142,862	142,593	143,155	144,311
Pocahontas	91,560	91,912	91,422	91,974
Winnebago	113,130	113,068	121,578	114,430
Wright	123,547	124,170	125,103	125,351
Webster	97,856	98,231	99,132	98,511
Hancock	123,478	123,858	124,395	124,138
Kossuth	114,910	115,281	115,316	115,528
Hardin	135,178	134,157	135,223	136,001
		Soja		
Franklin	88,186	89,368	89,145	89,513
Pocahontas	109,052	109,571	107,745	108,176
Winnebago	88,053	87,478	86,267	87,302
Wright	105,825	106,712	105,142	104,676
Webster	109,455	108,392	109,835	110,252
Hancock	102,876	103,413	102,240	101,808
Kossuth	101,862	101,159	101,379	100,808
Hardin	93,397	94,713	93,576	94,767

Tableau 6
Comparaison des c.v. a posteriori pour les données sur le maïs

	GI $(a_i, b_i), a_i = b_i$			
Comté	0,0001	0,001	0,01	0,1
		Maïs		
Franklin	0,129	0,124	0,128	0,125
Pocahontas	0,356	0,351	0,347	0,341
Winnebago	0,311	0,314	0,321	0,324
Wright	0,250	0,246	0,235	0,236
Webster	0,307	0,292	0,285	0,280
Hancock	0,145	0,148	0,148	0,142
Kossuth	0,109	0,110	0,107	0,104
Hardin	0,176	0,173	0,178	0,168
		Soja		
Franklin	0,239	0,233	0,231	0,227
Pocahontas	0,276	0,281	0,271	0,296
Winnebago	0,214	0,193	0,196	0,198
Wright	0,232	0,223	0,231	0,226
Webster	0,236	0,231	0,237	0,228
Hancock	0,169	0,165	0,168	0,161
Kossuth	0,148	0,145	0,142	0,135
Hardin	0,217	0,215	0,213	0,213

4. Conclusion et travaux futurs

Dans le présent article, nous avons étudié le modèle bien connu de Fay-Herriot dans les situations où il est supposé que σ_i^2, la variance d'erreur d'échantillonnage est inconnue et estimée au moyen de l'estimateur sans biais s_i^2, en utilisant l'approche hiérarchique bayésienne. L'approche HB complète avec échantillonnage de Gibbs tient compte automatiquement de l'incertitude supplémentaire associée à l'estimation de σ_i^2. Nous avons appliqué l'approche HB à l'analyse de deux ensembles de données d'enquête. Nos résultats montrent que l'approche HB proposée sous le modèle 1 donne d'assez bons résultats, que les tailles des échantillons régionaux soient de grande ou de petite taille. Lors de futurs travaux, l'approche de modélisation HB proposée pourrait être étendue aux modèles de niveau régional généraux étudiés par You et Rao (2002). Les applications de la nouvelle approche de modélisation HB comprennent l'estimation du sous-dénombrement au recensement décrite dans You, Rao et Dick (2004). Sous le modèle 1, il est possible d'obtenir les estimateurs HB des variances d'échantillonnage σ_i^2. Ces estimateurs HB de σ_i^2 peuvent alors être utilisés comme estimateurs lissés de rechange pour σ_i^2 dans les modèles d'échantillonnage. Les applications et évaluations des estimateurs HB des variances d'échantillonnage comprennent l'estimation du sous-dénombrement au recensement et l'estimation du taux de chômage dans le cadre de l'Enquête sur la population active (EPA) du Canada (You, Rao et Gambino 2003). Nous prévoyons aussi comparer l'approche HB à l'approche EBLUP telle qu'elle a été étudiée par Rivest et Vandal (2002), ainsi que par Wang et Fuller (2003).

Remerciements

Les auteurs tiennent à remercier deux examinateurs, un rédacteur adjoint, le rédacteur en chef délégué et le rédacteur en chef M.P. Singh, de leurs suggestions et commentaires constructifs. Les auteurs remercient aussi J.N.K. Rao, de l'Université Carleton, pour ses suggestions utiles, ainsi que Jack Gambino et Eric Rancourt, de Statistique Canada, pour leurs commentaires au sujet de la première version de l'article. Ces travaux ont été financés grâce aux ressources de financement global de la recherche de la Direction de la Méthodologie de Statistique Canada.

Bibliographie

Arora, V., et Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Datta, G.S., Lahiri, P., Maiti, T. et Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.

Fay, R.E., et Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 268-277.

Gelfand, A.E., et Smith, A.F.M. (1990). Sample-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 972-985.

Gelfand, A.E., et Smith, A.F.M. (1991). Gibbs sampling for marginal posterior expectations. *Communications In Statisitcs – Theory and Methods*, 20, 1747-1766.

Rao, J.N.K. (1999). Quelques progrès concernant l'estimation régionale fondée sur un modèle. *Techniques d'enquête*, 25, 199-212.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Rivest, L.P., et Vandal, N. (2002). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, 10-13 juillet, 2002, Ottawa, Canada.

Wang, J., et Fuller, W.A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y., et Rao, J.N.K. (2000). Estimation bayésienne hiérarchique des moyennes pour petites régions à l'aide de modèles à plusieurs niveaux. *Techniques d'enquête*, 26, 197-206.

You, Y., et Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 30, 1, 3-15.

You, Y., Rao, J.N.K. et Dick, P. (2004). Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, 6, 631-640.

You, Y., Rao, J.N.K. et Gambino, J. (2003). Estimation du taux de chômage fondée sur un modèle pour l'Enquête sur la population active du Canada : Une approche bayésienne hiérarchique. *Techniques d'enquête*, 29, 27-36.

Techniques d'enquête, juin 2006
Vol. 32, N° 1, pp. 115-125
Statistique Canada, N° 12-001-XIF au catalogue

Une stratégie rentable d'estimation du chômage au niveau provincial : Une approche d'estimation pour petits domaines

Ali-Reza Khoshgooyanfard et Mohammad Taheri Monazzah [1]

Résumé

L'objectif principal de l'article est de proposer une stratégie rentable d'estimation du taux de chômage intercensitaire au niveau provincial en Iran. Cette stratégie, qui tire parti des méthodes d'estimation pour petits domaines, s'appuie sur un échantillonnage unique au niveau national. Trois méthodes, basées respectivement sur un estimateur synthétique, un estimateur composite et un estimateur empirique bayésien, sont utilisées pour calculer les estimations d'intérêt indirectes pour 1996. Les résultats confirment non seulement que la stratégie proposée est appropriée, mais montrent aussi que l'estimateur composite et l'estimateur empirique bayésien produisent de bonnes estimations et ont des propriétés semblables.

Mots clés : Estimateur composite, estimateur fondé sur le plan de sondage, estimateur empirique bayésien, estimateur indirect, erreur non due à l'échantillonnage, estimateur synthétique, post-strate.

1. Introduction

Chaque année, des enquêtes par sondage sont réalisées en Iran afin d'obtenir les renseignements statistiques nécessaires à la prise de décisions et à l'élaboration des politiques. Cependant, ces enquêtes ne permettent pas de répondre à tous les besoins de données statistiques, pour deux raisons. Premièrement, les secteurs public et privé demandent des données statistiques complètes non seulement de niveau national et régional, mais aussi pour des petits domaines. En outre, ils ont besoin de cette information pour des périodes plus courtes qu'une année, disons mensuellement ou trimestriellement. Deuxièmement, en Iran, les enquêtes sont la source principale de données statistiques, mais des contraintes budgétaires limitent la réalisation d'enquêtes plusieurs fois par année au niveau des petits domaines. Ces deux facteurs obligent les organismes statistiques à concevoir des stratégies efficaces permettant de réaliser un juste équilibre entre le coût et la qualité de l'information statistique. Les travaux présentés ici représentent un effort en vue de relever ce défi en utilisant des méthodes d'estimation pour petits domaines.

Le but des méthodes d'estimation pour petits domaines est de produire des estimations acceptables pour certaines sous- populations, quelles qu'elles soient, dans le cadre d'un plan de sondage prévu pour l'« ensemble » de la population. Prenons l'exemple où on établit un plan de sondage en vue d'estimer les paramètres de population pour le « pays » et où, après la collecte des données, on estime ces paramètres d'après les données de l'échantillon national. Si l'on a besoin, simultanément, d'« estimations provinciales » des paramètres, il est impossible de réaliser des sondages provinciaux distincts. Les provinces sont des sous-populations non planifiées, en ce sens que le plan de sondage utilisé a été établi uniquement pour l'estimation des paramètres au niveau national, sans tenir compte du niveau provincial. Dans l'échantillon national, le nombre d'unités d'échantillonnage disponibles pour certaines provinces est faible ou nul. Par conséquent, il est impossible de produire des estimations acceptables pour ces provinces (sous-populations).

Avant la mise au point de méthodes d'estimation pour petits domaines, les estimations pour les sous-populations étaient obtenues par estimation directe fondée sur le plan de sondage. S'il existait des données provenant d'une sous-population particulière dans l'échantillon national, une estimation était calculée directement, conformément au plan de sondage national, en utilisant les « données disponibles ». Or, l'estimation directe peut différer sensiblement du paramètre réel de sous-population, à cause d'erreurs d'échantillonnage importantes dues à la petite taille de l'échantillon.

Les statisticiens et les démographes ont établi des moyens de produire des estimations pour ce genre de sous-populations. Ils ont proposé des estimateurs indirects, dont les applications se sont multipliées au cours des vingt dernières années. Néanmoins, les méthodes d'estimation pour petits domaines continuent de faire l'objet de nombreuses études. Consulter Purcell et Kish (1979, 1980), Ghosh et Rao (1994), Schaible (1995), Marker (1999), Pfeffermann (2002) et, surtout, Rao (2003a), pour la définition du problème et un examen des méthodes d'estimation pour petits domaines.

Pendant plusieurs années, le Centre statistique d'Iran (CSI) a procédé annuellement au tirage d'un échantillon national en grappes à un degré afin d'estimer le taux de chômage intercensitaire au niveau national. Pendant seize ans, un échantillon en grappes à un degré distinct a été

1. Ali-Reza Koshgooyanfard, The Center for Research, Studies and Program Assessments de l'IRIB. Courriel : khosh_ar@yahoo.com; Mohammad Taheri Monazzah, Banque centrale d'Iran. Courriel : Taheri53@yahoo.com.

sélectionné pour chaque province afin d'estimer les taux de chômage provinciaux. Le taux de chômage pour l'ensemble du pays était ensuite calculé par combinaison pondérée des estimations provinciales. La demande croissante d'estimations du taux de chômage au niveau provincial sur une base mensuelle, ou du moins saisonnière, et le manque de dossiers administratifs en Iran au niveau tant régional que national ont persuadé le CSI d'essayer d'adopter les méthodes d'estimation pour petits domaines comme élément central d'une stratégie révisée en vue de répondre aux besoins des provinces.

La stratégie révisée consiste à concevoir un plan de sondage uniquement au niveau national et à produire des estimations provinciales par des méthodes d'estimation pour petits domaines. Dans cette stratégie, une province représente un petit domaine. Cette stratégie requiert un échantillon de plus petite taille que celui résultant de l'agrégation des échantillons provinciaux. Si la stratégie révisée s'avère applicable en pratique, il sera possible de réduire la durée et le coût de la collecte des données et de produire des estimations provinciales mensuellement. Le plus petit échantillon est plus facile à gérer sur le terrain et les estimations sont moins affectées par les erreurs non dues à l'échantillonnage.

Le présent article vise à répondre aux questions suivantes :

1. Un échantillon national peut-il remplacer les échantillons provinciaux distincts pour estimer les taux de chômage provinciaux?
2. Parmi les trois méthodes d'estimation pour petits domaines, à savoir les estimateurs synthétique, composite et empirique bayésien, laquelle produit les meilleures estimations?

Afin de répondre empiriquement à ces deux questions, nous avons produit des estimations pour 1996, année pour laquelle les valeurs réelles des taux de chômage provinciaux sont connues grâce au Recensement de 1996. Par conséquent, nous pouvons calculer le biais réel de chaque estimation provinciale.

