Nº 12-001-XIF au catalogue

Techniques d'enquête

Décembre 2005

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 : 1 800 263-1136).

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.

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 sur les commandes et les abonnements

Le produit nº 12-001-XIF au catalogue est publié deux fois par année sous format électronique au prix de 23 $CAN l'exemplaire et de 44 $CAN pour un abonnement annuel. Pour obtenir un exemplaire ou s'abonner, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique Nos produits et services.

Ce produit nº 12-001-XPB au catalogue est aussi disponible en version imprimée standard au prix de 30 $CAN l'exemplaire et de 58 $CAN pour un abonnement annuel. Les frais de livraison supplémentaires suivant s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 $CAN	12 $CAN
Autres pays	15 $CAN	30 $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
  120, avenue Parkdale
  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 Statistique Canada > Offrir des services aux Canadiens.

TECHNIQUES D'ENQUÊTE

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

DÉCEMBRE 2005 • VOLUME 31 • NUMÉRO 2

Publication autorisée par le ministre
responsable de Statistique Canada

Février 2006

N° 12-001-XIE au catalogue

Périodicité: semestrielle

ISSN 1712-5685

Ottawa

Statistique Canada Statistics Canada

Canada

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

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

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, (smj@statcan.ca, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, 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 *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, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au 1 800 700-1033, par télécopieur au 1 800 889-9734 ou par Courriel: order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et statisticiens du Québec.

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 31, numéro 2, décembre 2005

Table des matières


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2005
Vol. 31, N° 2, pp. 121-122
Statistique Canada, N° 12-001-XPB au catalogue

Dans ce numéro

C'est avec une profonde tristesse que nous annonçons le décès récent de M.P. Singh, rédacteur en chef de la revue *Techniques d'enquête* depuis la publication du tout premier numéro en 1975. Le présent numéro débute par un bref article nécrologique en sa mémoire.

Ce numéro de *Techniques d'enquête* contient aussi le cinquième article de la série d'articles annuels sollicités dédiée à Joseph Waksberg. Une brève biographie de ce dernier est parue dans le numéro de juin 2001 de la revue, en même temps que le premier article de la série. Je tiens à remercier les membres du comité de sélection – Michael Brick, président, David Bellhouse, Gordon Brackstone et Paul Biemer – d'avoir choisi Jon Rao comme auteur de l'article Waksberg de cette année.

Dans son article intitulé « Interaction entre la théorie et la méthodologie des enquêtes par sondage : Une évaluation », Rao montre comment les progrès théoriques stimulent l'élaboration de méthodes d'enquête et comment la pratique des enquêtes force à remettre la théorie en question. Il commence par résumer 50 années de contributions, de 1920 à 1970, puis présente une discussion plus approfondie de faits nouveaux récents dans plusieurs domaines. Enfin, il donne plusieurs exemples de théories importantes qui ne sont pas encore appliquées à grande échelle en pratique.

Dans leur article, Fuller et Kim élaborent et étudient une méthode d'imputation hot-deck efficace sous l'hypothèse que les probabilités de réponse sont égales dans les cellules d'imputation. La méthode qu'ils proposent est fondée sur la notion d'imputation fractionnaire et s'appuie sur des techniques de régression pour obtenir une approximation de la version entièrement efficace de l'imputation fractionnaire. Ils élaborent une estimation de la variance pour les méthodes de rééchantillonnage et montrent que la méthode qu'ils proposent donne de bons résultats dans une étude en simulation.

L'article de Brick, Jones, Kalton et Valliant décrit la comparaison, au moyen d'une étude en simulation, de trois méthodes d'estimation de la variance en présence d'imputation hot-deck, à savoir la méthode assistée par modèle, la méthode du jackknife corrigée et la méthode d'imputation multiple. Le but de l'étude en simulation est d'étudier les propriétés de ces estimateurs de la variance quand les hypothèses sous-jacentes ne sont pas vérifiées. Les auteurs constatent que le taux de couverture des intervalles de confiance ne s'approche pas du niveau nominal quand les estimations ponctuelles sont biaisées parce que l'on omet de tenir compte des domaines d'intérêt à l'étape de l'imputation. Ils concluent en notant que les différences entre les estimateurs de la variance sont trop faibles et incohérentes pour qu'on puisse affirmer que l'un d'entre eux est supérieur aux autres en général.

Little et Vartivarian étudient l'effet de la pondération pour la non-réponse sur l'erreur quadratique moyenne (EQM) d'un estimateur de la moyenne de population. Ils corrigent la pondération pour tenir compte de la non-réponse en ajustant les poids de sondage au moyen de l'inverse des taux de réponse dans les cellules. Ils concluent que, pour réduire le biais de non-réponse, une covariable de repondération doit avoir deux caractéristiques : elle doit être corrélée à la probabilité de réponse, d'une part, et à la variable de résultat, d'autre part. Si cette deuxième caractéristique est vérifiée, la repondération peut aussi réduire la variance due à la non-réponse. Ils proposent des estimations de l'EQM et les utilisent pour définir un estimateur composite. Celui-ci donne de bons résultats lors d'une évaluation par étude en simulation.

O'Malley et Zaslavsky présentent des modèles de fonctions de variance et de covariance généralisées (FVCG) pour des moyennes multivariées de questions d'enquête ordonnées, dans le cas de données complètes ainsi que de données avec non-réponse structurée. Ils commencent par décrire l'élaboration et l'évaluation de leurs méthodes, puis ils illustrent ces dernières à l'aide de données provenant de la Consumer Assessments of Health Plans Study. Dans la conclusion, ils discutent de certaines questions liées à l'application des FVCG.

L'article de Singh, Shukla et Kundu décrit l'élaboration de modèles spatiaux et spatio-temporels pour l'estimation sur petits domaines, ainsi que pour l'estimation de l'erreur quadratique moyenne du meilleur prédicteur linéaire sans biais empirique (EBLUP) résultant. Ils appliquent leurs modèles aux données sur les dépenses de consommation mensuelles par habitant et concluent qu'ils peuvent être très efficaces s'il existe des corrélations importantes dues aux effets de quartier.