Le processus comprend les trois étapes qui suivent. Premièrement, un échantillon de 13 000 unités a été sélectionné pour l'ensemble du pays (le fichier de données du Recensement de 1996). La taille de l'échantillon a été déterminée au niveau national, puis répartie entre les provinces proportionnellement à leur population. La répartition fournit pour chaque province un échantillon qui permet d'estimer directement le taux de chômage provincial. Les estimations directes ne sont pas nécessairement acceptables pour toutes les provinces, à cause des erreurs d'échantillonnage importantes dues à la petite taille de l'échantillon pour certaines provinces. Deuxièmement, trois méthodes d'estimation pour petits domaines ont été appliquées pour produire des estimations indirectes pour chaque province. Troisièmement, les estimations indirectes ont été évaluées par comparaison aux valeurs réelles correspondantes, en se basant sur le calcul de l'erreur quadratique moyenne (EQM), de l'erreur absolue moyenne (EAH) et de l'erreur moyenne (EM).

Outre cette introduction, l'article contient trois autres sections. À la section 2, nous passons brièvement en revue les trois estimateurs utilisés, y compris les méthodes d'estimation, les EQM correspondantes et les propriétés des estimateurs. À la section 3, nous présentons les estimations et les méthodes de calcul correspondantes, et nous essayons d'évaluer les propriétés des estimateurs. À la section 4, nous présentons nos conclusions en ce qui concerne les estimateurs et les avantages de la stratégie d'estimation pour petits domaines, et nous formulons des recommandations.

2. Aperçu des estimateurs

Nous nous contentons de présenter brièvement les estimateurs indirects utilisés pour l'étude. Cependant, le lecteur trouvera une excellente discussion des méthodes d'estimation pour petits domaines dans Rao (2003a). Nous examinons d'abord l'estimateur synthétique, puis l'estimateur composite. Nous considérons aussi l'estimateur empirique bayésien (EB) à titre d'estimateur fondé sur un modèle.

2.1 Estimateur synthétique

Il existe une famille d'estimateurs sur petits domaines qui sont qualifiés de synthétiques, voir Rao (2003a, chapitre 4). Nous décrivons ici celui qui est le plus classique et le plus simple. Pour cet estimateur,

1. le pays est divisé en six post-strates en fonction de six groupes d'âge (voir tableau 1);
2. puis, le nombre de chômeurs est estimé dans chaque province, ce qui donne le numérateur de l'expression (1);
3. enfin, l'estimateur synthétique pour la i^{e} province est obtenu en divisant le nombre estimé de chômeurs dans la province i par la population économiquement active (PEA) de la province, c'est-à-dire

$$\hat{P}_i^S = \left(\sum_{j=1}^{6} N_{ij} \hat{P}_j \right) \Big/ N_i \qquad (1)$$

où $\hat{P}_j$ est une estimation directe fondée sur le plan de sondage du taux de chômage dans la post-strate j, N_i est la PEA de la province i et N_{ij} est la PEA dans l'intersection de la province i et de la post-strate j, c'est-à-dire la cellule (i, j). L'estimation synthétique pour la i^{e} province est calculée conformément à la définition officielle du taux de chômage en Iran.

L'estimation synthétique s'appuie sur toutes les données de l'échantillon national grâce à l'utilisation des estimations directes nationales du taux de chômage d'après les post-strates. Elle est basée sur les six estimations du taux de chômage par « post-strate » calculées sur l'ensemble des provinces, plutôt que sur les estimations spécifiques des six « cellules ». Par conséquent, ce processus revient à **emprunter de la force (information)**, puisque chaque province contribue à l'échantillon national grâce au regroupement des unités d'échantillonnage provinciales en vue de surmonter les problèmes posés pour chaque province par la petite taille de l'échantillon.

Cet estimateur a trois limites :

1. L'estimateur synthétique donne des résultats d'autant meilleurs que la variation entre les post-strates est faible. Autrement dit, le taux de chômage dans chaque groupe d'âge devrait être à peu près le même dans toutes les provinces. L'utilisation des estimations directes nationales par post-strate de façon uniforme pour toutes les provinces n'est admissible que sous cette hypothèse. Si l'hypothèse d'homogénéité n'est pas satisfaite, l'estimateur synthétique ne peut pas refléter les variations particulières au niveau des petits domaines et les estimations pourraient être gravement biaisées.

2. S'il existe plusieurs variables importantes pour la post-stratification, il est fréquent qu'on ne puisse pas les utiliser toutes dans l'estimateur synthétique, parce que la taille des échantillons des post-strates (après recoupement de plusieurs variables) est trop faible et produit des estimations directes inacceptables au niveau de la post-strate. Généralement parlant, un grand nombre de post-strates donne lieu à des estimations directes de mauvaise qualité pour certaines post-strates. Cette situation peut créer de sérieux problèmes lors de l'estimation synthétique, si elle est associée à une grande PEA dans une cellule.

3. La qualité des estimations de la PEA peut avoir une incidence sur les estimations synthétiques. Étant donné le manque de sources de données à jour, comme les dossiers administratifs, des estimations périmées de la PEA d'après les données du Recensement de 1986 sont utilisées ici pour produire les estimations synthétiques pour 1996.

2.2 Estimateur composite

L'estimateur composite pour la i^{e} province est une combinaison des estimateurs synthétique et direct pour cette province, à savoir

$$\hat{P}_i^C = W_i \hat{P}_i^D + (1 - W_i)\hat{P}_i^S \qquad (2)$$

où $\hat{P}_i^D$ est l'estimateur direct fondé sur le plan de sondage pour la i^{e} province et $0 \le W_i \le 1$. L'expression (2) est une amélioration de l'expression (1) grâce à l'exploitation des deux estimateurs. Autrement dit, dans l'estimateur composite, les écarts interprovinciaux sont pris en compte au moyen des estimations provinciales directes sans biais et l'instabilité de l'estimateur direct est réduite au moyen de l'estimateur synthétique.

Le poids W_i peut être spécifié de façon à réduire au minimum l'erreur quadratique moyenne de $\hat{P}_i^C$, $\text{EQM}(\hat{P}_i^C)$. Si l'on suppose que $\text{Cov}(\hat{P}_i^D, \hat{P}_i^S) \cong 0$, l'expression du poids se simplifie comme suit

$$W_i^{\text{opt}} = \frac{1}{\left(V(\hat{P}_i^D)/\text{EQM}(\hat{P}_i^S)\right) + 1} \qquad (3)$$

où $V(\hat{P}_i^D)$ et $\text{EQM}(\hat{P}_i^S)$ sont la variance de $\hat{P}_i^D$ et l'erreur quadratique moyenne de $\hat{P}_i^S$, respectivement. Dans l'expression (3), les poids des estimateurs direct et synthétique figurant dans (2) sont proportionnels aux EQM des deux estimateurs. Voir Schaible (1978) et Rao (2003a, page 58) pour les propriétés de l'estimateur et du poids.

En pratique, nous devrions estimer $\text{EQM}(\hat{P}_i^S)$ et $V(\hat{P}_i^D)$ pour générer une estimation du poids (3). S'il existe des données d'échantillon provenant de la i^{e} province, d'après le plan de sondage, nous pouvons calculer un estimateur fondé sur le plan de sondage sans biais de $V(\hat{P}_i^D)$ en utilisant uniquement les données de l'échantillon. Par conséquent, un seul estimateur est nécessaire pour $\text{EQM}(\hat{P}_i^S)$. Sous l'hypothèse que $\text{Cov}(\hat{P}_i^D, \hat{P}_i^S) \cong 0$, Ghosh et Rao (1994) ont proposé l'estimateur sans biais

$$\hat{\text{EQM}}(\hat{P}_i^S) = (\hat{P}_i^S - \hat{P}_i^D)^2 - \hat{V}(\hat{P}_i^D). \qquad (4)$$

Sous la même hypothèse, il est facile de montrer que

$$\text{EQM}(\hat{P}_i^C) = W_i^2 V(\hat{P}_i^D) + (1 - W_i)^2 \text{EQM}(\hat{P}_i^S). \qquad (5)$$

L'estimateur (4) produit parfois des estimations négatives pour certaines provinces, de sorte que le poids donné par l'expression (3) n'est plus calculable. Dans ce cas, au lieu de (3) et (4), nous avons utilisé, respectivement, le poids combiné donné par (6) et $\hat{\text{EQMM}} = (1/I')\sum_{i=1}^{I'} \hat{\text{EQM}}(\hat{P}_i^S)$, où I' est le nombre de petits domaines dont l'estimation de

l'EQM est positive (voir Gonzalez et Waksberg (1973) pour plus de précisions) :

$$W^C = \frac{1}{\left(\sum_i \hat{V}(\hat{P}_i^D) \Big/ \sum_i \hat{\text{EQM}}(\hat{P}_i^S)\right)+1}. \qquad (6)$$

En plus des expressions (3) et (6), Copas (1972), Ghosh et Rao (1994), ainsi que Thompsen et Holmoy (1998) proposent des poids de rechange.

2.3 Estimateur empirique bayésien (EB)

Plus d'attention a été accordée aux méthodes d'estimation pour petits domaines fondées sur un modèle qu'aux estimateurs synthétique et composite. Marker (1999) considère que les méthodes d'estimation pour petits domaines ont un élément commun exprimé au moyen de modèles de régression. La méthode EB rentre dans la catégorie des méthodes de régression. Considérons le modèle mixte suivant (voir Rao (2003a, page 76)) :

$$\boldsymbol{g} = X\boldsymbol{\beta} + \boldsymbol{\nu} + \boldsymbol{\varepsilon} \qquad (7)$$

où

$$\boldsymbol{g}' = (Ln\frac{\hat{P}_1^D}{1-\hat{P}_1^D},\ ...,\ Ln\frac{\hat{P}_I^D}{1-\hat{P}_I^D}),$$

X est une matrice de plan d'expérience de dimensions $I \times k$ de variables supplémentaires, $\underline{\beta}$ est un vecteur de dimensions $k \times 1$ de paramètres inconnus, et $\underline{\nu}$ et $\underline{\varepsilon}$ sont des vecteurs aléatoires de dimensions $I \times 1$ (I est le nombre de provinces). Supposons que :

1. $\boldsymbol{\nu}$ et $\boldsymbol{\varepsilon}$ sont indépendants;
2. $E(\boldsymbol{\varepsilon}) = 0$ et $\text{Var}(\boldsymbol{\varepsilon}) = \text{Diag}(d_1^2,...,d_I^2)$;
3. $\boldsymbol{\nu} \sim N(0, \Sigma)$, où $\boldsymbol{\Sigma} = \text{Diag}(t^2,\ ...,\ t^2)$.

Ghosh et Meeden (1997) montrent que l'estimation EB du i^{e} élément de $\boldsymbol{g}$ est :

$$\hat{g}_i^{\text{EB}} = \hat{W}_i \boldsymbol{x}_i' \hat{\beta} + (1-\hat{W}_i) g_i \qquad (8)$$

où $\boldsymbol{x}_i'$ et g_i sont la i^{e} ligne et la i^{e} composante de X et $\underline{g}$ respectivement, et $\hat{W}_i$ est une estimation de

$$W_i = \frac{d_i^2}{d_i^2 + t^2}. \qquad (9)$$

Par conséquent, l'estimation EB du i^{e} taux est :

$$\hat{P}_i^{\text{EB}} = \frac{\exp(\hat{W}_i \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} + (1-\hat{W}_i) g_i)}{1 + \exp(\hat{W}_i \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} + (1-\hat{W}_i) g_i)}. \qquad (10)$$

Il est évident que l'utilisation de (10) requiert deux estimations, celles de $\boldsymbol{\beta}$ et du poids donné par (9). Par ailleurs, ce poids repose sur les estimations de t^2 et de d_i^2. Par application de la méthode delta, $(g_i')^2 \hat{V}(\hat{P}_i^D)$ produit une estimation de d_i^2, où g_i' est la dérivée première de $g_i = Ln(\hat{P}_i^D / 1 - \hat{P}_i^D)$. En nous inspirant de Chand et Alexander (1995), nous obtenons les estimations de $\boldsymbol{\beta}$ et de t^2 en résolvant simultanément

$$\begin{cases} t^2 = (\boldsymbol{g} - X\boldsymbol{\beta})' V^{-1} (\boldsymbol{g} - X\boldsymbol{\beta}) / (I-k) \\ \boldsymbol{\beta} = (X' V^{-1} X)^{-1} X' V^{-1} \boldsymbol{g} \end{cases} \qquad (11)$$

où $V = \text{Diag}(d_1^2 + t^2, \ldots, d_I^2 + t^2)$. Soulignons que les équations (11) sont résolues par itération numérique en partant d'une valeur initiale pour t^2.

Les estimateurs EB et composite présentent des similarités, bien qu'ils soient obtenus en suivant des approches différentes. L'un et l'autre ont deux composantes, c'est-à-dire une composante directe ($\hat{P}_i^D$ dans (2) et g_i dans (8)) calculée d'après des données d'échantillon provincial et une composante indirecte ($\hat{P}_i^S$ dans (2) et $\boldsymbol{x}_i' \hat{\boldsymbol{\beta}}$ dans (8)) construite d'après les données de l'échantillon national et des renseignements supplémentaires. Les estimateurs (2) et (8) accordent tous deux plus de poids à la composante indirecte lorsqu'elle est fiable. Sinon, la composante directe reçoit plus de poids. Des précisions supplémentaires sont données dans Cressie (1989), Ghosh et coll. (1998) et Rao (2003 a, b).

3. Estimation pour l'Iran

Nous avons produit des estimations pour 1996, parce que les taux réels de chômage de 1996 sont connus pour chaque province grâce au Recensement de 1996. Par conséquent, il est possible de calculer le biais réel pour chaque estimation.

En 1996, le pays était constitué de 26 provinces. Cependant, nous n'en étudions que 21 ici, parce que les renseignements supplémentaires provenant du Recensement de 1986 n'étaient disponibles que pour 21 provinces dont les limites géographiques n'ont pas changé entre 1986 et 1996. Pour produire les trois estimations indirectes au niveau national, nous avons établi un plan de sondage en déterminant la taille d'échantillon nécessaire pour estimer le taux de chômage pour le pays dans son ensemble. Chaque province représente un petit domaine. L'échantillon national a été réparti entre les 26 provinces proportionnellement à la population de ces dernières afin de disposer de données d'échantillon provenant de chaque province (approche descendante). Cela nous a permis de calculer des estimations directes fondées sur le plan de sondage pour chaque province et les variances correspondantes requises pour les estimateurs EB et composite. Le plan de sondage permet de produire de bonnes estimations pour le pays dans son ensemble et pour certaines provinces.

3.1 Méthodes de calcul

Pour produire des estimations synthétiques, nous avons défini des post-strates en fonction de six groupes d'âge. Au tableau (1), nous présentons, pour chaque groupe, le taux de chômage basé sur l'échantillon national et sa valeur réelle correspondante basée sur le Recensement de 1996, ainsi que les erreurs absolues des estimations.

Tableau 1
Caractéristique des post-strates

Groupe d'âge	Taux estimé ($\hat{P}_j$)	Valeur réelle	Erreur absolue
10–15	0,3240	0,2826	0,0414
16–20	0,2402	0,2629	0,0227
21–25	0,1868	0,1856	0,0012
26–30	0,0811	0,0802	0,0009
31–50	0,0363	0,0366	0,0003
Plus de 50 ans	0,0653	0,0648	0,0005

Les estimations obtenues pour les deux premiers groupes sont entachées d'une très grande erreur. Par conséquent, dans l'expression (1), si une province fournit une grande PEA pour ces groupes d'âge, son estimateur synthétique pourrait ne pas donner de bons résultats. Nous avons utilisé les données du Recensement de 1986 pour calculer les PEA pour toutes les provinces et cellules (N_i et N_{ij} dans l'expression (1)), parce qu'en l'absence de données administratives, le recensement le plus rapproché de 1996 est la source principale de données à tout niveau.

Pour construire les estimations composites, nous avons réparti les provinces en deux groupes. Le premier comprend 14 provinces auxquelles nous avons appliqué le poids donné par l'expression (3) et le second, sept provinces auxquelles nous avons appliqué le poids commun $W^C = 0{,}873184$ fondé sur (6). Comme l'estimateur donné par l'expression (4) produit des estimations négatives de EQM($\hat{P}_i^S$) pour ces sept provinces, nous avons utilisé la moyenne des erreurs quadratiques moyennes (EQMM).

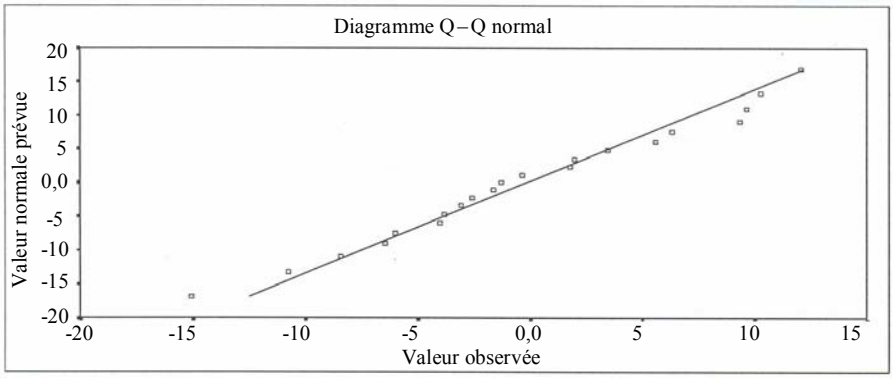
Pour construire les estimations EB, nous avons estimé d_i^2 en utilisant la méthode delta, puis t^2 suivant la méthode de Prasad et Rao (1990) en utilisant un programme SAS/IML (ce programme peut être obtenu auprès des auteurs). Ce programme nécessite une estimation initiale de t^2 que nous avons calculée par la méthode d'estimation du moment et avons obtenu $t^2 = 0{,}3117194$. Pour résoudre les équations (11), nous avons utilisé la matrice de plan d'expérience de dimensions 21×2 suivante, dont les première et deuxième colonnes contiennent des valeurs 1 et les PEA, respectivement :

$$X = \begin{bmatrix} 1 & 133\,449 \\ 1 & 141\,124 \\ 1 & 883\,653 \\ 1 & 795\,714 \\ \vdots & \vdots \\ 1 & 522\,976 \\ 1 & 162\,892 \end{bmatrix}$$

Les valeurs estimées de t^2 et $\underline{\beta}$ sont

$$\hat{t}^2 = 0.5596389, \hat{\boldsymbol{\beta}} = \begin{pmatrix} -2.066874 \\ -1.273 \times 10^{-7} \end{pmatrix}$$

Pour tester la normalité, nous avons utilisé un diagramme quantile-quantile (Q–Q plot) normal et un test de Shapiro-Wilk pour les résidus standardisés du modèle ajusté. Les points du diagramme Q–Q sont proches d'une droite et le test ne mène pas au rejet de l'hypothèse nulle de normalité (valeur $p = 0{,}851$).



3.2 Résultats

Nous présenterons les résultats en quatre parties. En premier lieu, nous examinons le biais sous forme de l'erreur et de l'erreur absolue en prenant pour critères l'erreur moyenne (EM) et l'erreur absolue moyenne (EAM). En deuxième lieu, nous comparons les erreurs quadratiques moyennes (EQM) calculées pour les diverses méthodes. En troisième lieu, nous évaluons l'efficacité des estimateurs indirects comparativement à l'estimateur direct. Enfin, nous analysons les poids des composantes directes dans les expressions (2) et (8). Tous les résultats sont illustrés au moyen de figures appropriées, mais des renseignements détaillés sont donnés au tableau 2.

Soit S_a la taille d'échantillon attribuée à une province particulière d'après l'échantillon national et S_r, la taille d'échantillon requise déterminée individuellement pour la province. Autrement dit, s'il existe un échantillon de taille S_r pour la province, il est possible de calculer une estimation directe pour cette dernière. Par conséquent, $(S_a/S_r)\times 100$ indique dans quelle mesure la taille d'échantillon disponible (S_a) est appropriée pour une province donnée. Cette mesure est utilisée sur l'axe horizontal de tous les diagrammes pour permettre la comparaison des effets de taille d'échantillon.

L'estimateur synthétique est celui dont l'EAM est la plus grande, celle-ci excédant même celle de l'estimateur direct (voir figure 1). Inversement, les EAM des estimateurs composite et EB sont les plus faibles et fort semblables. Si nous choisissons l'erreur moyenne (EM) comme critère, nous constatons que tous les estimateurs surestiment légèrement la valeur réelle. L'estimateur direct est celui dont l'EM est la plus faible, parce qu'il est sans biais. Les EM des estimateurs composite et EB sont proches, et celle de l'estimateur synthétique est la plus élevée.

Pour les estimateurs direct, composite et EB, l'erreur absolue est inférieure à 0,02 pour toutes les provinces pour lesquelles $S_a/S_r \geq 10\,\%$. Les erreurs absolues les plus élevées sont celles enregistrées pour la province d'Ilam et celle de Kohkiluyeh-o-Boyer Ahmad, qui ont les populations les plus faibles et un ratio S_a/S_r très petit. Les diagrammes obtenus pour ces trois estimateurs ont une allure relativement semblable. Il n'en est pas de même de l'estimateur synthétique, parce que les données de l'échantillon « national » sont utilisées uniquement pour produire les estimations synthétiques d'après les estimations directes par post-strate, de sorte que la taille de l'échantillon « national » (et non le ratio S_a/S_r) influe sur l'estimation synthétique calculée pour une province par la voie de la PEA dans la cellule. Autrement dit, si une post-strate ne possède pas « suffisamment » de données provenant de l'échantillon national pour produire des estimations directes acceptables et qu'une province fournit une grande PEA pour le calcul de l'estimation directe pour la post-strate, l'estimation synthétique pour la province est médiocre. Nous observons cette situation pour les provinces de Sistan-o-Balouchestan, de Bushehr, de Téhéran et de Lorestan, à cause des estimations directes de mauvaise qualité pour les deux premières post-strates (les groupes des 10 à 15 ans et des 16 à 20 ans) et de la forte population de jeunes dans ces provinces.

Tableau 2
Caractéristiques des provinces et des estimateurs

Province	PEA	S_a	S_r	S_a/S_r	ER^C	ER^{EB}	ER^S	EA^C	EA^{EB}	EA^S	EA^D	EQM^C	EQM^{EB}	EQM^S	EQM^D
Bushehr	133 449	146	4 550	3,2%	0,96	1,17	25,57	0,03300	0,01687	0,06501	0,02644	**0,0003030**	0,000368	0,0080483	0,0003148
Chaharmahal et Bakhtiyari*	141 124	203	4 063	5,0%	0,87	0,95	6,52	0,02136	0,02135	0,03644	0,02031	**0,0003813**	0,000417	0,0028670	0,0004397
Esfahan	883 653	1032	5 850	17,6%	0,90	1,00	9,56	0,01268	0,01421	0,00990	0,01504	**0,0000533**	0,000059	0,0005631	0,0000589
Fars	795 714	925	6 175	15,0%	0,91	0,99	9,69	0,00610	0,00886	0,02235	0,00904	**0,0000836**	0,000091	0,0008931	0,0000922
Gilan*	734 196	683	5 364	12,7%	1,04	0,97	17,25	0,00484	0,00460	0,01107	0,00393	0,0001728	**0,000162**	0,0028670	0,0001662
Hamedan	387 517	439	4 550	9,6%	0,77	1,00	3,36	0,01294	0,01701	0,00675	0,01880	**0,0001155**	0,000150	0,0005030	0,0001498
Hormozgan*	168 268	198	4 063	4,9%	0,84	0,93	5,12	0,01984	0,01734	0,02821	0,01862	**0,0004731**	0,000519	0,0028670	0,0005600
Ilam	84 210	111	4 063	2,7%	0,83	0,87	4,94	0,04901	0,05201	0,03395	0,06579	**0,0013919**	0,001450	0,0082747	0,0016734
Kerman*	312 768	450	5 200	8,7%	0,96	0,97	12,00	0,03615	0,03672	0,02864	0,03724	**0,0002283**	0,000231	0,0028670	0,0002389
Kermanshah	357 096	436	3 575	12,2%	0,75	0,96	3,07	0,00265	0,00928	0,02641	0,01210	**0,0002747**	0,000349	0,0011190	0,0003640
Khorasan	1 410 863	1 587	8 125	19,5%	0,70	0,99	2,36	0,00515	0,00193	0,01353	0,00160	**0,0000298**	0,000042	0,0000999	0,0000424
Khuzestan*	609 044	786	4 225	18,6%	1,03	0,97	16,83	0,01034	0,01247	0,00308	0,01140	0,0001760	**0,000166**	0,0028670	0,0001704
Kohkiluyeh-o-Boyer Ahmad	90 655	105	3 575	2,9%	0,83	0,86	4,83	0,05486	0,05932	0,02630	0,07165	**0,0013629**	0,001408	0,0079493	0,0016449
Kurdistan*	276 575	341	5 200	6,6%	0,91	0,95	9,22	0,03105	0,02814	0,03641	0,03027	**0,0002833**	0,000297	0,0028670	0,0003111
Lorestan	310 918	341	3 575	9,5%	0,86	0,95	6,22	0,00943	0,01383	0,04101	0,01754	**0,0004090**	0,000451	0,0029534	0,0004747
Mazandaran*	917 259	1 043	6 013	17,3%	1,30	0,98	33,57	0,00199	0,00188	0,00310	0,00183	0,0001112	**0,000084**	0,0028670	0,0000854
Semnan	110 166	121	4 713	2,6%	0,34	1,08	0,51	0,02776	0,01929	0,03661	0,01042	**0,0001534**	0,000491	0,0002317	0,0004542
Sistan-o-Balouchestan	272 752	318	4 875	6,5%	0,96	0,97	26,53	0,00431	0,00228	0,08606	0,00123	**0,0002519**	0,000254	0,0069347	0,0002614
Téhéran	2 343 290	2 913	8 125	35,9%	0,99	1,00	83,08	0,00605	0,00573	0,04767	0,00555	**0,0000209**	0,000021	0,0017530	0,0000211
Azerbaijan-e-gharbi (de l'ouest)	522 976	654	6 500	10,1%	0,46	0,98	0,85	0,00505	0,01247	0,00182	0,01309	**0,0000552**	0,000118	0,0001024	0,0001199
Yazd	162 892	207	5 038	4,1%	0,82	1,36	4,52	0,01414	0,00968	0,01008	0,01950	**0,0001299**	0,000215	0,0007164	0,0001586