Belsby, Bjørnstad et Zhang discutent de la modélisation en vue d'estimer le nombre de ménages de diverses tailles en présence de non-réponse non ignorable. Ils modélisent le mécanisme de réponse sachant la taille du ménage, en utilisant la taille enregistrée de la famille comme donnée supplémentaire. Ils décrivent d'abord l'élaboration de leurs méthodes de modélisation, puis produisent et évaluent des estimations à l'aide de données provenant de l'Enquête sur les dépenses de consommation en Norvège de 1992.

Nandram, Cox et Choi considèrent l'analyse de données catégoriques provenant d'un seul tableau à double entrée en présence de non-réponse partielle ainsi que totale ou, selon leur terminologie, de classification partielle. Ils proposent d'utiliser une méthode bayésienne pour modéliser divers scénarios de données manquantes sous les hypothèses d'ignorabilité et de non-ignorabilité. Ils illustrent leurs méthodes à l'aide de données bivariées incomplètement observées provenant de la National Health and Nutrition Examination Survey, où les variables pour lesquelles des données manquent sont la densité minérale osseuse et le revenu familial.

Dans la première de trois notes brèves publiées dans le présent numéro, Beaumont discute de l'utilisation de l'information sur le processus de collecte des données lors de la correction de la pondération pour tenir compte de la non-réponse. Puis, il donne un exemple tiré de l'Enquête sur la population active du Canada en utilisant le nombre de tentatives de prise de contact avec une unité étudiée. Un résultat important est que, si l'information sur le processus de collecte peut être considérée comme étant aléatoire, la méthode n'introduit aucun biais.

En partant de principes fondamentaux, Bustos dérive une forme explicite de la fonction de probabilité d'un échantillon ordonné. Puis, il montre comment on peut l'utiliser pour calculer les probabilités d'inclusion et offre des exemples pour des plans de sondage courants. Enfin, il donne la forme générale de la matrice des corrélations des unités d'échantillonnage, qui dépend uniquement des probabilités d'inclusion.

Enfin, dans son article, Wu passe brièvement en revue certains aspects théoriques de la méthode de la pseudo-vraisemblance empirique en échantillonnage et présente des algorithmes permettant de calculer l'estimateur du maximum de pseudo-vraisemblance empirique et de construire les intervalles de confiance des rapports de pseudo-vraisemblance empirique. Il donne des fonctions utilisant les logiciels statistiques R et S-PLUS pour faciliter l'implémentation de ces algorithmes dans le cas d'enquêtes réelles ou d'études en simulation.

Harold Mantel

Techniques d'enquête, décembre 2005
Vol. 31, N° 2, pp. 123
Statistique Canada, N° 12-001-XPB au catalogue

À la mémoire
M.P. Singh
(1941-2005)

Dr. Mangala P. Singh né en Inde le 26 décembre 1941, il avait obtenu un doctorat de l'Indian Statistical Institute en 1969, avec une spécialisation en échantillonnage d'enquêtes. Il s'était joint à Statistique Canada en 1970, où il avait atteint le poste de directeur de la division des méthodes d'enquêtes des ménages en 1994, poste qu'il a occupé jusqu'à son décès le 24 août 2005.

M.P., comme tout le monde l'appelait, était une figure de proue pour l'application des méthodes statistiques à Statistique Canada. Il était probablement plus étroitement associé à l'enquête sur la population active, l'une des enquêtes les plus importantes de l'agence. Il a dirigé la méthodologie de l'enquête sur la population active au cours de plusieurs remaniements durant les années 1970, 1980, 1990 et au début du 21 è siècle, en introduisant à chaque fois des innovations tout en s'assurant que les changements étaient valides et bien testés. Au cours des dernières années de sa carrière, il a veillé au développement de plusieurs enquêtes nouvelles et innovatrices dans le domaine de la santé et au développement de programmes statistiques dans les domaines des dépenses des ménages, de l'éducation et de la justice.

Le rôle de M.P. comme rédacteur en chef de la revue Techniques d'enquête a transformé le rôle des techniques d'enquête, à la fois au Canada et à l'étranger. M.P. a été le rédacteur fondateur de la revue , et au cours de 30 années, l'a fait évoluer jusqu'à ce qu'elle devienne un publication amirale de Statistique Canada. Grâce à sa capacité à attirer un réseau de rédacteurs adjoints et de contributeurs, Techniques d'enquête est maintenant reconnue comme une des revues prééminentes dans son domaine à travers le monde. Même au cours des dernières années, M.P. a continué d'innover, comme avec l'introduction de la série d'articles Waksberg et de la publication électronique.

M.P. a été une source de plusieurs autres "grandes idées" tout au long de sa carrière à Statistique Canada. Durant les années 1970, il a contribué à gagner des appuis pour l'idée d'un fonds permanent pour la recherche en méthodologie, et il a présidé personnellement le Comité de recherche et développement en méthodologie à ses débuts. Il a encouragé de nombreux chercheurs et s'est donné beaucoup de mal pour qu'ils se sentent chez eux à Statistique Canada. La soixantaine n'a pas du tout endigué le flot de ses idées. M.P. a déployé une énergie considérable au cours des quatre dernières années pour proposer une révision majeure de la façon d'effectuer les enquêtes sur les ménages au Canada. Comme résultat de ses efforts, on travaille à travers Statistique Canada à des façons de mettre en oeuvre sa vision, et son influence sur les enquêtes des ménages se fera sentir pour des années à venir.

M.P. était spécialement attaché à la recherche en statistique et à la profession statistique. Il était l'auteur de plus de 40 articles dans des revues internationales, le coéditeur de deux livres publiés chez Wiley&Sons, et a organisé plusieurs séances et fait des presentations lors de plusieurs conférences statistiques. Il a siégé sur plusieurs comités et ateliers de travail de la Société statistique du Canada, de l'International Statistical Institute et de l'American Statistical Association. Il a aussi été secrétaire du comité consultatif de Statistique Canada sur le méthodes statistiques. En retour, ila reçu les honneurs de la profession: il a été élu à l'International Statistical Institute en 1975, et en 1988 il est devenu Fellow de l'American Statistical Association.