* Dénote les provinces pour lesquelles l'expression (3) produit des estimations négatives de l'EQM.
PEA : Population économiquement active
S_a : Taille d'échantillon attribuée
S_r : Taille d'échantillon requise
ER : Efficacité relative
EA : Erreur absolue
EQM : Erreur quadratique moyenne (la valeur la plus faible est indiquée en caractères gras pour chaque province)
C, EB, S et D représentent les estimateurs composite, empirique bayésien, synthétique et direct, respectivement.

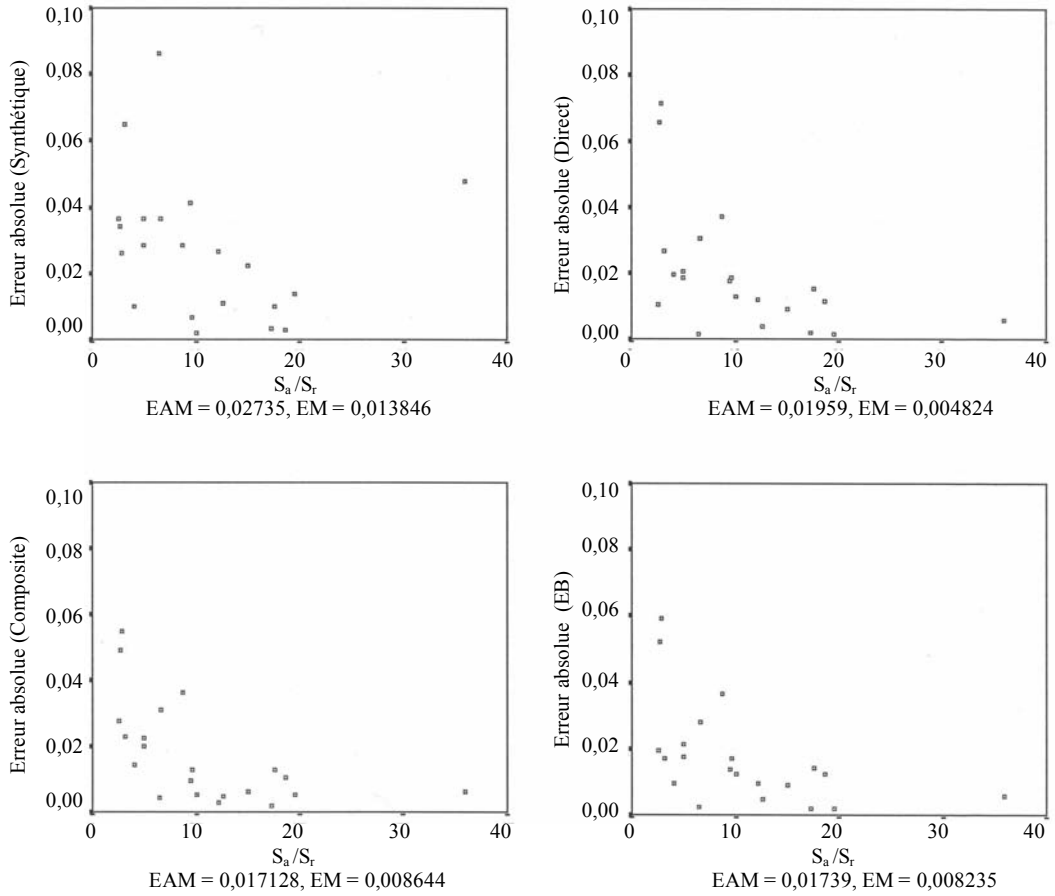


Figure 1. Erreurs absolues des estimation en fonctions du ratio S_a/S_r.

La valeur la plus faible de l'EQM est toujours donnée par l'estimateur composite ou l'estimateur EB (voir la figure 2). Cependant, l'EQM de l'estimateur composite est souvent plus faible que celle de l'estimateur EB. L'EQM de l'estimateur synthétique est systématiquement plus élevée que celle des autres estimateurs, même l'estimateur direct.

À mesure que le ratio S_a/S_r augmente, l'EQM diminue pour les estimateurs direct, composite et EB (voir la tendance décroissante à la figure 2). Cet effet est très important pour Téhéran (S_a/S_r = 36 %). De nouveau, les provinces d'Ilam et de Kohkiluyeh-o-Boyer Ahmad, qui toutes deux ont une faible population et un ratio S_a/S_r très petit, font exception pour les trois estimateurs. Le nuage de points de la figure 2 pour l'estimateur synthétique pourrait être trompeur, parce que nous avons utilisé l'EQMM au lieu de l'EQM pour sept provinces. Cependant, les valeurs pour les quatre provinces mentionnées précédemment (Sistan et Balouchestan, Bushehr, Téhéran et Lorestan) ne concordent pas non plus avec ce nuage de points. En règle générale, pour tout estimateur considéré ici, la relation entre l'EQM et le ratio S_a/S_r est d'autant plus forte que la dépendance à l'égard des estimations directes provinciales est grande.

Pour toutes les provinces, l'efficacité relative (ER) des trois estimateurs indirects comparativement à l'estimateur direct est souvent inférieure ou égale à l'unité pour les estimateurs composite et EB, et supérieure à l'unité pour l'estimateur synthétique. L'efficacité relative des estimations composites est bonne pour certaines provinces : Semnan (0,34), Azerbaijan de l'Ouest (0,46), Khorasan (0,70), Kermanshah (0,75) et Hamedan (0,77). Les moyennes des ER ($\overline{ER}^S = 13{,}6$, $\overline{ER}^C = 0{,}8595$ et $\overline{ER}^{EB} = 0{,}9951$) indiquent que l'estimateur composite est le plus efficace des trois estimateurs indirects. En outre, à la figure 3, à mesure que S_a/S_r augmente, ER^{EB} tend vers un. Comme la figure 2, la figure 3 pourrait être trompeuse pour l'estimateur synthétique.

Il est systématiquement accordé plus de poids à la composante directe, g_i, de l'estimateur donné par l'expression (8) qu'à la composante indirecte. Il en est ainsi pour l'estimateur composite, excepté pour les provinces de

Semnan et d'Azerbaijan de l'Ouest. Pour l'estimateur composite, Rao (2003a, page 58) déclare que « le poids optimal W_i^{opt} s'approche de zéro ou de un lorsque l'EQM de l'un des estimateurs qui le compose est beaucoup plus grande que celle de l'autre, c'est-à-dire quand la valeur de $f_i = EQM(\hat{P}_i^C) / EQM(\hat{P}_i^S)$ est grande ou faible. Dans ces conditions, l'estimateur dont l'EQM est la plus grande ajoute peu d'information et il est donc préférable d'utiliser la composante ayant l'EQM la plus petite. » Ce commentaire est illustré clairement pour les provinces de Bushehr ($W = 0{,}962355$, $ER^S = 25{,}27$), de Sistan et Balouchestan ($W = 0{,}963670$, $ER^S = 26{,}53$) et de Téhéran ($W = 0{,}988083$, $ER^S = 83{,}08$), parce que les estimations directes pour ces provinces ont une EQM plus faible que les estimations synthétiques. La figure 4 illustre clairement une relation ascendante entre le poids et le ratio S_a/S_r pour l'estimateur EB. Pour l'estimateur composite, le poids le plus faible et le poids le plus élevé correspondent aux provinces ayant le ratio S_a/S_r le plus faible et le ratio S_a/S_r le plus élevé, respectivement.

En général, l'estimateur synthétique produit des résultats médiocres, si l'on choisit pour critères l'EAM, l'EM, l'EQM et l'ER, même si les estimations synthétiques obtenues pour certaines provinces sont, individuellement, plus proches des valeurs réelles que les autres estimations. Toutefois, les estimations synthétiques ont été calculées dans les conditions les plus défavorables. Les valeurs des PEA appliquées pour les construire sont fondées sur le Recensement de 1986 (antérieur de dix ans à l'année pour laquelle les estimations sont produites). En outre, les estimations directes obtenues pour les deux premières post-strates sont assez différentes de celles obtenues pour les autres, ce qui produit de mauvaises estimations synthétiques.

Pour résoudre le premier problème, il faudrait établir des dossiers administratifs; pour le deuxième, l'estimation au niveau des post-strates devrait être prévue lors de l'élaboration du plan de sondage. Si l'on tient compte non seulement de l'estimation par post-strate, mais aussi de la classification des provinces lors de l'élaboration du plan de sondage, on pourra s'attendre à obtenir de bonnes estimations directes pour les post-strates. Par conséquent, on pourra aussi s'attendre à obtenir de bonnes estimations synthétiques pour les provinces. La classification des provinces peut accroître l'homogénéité grâce au regroupement des provinces semblables par classe et à l'utilisation des données d'échantillon provenant des provinces d'une classe donnée uniquement pour produire les estimations directes par post-strate en vue de construire les estimations synthétiques pour ces provinces.

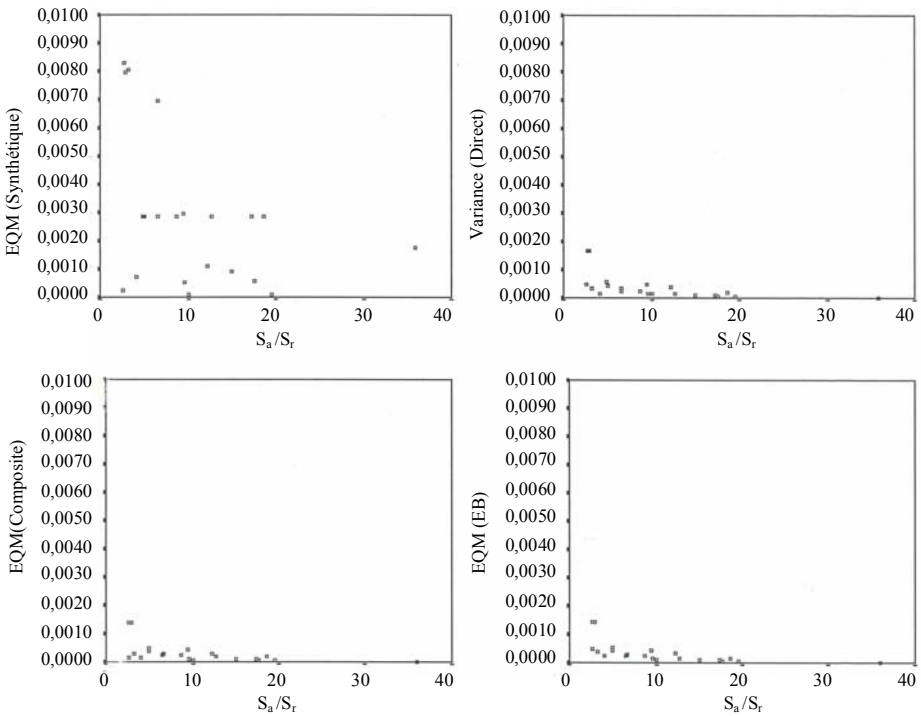


Figure 2. EQM des estimations en fonction du ratio S_a/S_r.

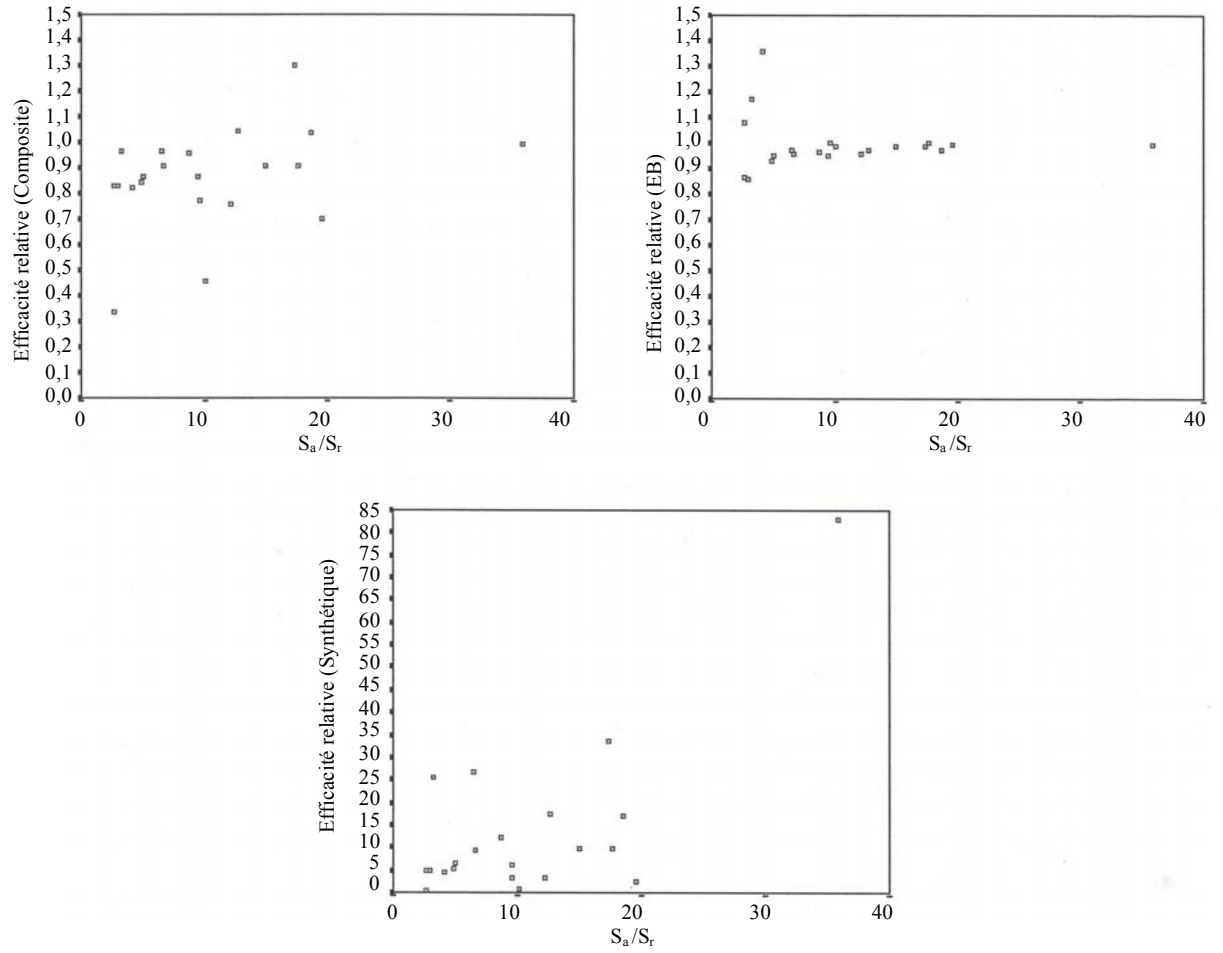


Figure 3. Efficacité relative (EQM estimée de l'estimateur indirect/variance estimée de l'estimateur direct) en fonction du ratio S_a/S_r (une échelle différente a été utilisée sur l'axe vertical pour le diagramme de l'estimateur synthétique pour le rendre plus lisible.

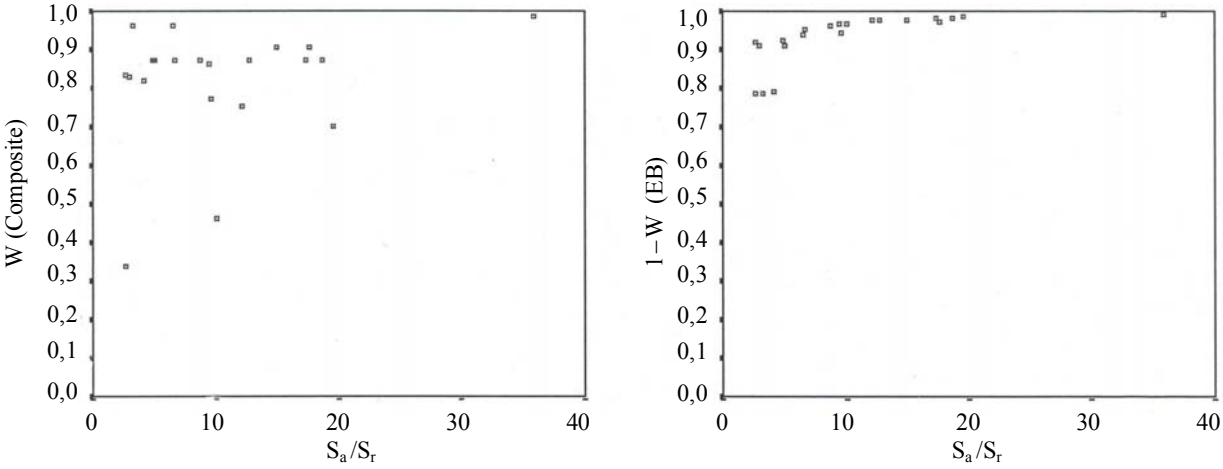


Figure 4. Poids des composantes directes des estimations composite et EB en fonction du ration S_a/S_r.

Les estimateurs composite et EB donnent habituellement de bons résultats quand le ratio S_a/S_r est égal ou supérieur à 10 % pour une province, parce que les composantes directes des estimateurs (2) et (8) sont relativement stables et reçoivent un poids plus important, particulièrement dans le cas de l'estimateur EB. Les provinces de Téhéran, de Khorasan, de Khuzestan et d'Esfahan sont de ce type, tandis que celles de Bushehr, d'Ilam, de Kohkilueyh et Buyer Ahmad et de Semnan ne le sont pas.

4. Conclusion

Dans les pays en voie de développement tels que l'Iran, il est fréquent que des dossiers administratifs ne soient disponibles ni pour les petites ni pour les grandes régions. Les enquêtes peuvent produire des estimations satisfaisantes pour les grandes régions, mais non pour les petites. Les recensements périodiques ne permettent pas de fournir toutes les données nécessaires à l'établissement de politiques efficaces et à une bonne planification. Ces limites donnent lieu à des lacunes dans les statistiques officielles. Par conséquent, les activités de planification statistique du Centre statistique d'Iran (SCI) visent à combler ces lacunes en utilisant de nouvelles méthodes et stratégies. Le présent article propose une stratégie rentable pour surmonter certaines de ces limites.

Les résultats de l'étude confirment l'idée qu'un plan d'échantillonnage de portée nationale peut remplacer des plans d'échantillonnage provinciaux distincts si l'on applique les méthodes d'estimation pour petits domaines appropriées. L'échantillon national considéré comprend près de 13 000 personnes, tandis que les 21 échantillons provinciaux distincts englobent, en tout, près de 100 000 personnes. Le tirage d'échantillons provinciaux est la méthode utilisée à l'heure actuelle par le CSI pour produire des estimations provinciales. L'utilisation d'un plan de sondage de portée nationale réduirait les coûts de plus de 80 %. En outre, il convient de souligner ce qui suit :

1. Bien que certaines méthodes d'estimation pour petits domaines ne s'appuient pas sur des données d'échantillon existantes provenant de tous les petits domaines (ou petites régions), la stratégie destinée à produire des estimations provinciales sera plus appropriée si les petits domaines d'intérêt sont définis a priori. L'échantillon national peut alors être réparti entre ces petits domaines afin de produire des estimations directes fondées sur le plan de sondage. Il est important d'ajuster le plan de sondage de façon à tenir compte des méthodes d'estimation pour petits domaines avant que la collecte de données ne débute. Comme le font remarquer Singh, Gambino et Mantel (1994, page 3),

   > « On devrait prendre conscience de la question des petites régions dès le début de la conception des plans de sondage pour les grandes enquêtes. Les plans d'échantillonnage devraient être conçus de manière que l'on puisse obtenir des données régionales fiables à l'aide d'estimateurs de plan ou de modèle. »

   Par conséquent, le CSI doit remanier les plans de sondage afin qu'ils reflètent les besoins des petites régions.
2. Les estimateurs utilisés pour l'estimation pour petits domaines donnent habituellement de bons résultats à mesure qu'augmente la taille d'échantillon. Pour améliorer les estimations provinciales, la taille de l'échantillon national peut être accrue de façon à obtenir des échantillons de plus grande taille pour chaque province. En outre, les provinces ayant des caractéristiques semblables, comme le taux de chômage, les variables sociodémographiques, et ainsi de suite, peuvent être regroupées. Des échantillons de taille distincte seraient alors déterminés pour chaque groupe.
3. L'ajout de variables supplémentaires appropriées, qui sont corrélées à la variable d'intérêt, joue un rôle essentiel dans l'amélioration des estimateurs.
   - Une seule variable (âge) a été utilisée dans l'estimateur synthétique pour diviser l'échantillon, mais on peut choisir une autre variable ou une combinaison de variables à cet effet. Les post-strates de l'estimateur synthétique devraient être formées en fonction de variables qui réduisent la variation dans chaque post-strate. Ces variables peuvent influer indirectement sur l'estimateur composite également.
   - Le modèle EB peut être amélioré en y ajoutant de meilleurs renseignements supplémentaires. Par conséquent, il est important de faire l'essai de diverses variables supplémentaires afin de découvrir le meilleur modèle. Dans les présents travaux, nous n'avons utilisé que la population économiquement active (PEA) comme variable indépendante dans le modèle, mais d'autres variables pourraient produire de meilleures estimations.
4. L'estimateur composite produit de relativement meilleurs résultats que les estimateurs synthétique et EB. Cependant, notre étude vise uniquement à donner une première idée de l'utilité des méthodes d'estimation pour petits domaines. D'autres travaux seront nécessaires en vue de mettre au point une méthodologie d'estimation pour petits domaines

générique pour l'Iran. En outre, les méthodes d'estimation pour petits domaines devraient être appliquées non seulement à l'estimation des taux de chômage, mais aussi à celle d'autres paramètres, et les estimations qu'elles produisent devraient être comparées à celles obtenues au moyen de plans d'échantillonnage distincts.

Remerciements

Les travaux de recherche décrits dans l'article ont été financés en partie par le Centre de recherche statistique de l'Iran. Les auteurs remercient les examinateurs et le rédacteur adjoint de leurs nombreux commentaires utiles et constructifs. Mes remerciements sincères vont à Jim Lepkowski pour son aide amicale. Les opinions exprimées sont celles des auteurs et ne reflètent pas forcément celles du Centre statistique d'Iran

Bibliographie

Chand, N., et Alexander, C.H. (1995). Indirect estimation of rates and rates for small areas with continuous measurement. Dans *Proceeding of the Section on Survey Research Methods*, American Statistical Association, 549-554.

Copas, J.B. (1972). Empirical Bayes methods and the repeated use of a standard. *Biometrika*, 59, 349-360.

Cressie, N. (1989). Empirical bayes estimation of undercount in the decennial census. *Journal of the American Statistical Association*, 84, 1033-1044.

Ghosh, M., Natarajan, K., Stroud, T.W.F. et Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93, 273-282.

Ghosh, M., et Rao, J.N.K. (1994). Small area estimation: An appraisal (avec discussion). *Statistical Science*, 9, 65-93.