Néanmoins, c'est son influence sur toute une génération de statisticians qui constitue son plus grand héritage. Il a été un mentor, un moniteur, un père et un ami pour tous ceux qui l'ont connu. Il a inspiré les autres à offrir le meilleur d'eux-mêmes, ce qu'ils ont fait. Il était toujours prêt pour un rire,un sourire et un mot amical d'encouragement. Il a consacré sa vie à la profession statistique et c'est à travers ceux qu'il a atteints qu'on peut mesurer sa veritable contribution.

Il laisse dans le deuil son épouse Savitri, ses deux filles Mala et Mamta, et son fils Rahul.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2005
Vol. 31, n° 2, pp. 125-126
Statistique Canada, N° 12-001-XPB au catalogue

Série Waksberg d'articles sollicités

La revue *Techniques d'enquête* a mis sur pied une série de communications sollicitées en l'honneur de Joseph Waksberg, qui a fait de nombreuses contributions importantes à la méthodologie d'enquête. Chaque année, un éminent chercheur est choisi pour rédiger un article pour la série de communications sollicitées de Waksberg. L'article examine les progrès et l'état actuel d'un thème important dans le domaine de la méthodologie d'enquête et reflète l'agencement de théorie et de pratique caractéristique des travaux de Waksberg. L'auteur reçoit une prime en argent qui provient d'une bourse de Westat, en reconnaissance des contributions de Joe Waksberg pendant ses nombreuses années de collaboration avec Westat. L'administration financière de la bourse est assurée par l'American Statistical Association. Gad Nathan, Wayne Fuller, Tim Holt, Norman Bradburn, Jon Rao et Alastair Scott sont les gagnants précédents. Les cinq premiers articles de la série sont déjà parus dans la revue *Techniques d'enquête*.

Précédents gagnants du prix Waksberg :

Gad Nathan (2001)
Wayne A. Fuller (2002)
Tim Holt (2003)
Norman Bradburn (2004)
J.N.K. Rao (2005)

Nominations:

L'auteur de l'article Waksberg de 2007 sera sélectionné par un comité de quatre personnes désignées par *Techniques d'enquête* et l'American Statistical Association. Les candidatures ou les suggestions de sujets doivent être envoyées à Gordon Brackstone, président du comité, à 78, chemin Charing, Ottawa (Ontario), Canada, K2G 4C9, par courriel à Gordon.brackstone@sympatico.ca ou par télécopieur au (613) 951-1394. Les candidatures et les suggestions de sujets doivent être reçues d'ici au 28 février 2006.

Article sollicité Waksberg 2005

Auteur: J.N.K. Rao

J.N.K. Rao est professeur distingué de recherche à l'Université Carleton d'Ottawa. Il a publié plusieurs articles sur une vaste étendue de sujets en théorie et méthodes de sondages et est auteur du livre de 2003 chez Wiley "Small Area Estimation". Son intérêt pour la recherche en échantillonnage inclue l'analyse de données d'enquêtes, l'estimation pour petites régions, les données manquantes et l'imputation, les méthodes de ré-échantillonnage et l'inférence à l'aide de la vraisemblance empirique. Son article du JASA de 1981 (avec A.J. Scott) sur l'analyse de données d'enquêtes a été sélectionné parmi les articles phares (landmark paper) de la théorie et des méthodes d'échantillonnage. Il est membre du Comité consultatif des méthodes statistiques de Statistique Canada depuis sa création il y a 20 ans. Il est fellow de la Société Royale du Canada et a reçu la médaille d'or de la Société statistique du Canada en 1994.

Membres du comité de sélection de l'article Waskberg (2005-2006)

Gordon Brackstone, (Président)
Wayne Fuller, *Iowa State University*
Sharon Lohr, *Arizona State University*

Présidents précédents :

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)
David A. Binder (2002 - 2003)
J. Michael Brick (2003 - 2004)
David R. Bellhouse (2004 - 2005)

Techniques d'enquête, décembre 2005
Vol. 31, N° 2, pp. 127-151
Statistique Canada, N° 12-001-XPB au catalogue

Évaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage

J.N.K. Rao [1]

Résumé

Une grande partie de la théorie des enquêtes par sondage a été motivée directement par des problèmes d'ordre pratique survenus au moment de la conception et de l'analyse des enquêtes. En revanche, la théorie des enquêtes par sondage a influencé la pratique, ce qui a souvent donné lieu à des améliorations importantes. Dans le présent article, nous examinons cette interaction au cours des 60 dernières années. Nous présentons également des exemples où une nouvelle théorie est nécessaire ou encore où la théorie existe sans être utilisée.

Mots clés : Analyse des données d'enquête; apports antérieurs; question d'inférence; méthodes de rééchantillonnage; estimation sur petits domaines.

1. Introduction

Dans cet article, je vais examiner l'inter-relation entre la théorie des sondages et la pratique dans les quelques 60 dernières années. Je vais couvrir une grande variété de sujets: les premières contributions significatives qui ont grandement influencé la pratique, les questions d'inférence, l'estimation par calage qui assure la cohérence aux totaux établis de variables auxiliaires, l'échantillonnage à probabilités inégales sans remplacement, l'analyse de données d'enquêtes, le rôle des méthodes de ré-échantillonnage, et l'estimation pour petits domaines. Je vais aussi présenter quelques exemples où il y a soit besoin d'une nouvelle théorie soit une théorie existante qui n'est pas tellement utilisée.

2. Quelques apports marquants : 1920 – 1970

La présente section rend compte de certains apports marquants à la théorie et aux méthodes des enquêtes par sondage, apports qui ont grandement influencé la pratique. Le statisticien norvégien A.N. Kiaer (1897) fut sans doute le premier à promouvoir l'échantillonnage (appelé « méthode représentative » à l'époque) plutôt qu'un dénombrement complet, quoique la plus ancienne référence à l'échantillonnage remonte au grand récit épique indien Mahabharata (Hacking 1975, page 7). Dans la méthode représentative, l'échantillon doit refléter la population mère finie; à cette fin, on procède par échantillonnage équilibré au moyen de la sélection raisonnée ou par échantillonnage aléatoire. On utilisa la méthode représentative en Russie dès 1900 (Zarkovic 1956) et, vers la même époque, Wright l'employa pour mener des enquêtes par sondage aux États-Unis. Dans les années 1920, on utilisait abondamment la méthode représentative, et l'Institut international de statistique joua un rôle de premier plan en créant en 1924 un comité chargé de produire un rapport sur cette méthode. Le rapport de ce comité portait sur des aspects théoriques et pratiques de la méthode d'échantillonnage aléatoire. Bowley (1926) contribua à ce rapport par ses travaux fondamentaux sur l'échantillonnage aléatoire stratifié avec répartition proportionnelle, qui permit de tirer un échantillon représentatif avec probabilités d'inclusion égales. Hubback (1927) prit conscience de la nécessité d'un échantillonnage aléatoire dans les enquêtes sur les cultures : « La seule façon d'arriver à une estimation satisfaisante consiste à établir une approximation de l'échantillonnage aléatoire aussi proche que les circonstances le permettent, car ainsi, non seulement on élimine les limites personnelles de l'expérimentateur, mais il devient possible de déterminer la probabilité avec laquelle les résultats d'un nombre donné d'échantillons se situeront à l'intérieur d'une étendue donnée par rapport à la moyenne arithmétique. Concrètement, il s'agit de trouver combien d'échantillons sont nécessaires pour assurer que la probabilité soit d'au moins 20:1 par rapport à la moyenne des échantillons à l'intérieur d'un maund de la vraie moyenne. » Cet énoncé contient deux observations importantes concernant l'échantillonnage aléatoire : 1) il évite les biais personnels dans la sélection d'un échantillon; 2) on peut déterminer la taille de l'échantillon pour satisfaire une marge d'erreur spécifiée par rapport à une chance de 1 sur 20. Mahalanobis (1946b) a observé que les travaux fondamentaux de R.A. Fisher sur la conception des expériences, menés à la Rothamsted Experimental Station, furent directement influencés par Hubback (1927).

Dans un article marquant, devenu un classique, Neyman (1934) a jeté les bases théoriques de l'échantillonnage

1. J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, (Ontario), Canada, K1S 5B6.

probabiliste (ou fondé sur le plan de sondage) en ce qui concerne l'inférence à partir d'échantillons d'enquête. Il a montré, avec des arguments théoriques et des exemples pratiques, que l'échantillonnage aléatoire stratifié était préférable à l'échantillonnage équilibré, car ce dernier peut donner de mauvais résultats si les hypothèses sous-jacentes du modèle sont violées. Neyman a également avancé, dans sa théorie de l'échantillonnage aléatoire stratifié sans remise, les notions d'efficience et de répartition optimale en assouplissant la condition des probabilités d'inclusion égales. En généralisant le théorème de Markov sur l'estimation par les moindres carrés, Neyman a prouvé que la moyenne stratifiée, $\bar{y}_{st} = \sum_h W_h \bar{y}_h$, était le meilleur estimateur de la moyenne de population, $\bar{Y} = \sum_h W_h \bar{Y}_h$, dans la classe linéaire d'estimateurs sans biais de forme $\bar{y}_b = \sum_h W_h \sum_i b_{hi} y_{hi}$, où $W_h, \bar{y}_h$ et $\bar{Y}_h$ sont le poids, la moyenne d'échantillon et la moyenne de population de la h^{e} strate ($h = 1, ..., L$), et b_{hi} est une constante associée à la valeur de l'élément y'_{hi} observée au moment du i^{e} tirage d'échantillon ($i = 1, ..., n_h$) dans la h^{e} strate. On a obtenu la répartition optimale ($n_1, ..., n_L$) de la taille de l'échantillon total, n, en minimisant la variance de $\bar{y}_{st}$ sous réserve de $\sum_h n_h = n$; on a découvert plus tard une preuve antérieure de la répartition de Neyman par Tschuprow (1923). Neyman a également proposé une inférence à partir de grands échantillons en fonction d'intervalles de confiance selon la théorie normale, de manière que la fréquence des erreurs dans les énoncés de confiance en fonction de tous les échantillons aléatoires stratifiés qu'il est possible de tirer n'excède pas la limite prescrite à l'avance « quelles que soient les propriétés inconnues de la population ». Une méthode d'échantillonnage qui satisfait l'énoncé de fréquence susmentionné est dite « représentative ». Il est à noter que Hubback (1927) avait déjà fait allusion à l'énoncé de fréquence associé à l'intervalle de confiance. Dans son dernier apport à la théorie des enquêtes par sondage, Neyman (1938) a étudié l'échantillonnage à deux phases de stratification et calculé la taille optimale des échantillons de première phase et de deuxième phase, n' et n, en minimisant la variance de l'estimateur sous réserve d'un coût donné $C = n'c' + nc$, où le coût par unité de deuxième phase, c, est élevé par rapport au coût par unité de la première phase, c'.