Ghosh, M., et Meeden, G. (1997*). Bayesian Methods for Finite Population Sampling.* Chapman & Hall, London.

Gonzalez, M.F., et Hoza, C. (1978). Small area estimation with application to unemployment and housing estimation. *Journal of the American Statistical Association*, 73, 7-15.

Gonzalez, M.F., et Waksberg, J. (1973). Estimation of the errors of synthetic estimates. Article présenté à la première réunion de International Association of Survey Statistician, Vienne, Autriche, 18-25 août.

Levy, P.S. (1971). The use of mortality data in evaluating synthetic estimates. Dans *Proceedings of the American Statistical Association*, *Social Statistics Section*, 328-331.

Marker, D.A. (1995). *Small area estimation*: *A Bayesian perspective.* Thèse non publiée, University of Michigan, Ann Arbor, Michigan.

Marker, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.

Pfeffermann, D. (2002). Small area estimation-New developments and directions. *Revue Internationale de Statistique*, 70, 125-143.

Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of the mean square error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Purcell, N.J., et Kish, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.

Purcell, N.J., et Kish, L. (1980). Postcensal estimates for local areas (or domains). *Revue Internationale de Statistique*, 48, 3-18.

Rao, J.N.K. (2003a). *Small Area Estimation.* New York: John Wiley & Sons, Inc.

Rao, J.N.K. (2003b). Some new developments in small area estimation. *Journal of the Iranian Statistical Society*, 2, 2, 145-169.

Schaible, W.L. (1978). Choosing weight for composite estimators for small area statistics. Dans *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 741-746.

Schaible, W.L. (1995). Éd. *Lecture Notes in Statistics*: *Indirect Estimators in U.S. Federal Programs*, New York: Springer.

Singh, M.P., Gambino, J. et Mantel, H.J. (1994). Les petites régions : problèmes et solutions. *Techniques d'enquête*, 20, 3-23.

Thompsen, I., et Holmoy A.M.K. (1998). Combining data from surveys and administrative record system: The Norwegian experience. *Revue Internationale de Statistique*, 66, 201-221.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, juin 2006
Vol. 32, N° 1, pp. 127-133
Statistique Canada, N° 12-001-XIF au catalogue

Effets de plan pour les échantillons à plans de sondage multiples

Siegfried Gabler, Sabine Häder et Peter Lynn [1]

Résumé

Dans certaines situations, le plan de sondage d'une enquête est assez complexe et comporte des plans fondamentalement différents pour divers domaines. L'effet de plan des estimations fondées sur l'échantillon total est une somme pondérée des effets de plan selon le domaine. Nous calculons les pondérations sous un modèle approprié et illustrons leur utilisation au moyen de données provenant de l'Enquête sociale européenne (European Social Survey ou ESS).

Mots clés : Stratification; mise en grappes; modèle des composantes de la variance; coefficient de corrélation intraclasse; probabilités de sélection.

1. Introduction

En recherche par sondage, l'application de plans de sondage complexes est fréquente. Ces plans possèdent des caractéristiques, telles la stratification, la mise en grappes et (ou) l'utilisation de probabilités d'inclusion inégales, qui donnent lieu à des « effets de plan ». L'effet de plan est une mesure qui représente l'effet du plan de sondage sur la variance d'une estimation. Fondé sur le plan de sondage, il est défini comme suit (voir Lohr 1999, page 239) :

$$deff(plan, statistique) = \frac{\mathrm{V}(\text{estimation d'après le plan d'échantillonnage})}{\mathrm{V}\left(\begin{array}{c}\text{estimation d'après un eas avec le même} \\ \text{nombre d'unités d'observation}\end{array}\right)}$$

où eas indique un échantillon aléatoire simple. Le recours à la mise en grappes et (ou) à des probabilités d'inclusion inégales produit habituellement des effets de plan dont la valeur est supérieure à 1,0; autrement dit, la variance d'une estimation est plus grande que celle de l'estimation établie d'après un échantillon aléatoire simple contenant le même nombre d'observations. La prise en compte des effets de plan est très importante lorsqu'on décide d'avance de la taille de l'échantillon d'une enquête. Ainsi, si l'on prévoit mener une enquête comparative entre plusieurs pays, il est très utile de disposer d'information sur les effets de plan pour ces pays. Il est alors possible de choisir les tailles d'échantillon nettes de façon que la précision des estimations soit approximativement uniforme. Pour cela, la taille d'échantillon qui serait nécessaire sous eas (taille effective d'échantillon) pour obtenir un degré donné de précision doit être multipliée par l'effet de plan prévu.

L'Enquête sociale européenne (ESS, voir www.europeansocialsurvey.com) est un programme d'enquête dans lequel les effets de plan sont pris en compte pour le calcul des tailles nettes d'échantillon, en cherchant à obtenir la même taille effective d'échantillon pour chaque pays (n_{eff} = 1 500). Des 22 pays qui ont participé au premier cycle de l'ESS, trois seulement, le Danemark, la Finlande et la Suède, ont utilisé un plan de sondage avec probabilités de sélection égales, sans mise en grappes (eas). Pour tous les autres, il a fallu prédire l'effet de plan avant l'étude. On peut utiliser, pour cela, une approche fondée sur un modèle (voir Gabler, Häder et Lahiri 1999) qui fait la distinction entre l'effet de plan dû à un échantillonnage avec probabilités d'inclusion inégales (1er terme) et l'effet de plan dû à la mise en grappes (2e terme) :

$$deff = m\frac{\sum_{i=1}^{I} m_i w_i^2}{\left(\sum_{i=1}^{I} m_i w_i\right)^2} \times [1 + (b^* - 1)\rho] = deff_p \times deff_c \qquad (1)$$

où m_i représente les répondants dans la i^{e} classe de probabilités de sélection, chacun recevant un poids de w_i, ρ est le coefficient de corrélation intragrappe et

$$b^* = \frac{\sum_{c=1}^{C}\left(\sum_{j=1}^{b_c} w_{cj}\right)^2}{\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}^2}$$

où b_c est le nombre d'observations dans la grappe $c\,(c = 1, \ldots, C)$ et w_{cj} est le poids de sondage de l'élément j dans la grappe c. (Il s'agit évidemment d'une simplification reposant sur l'hypothèse qu'il n'existe aucune association entre y et w_i, ou entre w_i et b^*, et ne tenant compte d'aucun effet de stratification, qui aura tendance à être avantageuse et modeste. Voir Lynn, Gabler, Häder et Laaksonen (2007, à paraître), ainsi que Park et Lee (2004) pour une discussion de la sensibilité des prédictions de *deff* à

1. Siegfried Gabler et Sabine Häder, Zentrum für Umfragen, Methoden und Analysen (ZUMA), Postfach 12 21 55, 68072 Mannheim, Allemagne. Courriel : gabler@zuma-mannheim.de; Peter Lynn, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, Royaume-Uni. Courriel : plynn@essex.ac.uk.

ces hypothèses; voir Lynn et Gabler (2005) pour une discussion de divers moyens de prédire $deff_c$).

Dans certains pays, les plans de sondage utilisés étaient encore plus compliqués, comprenant des plans fondamentalement différends dans chacun des deux domaines indépendants. Au Royaume-Uni, par exemple, il s'agissait du mélange d'un plan par grappes avec probabilités d'inclusion inégales (en Grande-Bretagne) et d'un échantillon sans mise en grappes (en Irlande du Nord). En Pologne, des échantillons aléatoires simples ont été sélectionnés dans un domaine (villes grandes et moyennes), tandis qu'un plan à deux degrés avec mise en grappes a été appliqué au deuxième domaine (toutes les autres régions). En Allemagne, un échantillon par grappes avec probabilités de sélection égales a été sélectionné dans chaque domaine (Allemagne de l'Ouest, y compris Berlin Ouest; Allemagne de l'Est), mais les fractions d'échantillonnage n'étaient pas les mêmes dans les deux domaines.

La question de la prédiction des effets de plan s'est donc posée pour ces échantillons à plan de sondage double. Comme nous le montrons plus loin, il ne s'agit pas simplement d'une combinaison convexe des effets de plan pour les divers domaines, à part dans des cas spéciaux. Nous présentons une solution générale pour les échantillons à plans de sondage multiples à la section 2, ainsi que des exemples d'application de cette solution afin de prédire les effets de plan avant le travail sur le terrain (section 3) et après le travail sur le terrain (section 4). À la section 5, nous concluons par une discussion.

2. Effets de plan pour les échantillons à plans de sondage multiples

Soit $\{C_1, \ldots, C_K\}$ une partition des grappes en K domaines. Alors, $C\bar{b} = \sum_{c=1}^{C} b_c = \sum_{k=1}^{K} \sum_{c \in C_k} b_c = \sum_{k=1}^{K} m_k = m$, où $m_k = \sum_{c \in C_k} b_c$ est le nombre d'observations dans le k^{e} domaine de grappes. Soit y_{cj} l'observation pour l'unité d'échantillonnage j dans la grappe $c (c = 1, \ldots, C; j = 1, \ldots, b_c)$. L'estimateur habituel fondé sur le plan de sondage de la moyenne de population est

$$\bar{y}_w = \frac{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj} y_{cj}}{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj}} = \sum_{k=1}^{K} \frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}}{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj}} \bar{y}_w^{(k)}$$

où

$$\bar{y}_w^{(k)} = \frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj} y_{cj}}{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}}.$$

Nous postulons le modèle M1 suivant :

$$\left.\begin{aligned} E(y_{cj}) &= \mu \\ \text{Var}(y_{cj}) &= \sigma^2 \end{aligned}\right\} \quad \text{pour} \quad c = 1, \ldots, C; j = 1, \ldots, b_c$$
$$\text{Cov}(y_{cj}, y_{cj'}) = \begin{cases} \rho_k \sigma^2 \text{ si } c = c' \in C_k; j \neq j' \\ 0 \quad \text{autrement} \end{cases} \quad k = 1, \ldots, K. \tag{2}$$

Le modèle M1 convient pour tenir compte de l'effet de grappe avec divers types de grappes et généralise une approche antérieure (voir, par exemple, Gabler et coll. 1999). Des modèles plus généraux sont décrits dans Rao et Kleffe (1988, page 62). Nous définissons l'effet de plan (par rapport au modèle) comme étant $deff = \text{Var}_{M1}(\bar{y}_w) / \text{Var}_{M2}(\bar{y})$, où $\text{Var}_{M1}(\bar{y}_w)$ est la variance de $\bar{y}_w$ sous le modèle M1 et $\text{Var}_{M2}(\bar{y})$ est la variance de la moyenne globale d'échantillon $\bar{y}$, définie comme étant $\sum_{c=1}^{C} \sum_{j=1}^{b_c} y_{cj} / m$, calculée sous le modèle M2 suivant :

$$\left.\begin{aligned} E(y_{cj}) &= \mu \\ \text{Var}(y_{cj}) &= \sigma^2 \end{aligned}\right\} \quad \text{pour} \quad c = 1, \ldots, C; j = 1, \ldots, b_c$$
$$\text{Cov}(y_{cj}, y_{c'j'}) = 0 \quad \text{pour tous les} \quad (c, j) \neq (c', j'). \tag{3}$$

Soulignons que le modèle M2 est approprié sous échantillonnage aléatoire simple et donne l'expression habituelle $\text{Var}_{M2}(\bar{y}) = \sigma^2 / m$.