Au cours des années 1930, la demande d'information a connu une croissance rapide et l'on a pris conscience des avantages de l'échantillonnage probabiliste–portée accrue, réduction de coût, plus grande vitesse et caractéristiques indépendantes d'un modèle–, d'où une augmentation du nombre et du type d'enquêtes menées par échantillonnage probabiliste et couvrant de grandes populations. La presque totalité des statisticiens d'enquête ont adopté l'approche de Neyman. En outre, cette dernière a inspiré divers ajouts importants, motivés surtout par des critères d'ordre pratique et d'efficience. L'article marquant de Cochran (1939) présente plusieurs résultats importants : le recours à l'analyse de variance pour estimer l'amélioration de l'efficience due à la stratification, l'estimation des composantes de la variance dans l'échantillonnage à deux degrés en vue d'études futures sur un sujet semblable, le choix de l'unité d'échantillonnage, l'estimation par régression sous échantillonnage à deux phases et l'effet des erreurs dans la taille des strates. Dans cet article, Neyman a également proposé le concept de superpopulation : « La population finie doit être considérée comme un échantillon aléatoire d'une population infinie. » Il est intéressant de noter qu'à l'époque, Cochran n'était pas d'accord avec le concept traditionnel de population fixe : « En outre, il est loin d'être réaliste de considérer la population comme un lot fixe de nombres connus. » Cochran (1940) a proposé l'estimation par quotient pour les enquêtes par sondage, mais Laplace (1820) avait déjà utilisé l'estimateur par quotient. Dans un autre article marquant, Cochran (1942) a formulé la théorie de l'estimation par régression. Il a calculé la variance conditionnelle de l'estimateur par régression habituel pour un échantillon fixe ainsi qu'un estimateur échantillon de cette variance, en supposant un modèle de régression linéaire $y = \alpha + \beta x + e$, où e a une moyenne nulle et une variance constante dans les séries statistiques dans lesquelles x est fixe. Il a également noté que l'estimateur par régression restait sans biais (par rapport au modèle) sous échantillonnage non aléatoire, à condition que le modèle de régression linéaire hypothétique soit correct. Il a calculé le biais moyen en présence d'écarts par rapport au modèle (notamment dans le cas de la régression quadratique) pour l'échantillonnage aléatoire simple à mesure que la taille de l'échantillon n augmente. Cochran a ensuite étendu ses résultats à la régression pondérée et calculé le résultat d'optimalité, aujourd'hui bien connu, pour l'estimateur par quotient; selon lui, il s'agit de « la meilleure estimation linéaire sans biais si la valeur moyenne et la variance changent proportionnellement à x ». Dans les travaux récents, ce dernier modèle est appelé modèle par quotient. Madow et Madow (1944) et Cochran (1946) ont comparé la variance prévue sous un modèle de superpopulation pour étudier analytiquement l'efficience relative de l'échantillonnage systématique et de l'échantillonnage aléatoire stratifié. Cet article a incité d'autres chercheurs à mener des travaux sur l'utilisation de modèles de superpopulation dans le choix de stratégies d'échantillonnage probabiliste, ainsi que sur l'inférence dépendante d'un modèle et l'inférence assistée par un modèle (voir la section 3).

En Inde, Mahalanobis a fait un apport innovateur à la théorie de l'échantillonnage en formulant des fonctions de coût et de variance pour la conception d'enquêtes. Son

article marquant (Mahalanobis 1944) présente des résultats théoriques probants sur la conception efficace d'enquêtes par sondage et leurs applications pratiques, notamment dans le cas d'enquêtes sur les surfaces cultivées et le rendement des cultures. Maintenant bien connue, la répartition optimale sous échantillonnage aléatoire stratifié où le coût par unité varie d'une strate à l'autre est obtenue sous forme de cas particulier de sa théorie générale. Dès 1937, Mahalanobis a utilisé des plans de sondage à plusieurs degrés pour les enquêtes sur le rendement des cultures avec, comme unités d'échantillonnage aux quatre degrés d'échantillonnage, des villages, des grilles à l'intérieur des villages, des parcelles à l'intérieur des grilles et des coupes de tailles et de formes différentes (Murthy 1964). Il a également utilisé un plan de sondage à deux phases pour estimer le rendement de l'écorce de quinquina. Il a joué un rôle de premier plan dans l'établissement de la National Sample Survey (NSS) de l'Inde, la plus vaste enquête polyvalente permanente : un personnel à temps plein effectue des interviews sur place pour des enquêtes socioéconomiques et des mesures physiques pour des enquêtes sur les cultures. Plusieurs éminents statisticiens d'enquête, dont D.B. Lahiri et M.N. Murthy, ont collaboré à la NSS.

P.V. Sukhatme, qui a étudié avec Neyman, a également fait un apport innovateur à la conception et à l'analyse d'enquêtes agricoles à grande échelle en Inde, en utilisant l'échantillonnage stratifié à plusieurs degrés. À partir de 1942–1943, il a mis au point des plans de sondage efficients pour mener des enquêtes nationales sur les cultures de blé et de riz et a obtenu un degré élevé de précision pour les estimations nationales ainsi qu'une marge d'erreur raisonnable pour les estimations par district. L'approche de Sukhatme différait de celle de Mahalanobis, qui utilisait des parcelles de très petite taille pour les coupes-témoins et employait des enquêteurs *ad hoc*. Sukhatme (1947) et Sukhatme et Panse (1951) ont démontré que l'utilisation d'une petite parcelle pourrait donner des estimations biaisées à cause de la tendance à placer des plantes de bornage à l'intérieur de la parcelle lorsqu'il y a un doute. Ils ont également souligné que le recours à des enquêteurs *ad hoc*, qui se déplacent rapidement d'un endroit à l'autre, obligeait à mesurer uniquement les parcelles de champs échantillonnés qui sont prêts à moissonner à la date de la visite, ce qui est contraire au principe de l'échantillonnage aléatoire. La solution de Sukhatme consistait à utiliser de grandes parcelles pour éviter les biais liés au bornage et à confier les coupes-témoins à l'organisme public local chargé du revenu ou de l'agriculture.

De 1940 à 1970, les statisticiens d'enquête du U.S. Census Bureau, sous la direction de Morris Hansen, William Hurwitz, William Madow et Joseph Waksberg, ont fait des apports fondamentaux à la théorie et à la pratique des enquêtes par sondage, et bon nombre de leurs méthodes sont encore largement utilisées dans la pratique. Hansen et Hurwitz (1943) ont formulé la théorie de base de l'échantillonnage stratifié à deux degrés, une seule unité primaire d'échantillonnage (UPÉ) à l'intérieur de chaque strate étant tirée avec probabilité proportionnelle à la taille (échantillonnage PPT) puis sous-échantillonnée à un rythme qui assure l'autopondération (probabilités de sélection globales égales) à l'intérieur des strates. Cette approche permet de confier aux intervieweurs des charges de travail à peu près égales, ce qui est souhaitable dans le contexte des enquêtes sur le terrain. Elle permet aussi de réduire considérablement la variance en neutralisant la variabilité due à la taille inégale des UPÉ sans vraiment stratifier selon la taille, ce qui permet la stratification selon d'autres variables pour réduire la variance. Par contre, les charges de travail peuvent varier considérablement si les UPÉ sont sélectionnées par échantillonnage aléatoire simple, puis sous-échantillonnées au même rythme à l'intérieur de chaque strate. Aujourd'hui, on utilise abondamment l'échantillonnage PPT des UPÉ dans la conception d'enquêtes à grande échelle, mais on sélectionne dans chaque strate deux ou plusieurs UPÉ sans remise, de sorte que les probabilités d'inclusion des UPÉ sont proportionnelles à la taille (voir la section 5).