De façon fort semblable à Gabler et coll. (1999), nous notons que

$$\text{Var}_{M1}\left(\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj} y_{cj} \right) = \sigma^2 \sum_{k=1}^{K} \sum_{c \in C_k} \left\{ \sum_{j=1}^{b_c} w_{cj}^2 + \rho_k \sum_{j \neq j'}^{b_c} w_{cj} w_{cj'} \right\}. \tag{4}$$

Donc

$$deff = \sum_{k=1}^{K} \left(\frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}}{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj}} \right)^2 \frac{m}{m_k} deff_k \tag{5}$$

où

$$deff_k = m_k \frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}^2}{\left(\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj} \right)^2} \times [1 + (b_k^* - 1)\rho_k] = deff_{pk} \times deff_{ck},$$

et

$$b_k^* = \frac{\sum_{c \in C_k} \left(\sum_{j=1}^{b_c} w_{cj} \right)^2}{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}^2}.$$

On peut voir que *deff* n'est pas une combinaison convexe des effets de plan spécifiques $\{deff_k\}$, sauf dans des cas particuliers. Nous considérons ici quatre scénarios raisonnables, chacun représentant une simplification du cas général. La combinaison ne devient convexe que dans deux de ces scénarios (1 et 4) :

Scénario 1 : *Même pondération pour toutes les unités*

Si $w_{cj} = 1$ pour tous c, j, alors l'expression (5) se simplifie comme suit :

$$deff = \sum_{k=1}^{K} \frac{m_k}{m} deff_k. \qquad (6)$$

Scénario 2 : *Même pondération des unités dans chaque domaine*

Si $w_{cj} = w_k$ pour tous $c \in C_k, j$, alors l'expression (5) devient :

$$deff = \sum_{k=1}^{K} \left(\frac{m_k w_k}{\sum_{k=1}^{K} m_k w_k} \right)^2 \frac{m}{m_k} deff_k. \qquad (7)$$

Scénario 3 : *Taille d'échantillon pondéré proportionnelle à la taille de la population du domaine*

Si

$$\frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}}{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj}} = \frac{N_k}{N},$$

où N_k est la taille de population dans le domaine k; $N = \sum_{k=1}^{K} N_k$, alors l'expression (5) devient :

$$deff = \sum_{k=1}^{K} \left(\frac{N_k}{N} \right)^2 \frac{m}{m_k} deff_k. \qquad (8)$$

Scénario 4 : *Taille d'échantillon non pondéré proportionnelle à la taille de population du domaine*

Si

$$\frac{m}{m_k} = \frac{N}{N_k},$$

alors l'expression (8) devient :

$$deff = \sum_{k=1}^{K} \frac{N_k}{N} deff_k. \qquad (9)$$

3. Application à la prédiction de *Deff*

Lors du premier cycle de l'ESS, le plan de sondage était une combinaison de deux plans différents pour 5 des 22 pays, à savoir le Royaume-Uni, la Pologne, la Belgique, la Norvège et l'Allemagne. Nous pouvons appliquer la formule générale (5) des effets de plan pour échantillon à plans multiples à chacun de ces cas, où $K = 2$. Pour certains d'entre eux, nous pouvons utiliser indifféremment l'une des expressions simplifiées (6) à (9). Ici, nous illustrons comment la formule serait utilisée pour prédire les effets de plan avant le travail sur le terrain en vue d'établir la taille d'échantillon nette (répondants) requise pour atteindre une précision d'estimation préétablie. Dans chaque cas, l'approche consiste à prédire $\{deff_k\}$ en utilisant (1) pour chaque *k*, puis à utiliser (5) pour prédire *deff*. Afin de prédire $\{deff_k\}$, nous utilisons les valeurs observées de $\{w_{cj}\}$ provenant de l'échantillon de répondants du premier cycle de l'ESS pour estimer b^*, m_i et w_i. En d'autres termes, nous pourrions considérer que ces prédictions sont faites pour une future enquête basée sur le même plan de sondage (par exemple, un futur cycle de l'ESS). À titre d'exemple, nous supposons que $\rho_k = 0{,}02 \; \forall \, k$ avec un plan de sondage par grappes et $\rho_k = 0{,}00 \; \forall \, k$ avec un plan sans mise en grappes (0,02 dans l'ESS dans les cas où l'on ne disposait pas d'estimations d'après des enquêtes antérieures). Ici, nous nous concentrons sur l'application de l'équation (5). Pour une description plus détaillée des plans de sondage, voir Häder, Gabler, Laaksonen et Lynn (2003). Nous choisissons comme exemples trois des pays participants à l'ESS, la Pologne, le Royaume-Uni et l'Allemagne, car ils ont utilisé des plans de sondage multiples dont les différences entre domaines ne sont pas les mêmes. Les plans de sondage utilisés par la Norvège et la Belgique étaient semblables à celui de la Pologne, avec probabilités d'inclusion égales pour toutes les unités, mais mise en grappes dans un domaine et non dans l'autre.

3.1 Pologne

En Pologne, le premier domaine couvrait la population des villes de 100 000 habitants et plus. Dans ce domaine, des personnes ont été sélectionnées par EAS d'après le registre de population (base de données PESEL) dans chaque région, avec application d'une fraction d'échantillonnage légèrement différente selon la région afin de refléter les différences attendues de taux de réponse. Ce domaine comprenait 42 villes qui représentaient environ 31 % de la population cible.

Le deuxième domaine correspondait au reste de la population, c'est-à-dire les personnes vivant dans les villes de 99 999 habitants et moins et dans les régions rurales. Cette partie de l'échantillon a été stratifiée et mise en grappes (158 grappes). L'échantillonnage de cette deuxième partie a été réalisé selon un plan à deux degrés où les UPE ont été sélectionnées avec probabilité proportionnelle à la taille. La définition de l'UPE n'était pas la même pour les régions urbaines que pour les régions rurales. Pour les premières, l'UPE correspondait à une ville, tandis que pour

les secondes, il s'agissait d'un village. À la deuxième phase, une grappe de 12 répondants a été sélectionnée par eas dans chaque UPE.

Dans le premier domaine, $\rho_1 = 0$ et $deff_{c1} = 1$. La variation modérée des probabilités de sélection mène à $deff_{p1} = 1{,}005$ et, par conséquent, $deff_1 = deff_{c1} \cdot deff_{p1} = 1{,}005$. Dans le deuxième domaine, l'effet de plan dû à la mise en grappes prévu est $deff_{c2} = 1{,}18$ (d'après la prédiction que $b^* = 10{,}07$) et $deff_{p2} = 1{,}01$, ce qui donne $deff_2 = deff_{c2} \cdot deff_{p2} = 1{,}19$. La substitution de ces valeurs à $deff_k$ dans (5) donne un effet de plan prévu égal à $deff = 1{,}17$.

Le plan de sondage appliqué en Pologne ne diffère que légèrement du scénario 2 et, dans ce cas, nous voyons que l'expression plus simple, (7), donne une prédiction raisonnable si nous calculons la valeur approximative des pondérations comme suit. Le domaine 1 contient 37,3 % de l'échantillon brut et 31 % de la population cible. Donc

$$w_1 = \frac{N_1 / N}{n_1 / n} = \frac{0{,}310}{0{,}373} = 0{,}831$$

et

$$w_2 = \frac{N_2 / N}{n_2 / n} = \frac{0{,}690}{0{,}627} = 1{,}100,$$

respectivement, où n_k est la taille de l'échantillon sélectionné dans le domaine k; $\sum_{k=1}^{K} n_k$.

Maintenant, nous pouvons appliquer l'expression (7) pour calculer l'effet de plan prévu pour les estimations pour la Pologne : $deff = (0{,}194 \cdot 1{,}005) + (0{,}821 \cdot 1{,}19) = 1{,}17$.

3.2. Royaume-Uni

Au Royaume-Uni, le plan de sondage de l'ESS n'était pas le même pour la Grande-Bretagne (Angleterre, Pays de Galles, Écosse) que pour l'Irlande du Nord. En Grande-Bretagne, on a utilisé un plan stratifié à trois degrés avec probabilités de sélection inégales. À la première étape, 162 petits secteurs, appelés « secteurs de code postal » ont été sélectionnés systématiquement avec probabilité proportionnelle au nombre d'adresses dans le secteur, après stratification implicite selon la région et la densité de population. À la deuxième étape, 24 adresses ont été sélectionnées dans chaque secteur, ce qui a produit un échantillon avec probabilités égales d'adresses. À la troisième étape, une personne de 15 ans ou plus a été sélectionnée à chaque adresse échantillonnée au moyen d'une grille de Kish.

Pour l'Irlande du Nord, un échantillon aléatoire simple de 125 adresses a été tiré d'après la liste de propriétés privées de l'Agence d'évaluation foncière (Valuation and Land Agency). Puis, une personne de 15 ans ou plus a été sélectionnée à chaque adresse échantillonnée en utilisant une grille de Kish. Donc, l'échantillon du Royaume-Uni est mis en grappes dans un domaine, mais non dans l'autre. Dans les deux domaines, les probabilités de sélection sont inégales.

Pour la Grande-Bretagne, nous avons prédit $deff_{c1} = 1{,}20$ (d'après une prédiction de $b^* = 11{,}11$) et $deff_{p1} = 1{,}22$, de sorte que $deff_1 = 1{,}46$. Pour l'Irlande du Nord, nous obtenons les prédictions $deff_{c2} = 1$ (par définition) et $deff_{p2} = 1{,}27$, de sorte que $deff_{p2} = 1{,}27$. D'après l'expression (5), $deff = 0{,}978 \cdot 1{,}46 + 0{,}023 \cdot 1{,}27 = 1{,}460$. Il convient de souligner que les tailles des échantillons sélectionnés dans les deux domaines ont été choisies de façon à obtenir des tailles nettes d'échantillon à peu près proportionnelles aux tailles de population. Autrement dit, la simplification correspondant au scénario 4 est vérifiée approximativement. Si nous utilisons l'expression (9), nous obtenons $deff = N_1 / N \; deff_1 + N_2 / N \quad deff_2 = 0{,}97 \cdot 1{,}46 + 0{,}03 \cdot 1{,}27 = 1{,}457$, ce qui démontre que cette expression est une approximation raisonnable de (5) dans ce cas.

3.3. Allemagne

En Allemagne, des échantillons indépendants ont été sélectionnés dans deux domaines, c'est-à-dire l'Allemagne de l'Ouest, y compris Berlin Ouest et l'Allemagne de l'Est, y compris Berlin Est. Dans chaque domaine, on a sélectionné un échantillon par grappes avec probabilités égales, mais en utilisant une plus grande fraction d'échantillonnage pour l'Allemagne de l'Est.

À la première étape, 100 communautés (grappes) ont été sélectionnées pour l'Allemagne de l'Ouest et 50 pour l'Allemagne de l'Est avec probabilité proportionnelle à la taille de la population de la communauté (personnes de 15 ans et plus). Le nombre de communautés sélectionnées dans chaque strate a été déterminé par une méthode d'arrondissement contrôlé. Le nombre de points d'échantillonnage était de 108 à l'Ouest et de 55 à l'Est (pour certaines grandes collectivités, on a utilisé plus d'un point d'échantillonnage). À la deuxième étape, pour chaque point d'échantillonnage, on a sélectionné un nombre égal de personnes par échantillonnage aléatoire systématique, d'après les registres locaux des bureaux d'enregistrement des résidents.

Puisque le plan de sondage est autopondéré aussi bien pour l'Allemagne de l'Est que de l'Ouest, mais que la répartition est disproportionnelle, nous pouvons appliquer le scénario 2 et utiliser l'expression (7), où

$$w_1 = w_{\text{EST}} = \frac{N_{\text{EST}}}{N} \frac{n}{n_{\text{EST}}} = 0{,}567$$

et

$$w_2 = w_{\text{OUEST}} = \frac{N_{\text{OUEST}}}{N} \frac{n}{n_{\text{OUEST}}} = 1{,}257.$$

(Nous notons qu'il est courant, dans certaines enquêtes, de rééchelonner les pondérations afin que leur somme soit égale à la taille de population. Cette pratique n'aurait aucune incidence ici, car l'expression (5) ne comprend que des ratios de sommes de pondérations).

Pour chaque domaine, nous avons prédit les effets de plan $deff_{c1} = 1{,}39$ et $deff_{c2} = 1{,}35$, respectivement (d'après les prédictions que $b^* = 20{,}56$ et 18,65, respectivement), si bien qu'il découle de (7) que

$$deff = 0{,}120 \cdot 1{,}39 + 0{,}991 \cdot 1{,}35 = 1{,}51.$$

Il convient de souligner que dans ce cas, toute combinaison convexe des effets de plan de domaine produira une prédiction de *deff* comprise entre 1,35 et 1,39. Par exemple, (6) donnerait $deff = 1{,}36$. Ce résultat ne tient pas compte des différences de probabilité de sélection *entre* les domaines. Dans le cas du plan de sondage examiné ici, où la *seule* différence de plan de sondage entre les domaines est la différence de probabilité de sélection, *deff* pourrait aussi être prédit en prenant la combinaison convexe et en la multipliant par la prédiction de $deff_p$ pour le premier terme de l'expression (1), c'est-à-dire $deff = 1{,}36 \cdot 1{,}09 = 1{,}49$. Cette méthode n'est toutefois équivalente que dans le cas particulier où les $\{deff_k\}$ sont égaux, et approximativement équivalente ici, où la variation est faible.