Bon nombre d'enquêtes à grande échelle sont répétées au fil du temps, comme l'Enquête sur la population active (EPA) du Canada, menée chaque mois, et la Current Population Survey (CPS) des États-Unis, avec remise partielle des unités finales (appelée aussi échantillonnage par renouvellement). Dans le cas de l'EPA, par exemple, l'échantillon de ménages est divisé en six groupes de renouvellement (échantillons constants ou panels) et un groupe de renouvellement reste dans l'échantillon pendant six mois consécutifs, puis est retiré de l'échantillon, ce qui donne un chevauchement de cinq sixièmes entre deux mois consécutifs. Dans la foulée des travaux initiaux de Jessen (1942) sur l'échantillonnage à deux reprises avec remise partielle des unités, Yates (1949) et Patterson (1950) ont jeté les bases théoriques de la conception et de l'estimation d'enquêtes à passages répétés et démontré qu'on pouvait améliorer l'efficience de l'estimation de niveau et de changement en tirant parti des données antérieures. Hansen, Hurwitz, Nisseslson et Steinberg (1955) ont mis au point des estimateurs plus simples, appelés estimateurs composites K, applicables aux plans d'échantillonnage stratifié à plusieurs degrés avec échantillonnage PPT au premier degré. Rao et Graham (1964) ont étudié des politiques de remise optimale pour les estimateurs composites K. On a également proposé divers ajouts. On a utilisé des estimateurs composites dans le cas de la CPS et d'autres enquêtes permanentes à grande échelle. Encore récemment, l'EPA du Canada a adopté l'estimation composite, appelée estimation composite par

régression, qui utilise l'information sur l'échantillon obtenue au cours des mois précédents et qui peut être mise en œuvre avec un logiciel de poids de régression (voir la section 4).

Keyfitz (1951) a proposé une méthode ingénieuse pour obtenir de meilleures mesures de la taille des UPÉ dans les enquêtes permanentes fondées sur les plus récents dénombrements censitaires. Sa méthode permet de maximiser la probabilité de chevauchement avec l'échantillon antérieur d'une UPÉ par strate, ce qui réduit les coûts d'opération sur le terrain tout en améliorant l'efficience grâce aux meilleures mesures de la taille dans l'échantillonnage PPT. L'EPA du Canada et d'autres enquêtes permanentes ont utilisé la méthode de Keyfitz. Raj (1956) a formulé le problème de l'optimisation comme un « problème de transport » dans la programmation linéaire. Kish et Scott (1971) ont étendu la méthode de Keyfitz aux mesures changeantes des strates et de la taille. Ernst (1999) a brossé un excellent tableau de l'évolution, au cours des 50 dernières années, de la coordination d'échantillons (qui consiste à maximiser ou minimiser le chevauchement des échantillons) au moyen d'algorithmes de transport et de méthodes connexes; voir aussi Mach, Reiss et Schiopu-Kratina (2005) en ce qui concerne les applications aux enquêtes-entreprises avec création et suppression d'entreprises.

Dalenius (1957, chapitre 7) a étudié le problème de la stratification optimale d'un nombre donné de strates, L, dans le cadre de la répartition de Neyman. Dalenius et Hodges (1959) ont obtenu une approximation simple de la stratification optimale, appelée méthode de la fonction cumulative de la racine carrée des fréquences (cum $\sqrt{f}$), qui est abondamment utilisée dans la pratique. Pour les populations très asymétriques dont un petit nombre d'unités comptent pour une forte proportion du total Y, comme les populations d'entreprises, une stratification efficace nécessite une strate à tirage complet ($n_1 = N_1$) de grandes unités et des strates à tirage partiel d'unités moyennes et petites. Lavallée et Hidiroglou (1988) et Rivest (2002) ont mis au point des algorithmes pour déterminer les bornes de stratification en utilisant la méthode puissance (Fellegi 1981; Bankier 1988) et la répartition de Neyman pour les strates à tirage partiel. Aujourd'hui, Statistique Canada et d'autres organismes utilisent ces algorithmes pour les enquêtes-entreprises.

Avant 1950, la recherche portait sur l'estimation de totaux et de moyennes de population pour la population entière et de grandes sous-populations planifiées, comme des États ou des provinces. Or, les utilisateurs s'intéressent également aux totaux et aux moyennes de sous-populations non planifiées (appelées aussi domaines), comme les groupes d'âge-sexe à l'intérieur d'une province, ainsi qu'à des paramètres autres que les totaux et les moyennes, comme les médianes et d'autres quantiles, par exemple le revenu médian. Hartley (1959) a formulé une théorie simple et unifiée de l'estimation par domaine, applicable à n'importe quel plan de sondage et nécessitant uniquement les formules-types pour l'estimateur du total et son estimateur de variance, dénotés respectivement $\hat{Y}(y)$ et $v(y)$ dans la notation d'opérateur. Il a introduit deux variables synthétiques ${}_j y_i$ et ${}_j a_i$ qui prennent respectivement les valeurs y_i et 1 si l'unité i appartient au domaine j et qui sont nulles dans le cas contraire. Alors, on obtient simplement les estimateurs du total de domaine ${}_j Y = Y({}_j y)$ et de la taille de domaine ${}_j N = Y({}_j a)$ à l'aide des formules pour $\hat{Y}(y)$ et $v(y)$ en remplaçant respectivement y_i par ${}_j y_i$ et ${}_j a_i$. De même, on obtient les estimateurs des moyennes de domaine et des différences de domaine ainsi que leurs estimateurs de variance à l'aide des formules de base pour $\hat{Y}(y)$ et $v(y)$. Durbin (1968) a également obtenu des résultats semblables. Aujourd'hui, on pratique couramment l'estimation par domaine en utilisant l'ingénieuse méthode de Hartley.

Pour l'inférence concernant des quantiles, Woodruff (1952) a proposé une méthode simple et ingénieuse pour obtenir un intervalle de confiance de niveau $(1-\alpha)$ sous des plans d'échantillonnage généraux, en utilisant uniquement la fonction de distribution estimative et son erreur-type (voir l'ouvrage de Lohr (1999), pages 311 à 313). Il est à noter qu'on obtient simplement ces dernières à l'aide des formules pour un total en remplaçant y par une variable indicatrice. En mettant sur le même pied l'intervalle de Woodruff et un intervalle selon la théorie normale à l'égard du quantile, on peut aussi obtenir une formule simple pour l'erreur-type du p^{e} estimateur de quantile, soit la moitié de la longueur de l'intervalle divisé par le point supérieur $\alpha/2$ de la distribution normalisée $N(0, 1)$ qui égale 1,96 si $\alpha = 0{,}05$ (Rao et Wu 1987; Francisco et Fuller 1991). L'intervalle de Woodruff possède une propriété étonnante : il donne de bons résultats même lorsque p est petit ou grand et que la taille de l'échantillon est moyenne (Sitter et Wu 2001).

On s'est rendu compte de l'importance des erreurs de mesure dès les années 1940. Dans un article influent, Mahalanobis (1946a) a mis au point la technique des sous-échantillons superposés (appelée échantillonnage répété par Deming 1960). En Inde, on a beaucoup utilisé cette méthode dans les enquêtes par sondage à grande échelle pour évaluer les erreurs d'échantillonnage et les erreurs de mesure. L'échantillon est tiré sous forme de deux ou plusieurs sous-échantillons indépendants selon le même plan de sondage, de sorte que chaque sous-échantillon fournit une estimation valide du total ou de la moyenne. Les sous-échantillons sont attribués à des intervieweurs différents (ou à des équipes différentes), ce qui produit une estimation valide de la

variance totale qui tient compte de la variance de réponse corrélée due aux intervieweurs. Les sous-échantillons superposés entraînent une augmentation des frais de déplacement des intervieweurs, mais on peut les réduire en modifiant les affectations des intervieweurs. Hansen, Hurwitz, Marks et Mauldin (1951), Sukhatme et Seth (1952) et Hansen, Hurwitz et Bershad (1961) ont formulé des théories de base sous des modèles d'erreur de mesure additive et décomposé la variance totale en trois éléments : la variance d'échantillonnage, la variance de réponse simple et la variance de réponse corrélée. On a montré que la variance de réponse corrélée due aux intervieweurs était de l'ordre de k^{-1} sans égard à la taille de l'échantillon, k étant le nombre d'intervieweurs. Par conséquent, elle peut dominer la variance totale si k n'est pas un nombre élevé. Lors du recensement de 1950 aux États-Unis, l'étude de la variance due aux intervieweurs a montré que cette composante était effectivement grande pour les petits domaines. C'est en partie pour cette raison que lors du recensement de 1960, on a adopté l'autodénombrement par la poste pour réduire cette composante de la variance (Waksberg 1998). Il s'agit d'un exemple éloquent de l'influence de la théorie sur la pratique. Fellegi (1964) a proposé de combiner la superposition et la répétition pour estimer la covariance entre l'écart d'échantillonnage et l'écart de réponse. Cette composante est souvent négligée dans la décomposition de la variance totale, mais elle pourrait être appréciable dans la pratique.

Le concept de l'effet du plan de sondage (EPS), dû à Leslie Kish (voir Kish 1965, section 8.2), constitue un autre jalon de la méthodologie des enquêtes par sondage. L'effet du plan de sondage est le ratio de la variance réelle d'une statistique sous le plan de sondage spécifié à la variance qui serait obtenue sous échantillonnage aléatoire simple de même taille. Ce concept est particulièrement utile dans la présentation et la modélisation des erreurs d'échantillonnage, ainsi que dans l'analyse des données d'enquête complexes faisant intervenir la mise en grappes et les probabilités de sélection inégales (voir la section 6).

Le lecteur trouvera dans Kish (1995), Kruskal et Mosteller (1980), Hansen, Dalenius et Tepping (1985) et O'Muircheartaigh et Wong (1981) un examen des apports marquants à la théorie et aux méthodes des enquêtes par sondage.

3. Questions d'inférence

3.1 Cadre unifié fondé sur le plan de sondage

Au départ, l'élaboration de la théorie de l'échantillonnage a progressé de manière plus ou moins inductive, quoique Neyman (1934) ait étudié la meilleure estimation linéaire sans biais pour l'échantillonnage aléatoire stratifié. On a envisagé des stratégies (plan de sondage et estimation) qui semblaient raisonnables et l'on a soigneusement étudié des propriétés relatives au moyen de méthodes analytiques ou empiriques, en comparant surtout des erreurs quadratiques moyennes, et parfois aussi des erreurs quadratiques moyennes ou des variances prévues sous des modèles de superpopulation plausibles, comme nous le mentionnons dans la section 2. On n'a pas insisté sur une estimation sans biais sous un plan de sondage donné, car elle « entraîne souvent une erreur quadratique moyenne beaucoup plus grande que nécessaire » (Hansen, Hurwitz et Tepping 1983). On a plutôt jugé que la cohérence avec le plan de sondage était nécessaire pour les grands échantillons. Les ouvrages classiques de Cochran (1953), Deming (1950), Hansen, Hurwitz et Madow (1953), Sukhatme (1954) et Yates (1949), fondés sur l'approche susmentionnée, ont grandement influencé la pratique des enquêtes. Pourtant, les statisticiens universitaires accordaient peu d'attention à la théorie de l'échantillonnage traditionnelle, peut-être parce qu'il lui manquait un cadre théorique formel et qu'elle n'était pas intégrée à la théorie statistique courante. Plusieurs départements de statistique nord-américains de prestige n'offraient pas de cours supérieurs en théorie de l'échantillonnage.