4. Application à l'estimation de *Deff*

Nous allons maintenant illustrer l'utilisation de l'expression (5) pour estimer les effets de plan après le travail sur le terrain. Nous présentons des estimations pour cinq variables démographiques/de comportement et un ensemble de 24 mesures d'attitude provenant du premier cycle de l'Enquête sociale européenne (ESS) pour les trois mêmes pays qu'à la section 3. Aux fins de comparaison, nous présentons aussi les estimations que l'on obtiendrait en utilisant les expression plus simples (6), (8) et (9). Les résultats montrent que les estimations de *deff* diffèrent considérablement selon la variable. Cette situation, qui est prévisible, reflète la variation de l'association de y avec les grappes et les probabilités de sélection. Mais ici, nous nous intéressons principalement aux différences entre les méthodes d'estimation pour une même variable.

Dans le cas de l'Allemagne, nous voyons que les estimateurs (6) et (9), qui ne tiennent pas compte de la variation de la pondération et des taux d'échantillonnage entre les deux domaines, sous-estiment *deff* pour toutes les variables. L'estimateur (8), qui repose uniquement sur l'hypothèse que les taux de réponse sont égaux dans chaque domaine, produit des estimations fort semblables à (5). Pour la Pologne, les trois estimateurs simplifiés sous-estiment *deff*, quoique (6) pourrait produire des résultats légèrement meilleurs que les deux autres. Pour ce qui est du Royaume-Uni, nous obtenons le résultat remarquable que les quatre estimateurs produisent des estimations presque identiques pour chaque variable. L'hypothèse qui sous-tend (9) (et par conséquent également celle qui sous-tend (8)) tient pour le Royaume-Uni et, bien que les pondérations soient loin d'être égales, leur distribution est fort semblable dans chaque domaine. Il convient de souligner que (6) est vérifié sous une hypothèse plus faible que

$$\frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}}{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj}} = \frac{m_k}{m},$$

c'est-à-dire que la part des pondérations dans chaque strate est égale à la part des unités d'échantillonnage. Il est frappant que ces relations entre les estimateurs soient convergentes pour l'ensemble des variables considérées.

5. Discussion et conclusion

L'expression (5) offre un moyen approprié de combiner les effets de plan pour des domaines pour lesquels les plans de sondage sont fondamentalement différents. Elle peut être appliquée en estimant l'effet de plan *deff* de la manière habituelle pour chaque domaine, puis en les combinant d'après les données sur la pondération et l'appartenance des unités d'échantillonnage au domaine. L'utilisation de (5) pour prédire les effets de plan *deff* avant qu'une enquête soit réalisée est à peine plus difficile. Elle nécessite la prédiction de la part des pondérations dans l'échantillon de répondants dans chaque domaine, en plus d'une méthode de prédiction des *deff* propres aux divers plans de sondage.

Nous avons montré à la section 4 qui précède que l'utilisation d'autres méthodes, plus simples, de combinaison des *deff* de domaine ne produit pas toujours de bonnes estimations. Plus précisément, l'utilisation d'une combinaison convexe aura tendance à causer une sous-estimation dont l'importance dépend de l'écart par rapport aux hypothèses qui sous-tendent les expressions simplifiées. Dans notre exemple empirique, les écarts étaient modérés, mais il est facile d'imaginer des plans de sondage où les variations des probabilités de sélection moyennes ou de la distribution des poids de sondage selon le domaine sont plus importantes. Nous devrions par conséquent recommander de n'utiliser les estimateurs (6) à (9) que si les hypothèses sont vraiment vérifiées ou que les données sur le plan de sondage nécessaires pour calculer (5) ne sont pas disponibles, auquel cas l'analyste devrait au moins tenir compte arbitrairement d'une sous-estimation en se basant sur sa connaissance du plan de sondage.

Tableau 1
Estimations de *Deff* pour les moyennes sous quatre estimateurs pour trois pays

	Allemagne				Royaume-Uni				Pologne			
Estimateur :	(5)	(6)	(8)	(9)	(5)	(6)	(8)	(9)	(5)	(6)	(8)	(9)
Variables démographiques/de comportement												
N^bre de personnes dans le ménage	1,87	1,85	1,87	1,74	1,66	1,66	1,66	1,66	1,51	1,43	1,41	1,42
N^bre d'années d'études	3,25	2,80	3,25	2,88	2,81	2,79	2,80	2,79	1,77	1,66	1,63	1,64
Revenu net du ménage	2,46	2,15	2,46	2,19	2,82	2,80	2,80	2,80	2,16	2,00	1,95	1,98
Temps passé à regarder la TV	2,08	1,86	2,08	1,87	2,04	2,03	2,03	2,03	1,31	1,26	1,25	1,25
Temps passé à lire le journal	1,79	1,62	1,79	1,61	1,35	1,35	1,35	1,35	1,73	1,63	1,60	1,61
Mesures d'attitude												
Discrimination selon la race	1,16	1,03	1,16	1,04	1,92	1,92	1,92	1,92	1,02	1,01	1,01	1,01
Discrimination selon la religion	1,22	1,05	1,22	1,08	1,26	1,26	1,26	1,26	1,07	1,05	1,05	1,05
État de bonheur général	2,56	2,11	2,55	2,23	1,56	1,55	1,56	1,55	1,49	1,42	1,40	1,41
Confiance dans les autres	2,20	1,96	2,20	1,98	1,85	1,84	1,84	1,84	1,66	1,57	1,54	1,55
Confiance dans le Parlement européen	1,83	1,59	1,83	1,62	1,50	1,50	1,50	1,50	1,43	1,37	1,35	1,36
Confiance dans le système juridique	2,07	1,72	2,07	1,81	1,37	1,37	1,37	1,37	1,42	1,36	1,34	1,35
Confiance dans la police	1,92	1,63	1,92	1,69	1,24	1,24	1,24	1,24	1,24	1,20	1,19	1,19
Confiance dans les politiciens	1,75	1,62	1,75	1,59	1,38	1,38	1,38	1,38	1,63	1,54	1,51	1,53
Confiance dans le Parlement	1,64	1,48	1,64	1,48	1,45	1,45	1,45	1,45	1,13	1,10	1,10	1,10
Échelle gauche-droite	1,70	1,65	1,70	1,58	1,48	1,47	1,48	1,48	1,31	1,26	1,25	1,25
Satisfaction à l'égard de la vie	2,06	1,74	2,06	1,81	1,68	1,67	1,67	1,67	1,30	1,25	1,24	1,25
Satisfaction à l'égard du système d'éducation	3,03	2,89	3,03	2,79	1,37	1,37	1,37	1,37	1,40	1,34	1,32	1,33
Satisfaction à l'égard du système de santé	3,76	3,21	3,76	3,32	1,65	1,64	1,64	1,64	1,65	1,56	1,53	1,54
Attitude religieuse	1,94	1,75	1,94	1,75	1,57	1,56	1,56	1,56	1,73	1,63	1,60	1,61
Attitude à l'égard des immigrants	2,77	2,68	2,77	2,57	1,92	1,92	1,92	1,92	1,89	1,76	1,73	1,74
Appuie une loi contre la discrimination ethnique	2,82	2,85	2,82	2,66	1,73	1,72	1,72	1,72	2,57	2,36	2,29	2,33
Importance de la famille	2,17	1,99	2,17	1,97	1,19	1,19	1,19	1,19	1,21	1,17	1,17	1,17
Importance des amis	2,31	2,09	2,31	2,08	1,34	1,34	1,34	1,34	1,54	1,46	1,44	1,45
Importance du travail	2,20	2,16	2,20	2,05	1,90	1,89	1,89	1,89	1,69	1,59	1,57	1,58
Aide les personnes moins bien nanties	2,70	2,47	2,70	2,45	1,35	1,35	1,35	1,35	1,78	1,67	1,64	1,66
Respecte toujours la loi	2,43	2,21	2,43	2,20	1,53	1,52	1,52	1,52	2,11	1,96	1,91	1,93
Activisme politique	3,26	2,83	3,26	2,89	1,94	1,94	1,94	1,94	2,16	2,00	1,96	1,98
Libéralisme	2,28	2,18	2,28	2,10	1,78	1,77	1,78	1,78	1,75	1,64	1,61	1,63
Participation à des groupes	3,75	3,04	3,75	3,24	2,26	2,25	2,25	2,25	1,82	1,71	1,68	1,69

Une question importante qui dépasse le cadre du présent article est celle de savoir comment traiter la non-réponse lors de la prédiction ou de l'estimation des effets de plan pour les échantillons à plans multiples. Les expressions présentées à la section 2 se rapportent au nombre d'observations (unités d'échantillonnage répondantes) dans chaque domaine, m_k, et les calculs présentés aux sections 3 et 4 sont fondés sur les nombres prévus et réels d'observations, respectivement. Cependant, l'interprétation naturelle des différences entre les quatre scénarios de la section 2 pourrait se faire en fonction du plan de sondage, où les pondérations sont les poids de sondage. Donc, le scénario 2, par exemple, aurait trait à un plan de sondage avec *probabilités de sélection égales* dans les domaines, mais où la fraction d'échantillonnage peut varier selon le domaine. Cependant, dans la plupart des applications réelles, il y aura une non-réponse qui pourrait fort bien différer entre les domaines, ainsi que dans les domaines, situation qui est souvent reflétée par un ajustement du poids de sondage. Donc, la simplification du scénario 2 ne serait applicable que si l'ajustement pour la non-réponse était constant dans les domaines, outre la sélection avec *probabilités égales* dans les domaines.

Le scénario 3, s'il est interprété uniquement par rapport au plan de sondage, devrait être vérifié pour tout plan de sondage bien spécifié dans lequel les domaines forment des strates explicites. L'expression (8) est par conséquent équivalente à l'expression (5), en l'absence de non-réponse. En présence de non-réponse, le scénario 3 exige que les taux de réponse (pondérés par les poids de sondage) soient égaux dans chaque domaine. De même, le scénario 4 demande que le taux d'inclusion net (produit du taux de couverture, de la fraction d'échantillonnage et du taux de réponse) soit le même dans chaque domaine, tandis qu'une interprétation basée sur le plan de sondage ne tiendrait pas compte de la composante du taux de réponse.

La recherche de moyens appropriés d'intégrer l'ajustement pour la non-réponse dans l'estimation de l'effet de

plan et, en particulier, l'effet que cette correction pourrait avoir sur l'estimation dans le cas d'échantillons à plans multiples, semble être un domaine qui mérite d'être exploré lors de futures études.

Remerciements

Le troisième auteur remercie la ZUMA pour le poste de professeur invité qui lui a permis de trouver le temps et les conditions propices pour rédiger le présent article et remercie aussi de son soutien le UK Longitudinal Studies Centre de l'Université d'Essex, qui est financé par la subvention H562255004 de l'Economic and Social Research Council du Royaume-Uni.

Bibliographie

Gabler, S., Häder, 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, 119-120.

Häder, S., Gabler, S., Laaksonen, S. et Lynn, P. (2003). The sample. Chapître 2 dans *ESS 2002/2003: Rapport technique*. http://www.europeansocialsurvey.com.

Lohr, S.L. (1999). *Sampling*: *Design and Analysis*. Pacific Grove: Duxbury Press.

Lynn, P., et Gabler, S. (2005). Approximations de *b** dans la prévision des effets du plan dus à la mise en grappes. *Techniques d'enquête*, 31, 109-113.

Lynn, P., Gabler, S., Häder, S. et Laaksonen, S. (2007, a paraître). Methods for achieving equivalence of samples in cross-national surveys. *Journal of Official Statistics,* accepté.

Park, I., et Lee, H. (2004). Effets de plan pour les estimateurs pondérés de la moyenne et du total sous échantillonnage complexe. *Techniques d'enquête*, 30, 205-216.

Rao, C.R., et Kleffe, J. (1988). *Estimation of Variance Components and Applications*. Amsterdam: North-Holland.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

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 21, No. 4, 2005

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 33, No. 4, December/décembre 2005

Volume 34, No. 1, March/mars 2006

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 doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word. Une version papier pourrait être requise pour les formules et graphiques.

1. Présentation

1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
1.4 Les remerciements doivent paraître à la fin du texte.
1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) *etc*.
3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
3.5 Distinguer clairement les caractères ambigus (comme w, ω ; o, O, 0; l, 1).
3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.

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, page 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.