Dans les années 1950, on a élaboré des cadres et des approches théoriques formels pour intégrer la théorie de l'échantillonnage à l'inférence statistique courante dans des conditions quelque peu idéalistes axées sur les erreurs d'échantillonnage, en supposant l'absence d'erreurs de mesure ou de réponse ainsi que de non-réponse. Horvitz et Thompson (1952) ont apporté une contribution de base à l'échantillonnage avec probabilités de sélection arbitraires en formulant trois sous-classes d'estimateurs linéaires sans biais d'un total Y, dont la classe de Markov étudiée par Neyman. Une autre sous-classe avec poids de sondage d_i lié à une unité d'échantillonnage i et dépendant uniquement de i admettait l'estimateur bien connu avec poids inversement proportionnel à la probabilité d'inclusion π_i comme seul estimateur sans biais. Narain (1951) ayant également découvert cet estimateur, on devrait l'appeler l'estimateur de Narain-Horvitz-Thompson (NHT) au lieu de l'estimateur HT comme on l'appelle couramment. Pour l'échantillonnage aléatoire simple, la moyenne d'échantillon est le meilleur estimateur linéaire sans biais (best linear unbiased estimator ou BLUE) de la moyenne de population dans les trois sous-classes, mais ce n'est pas suffisant pour prétendre que la moyenne d'échantillon est le meilleur de tous les estimateurs linéaires sans biais. Godambe (1955) a proposé une classe générale d'estimateurs linéaires sans biais d'un total Y en supposant des données-échantillons $\{(i, y_i), i \in s\}$ et un poids dépendant de l'unité d'échantillonnage i ainsi que des autres unités échantillonnées s,

c'est-à-dire un poids de forme $d_i(s)$. Il a alors établi que l'estimateur BLUE n'existait pas dans la classe générale

$$\hat{Y} = \sum_{i \in s} d_i(s) y_i, \qquad (1)$$

même sous échantillonnage aléatoire simple. Ce résultat théorique négatif important a été, dans une grande mesure, négligé pendant une dizaine d'années. Godambe a également établi un résultat positif en liant y à une mesure de taille x au moyen d'un modèle de régression de superpopulation passant par l'origine avec variance d'erreur proportionnelle à x^2, puis en montrant que l'estimateur NHT sous un plan de sondage à taille fixe où π_i est proportionnel à x_i minimisait la variance prévue de la classe sans biais (1). Ce résultat montre clairement les conditions du plan pour l'utilisation de l'estimateur NHT. Rao (1966) a constaté les limites de l'estimateur NHT dans le contexte d'enquêtes avec échantillonnage PPT et caractéristiques multiples. Ici, l'estimateur NHT s'avère très inefficient lorsqu'une caractéristique y n'est pas liée (ou qu'elle est faiblement liée) à la mesure de taille x (comme le dénombrement de volailles y et la taille de la ferme x dans une enquête sur les fermes). Rao a proposé pour ces cas d'autres estimateurs efficients qui font abstraction des poids NHT. En faisant abstraction des résultats susmentionnés, des spécialistes de l'échantillonnage ont avancé plus tard certains critères théoriques pour affirmer qu'il fallait utiliser l'estimateur NHT pour tout plan de sondage. En prenant l'exemple amusant des éléphants d'un cirque, Basu (1971) a illustré la futilité de ces critères. Il a construit un « mauvais » plan dans lequel π_i n'était pas lié à y_i pour démontrer que l'estimateur NHT produisait des estimations absurdes, ce qui a incité le célèbre statisticien bayésien Dennis Lindley à conclure que ce contre-exemple détruisait la théorie des enquêtes par sondage fondées sur le plan de sondage (Lindley 1996). Cette conclusion est plutôt malheureuse, car NHT et Godambe ont clairement énoncé les conditions du plan pour une utilisation appropriée de l'estimateur NHT, et Rao (1966) et Hajek (1971) ont proposé d'autres estimateurs pour composer respectivement avec les caractéristiques multiples et les mauvais plans. Il est intéressant de noter que les mêmes critères théoriques ont abouti à un mauvais estimateur de variance de l'estimateur NHT comme choix « optimal » (Rao et Singh 1973).

On a aussi tenté d'intégrer la théorie des enquêtes par sondage à l'inférence statistique courante au moyen de la fonction de vraisemblance. Godambe (1966) a montré que la fonction de vraisemblance d'après les données-échantillons $\{(i, y_i),\ i \in s\}$, en considérant comme paramètre le vecteur N des valeurs y inconnues, ne fournissait pas d'information sur les valeurs non observées de l'échantillon ni, par conséquent, sur le total Y. Cette caractéristique non informative de la fonction de vraisemblance est due à la propriété d'étiquette qui traite les unités de population N essentiellement comme des post-strates N. On peut contourner cette difficulté en employant la méthode bayésienne et en supposant des valeurs antérieures informatives (échangeables) sur le vecteur paramètre (Ericson 1969). Une autre solution (fondée sur le plan de sondage) consiste à faire abstraction de certains aspects des données-échantillons pour rendre l'échantillon non unique et arriver ainsi à une fonction de vraisemblance informative (Hartley et Rao 1968; Royall 1968). Par exemple, sous échantillonnage aléatoire simple, en supprimant les étiquettes i et en considérant les données $\{(i, y_i),\ i \in s\}$ en l'absence d'information liant i à y_i, on obtient la moyenne d'échantillon comme estimateur du maximum de vraisemblance de la moyenne de population. En supposant des distributions antérieures non informatives, l'estimation bayésienne produit des résultats semblables à ceux obtenus par Ericson (1969) mais, contrairement à l'estimation d'Ericson, elle dépend du plan de sondage. Dans le cas où y_i est un vecteur qui comprend des variables auxiliaires avec totaux connus, Hartley et Rao (1968) ont montré que sous échantillonnage aléatoire simple, l'estimateur du maximum de vraisemblance était à peu près égal à l'estimateur par régression traditionnel du total. Cet article a été le premier à montrer comment intégrer des totaux de population auxiliaire connus à un cadre de vraisemblance. Pour l'échantillonnage aléatoire stratifié, on fait abstraction des étiquettes à l'intérieur des strates, mais pas des étiquettes de strate, à cause des différences connues entre les strates. L'estimateur du maximum de vraisemblance ainsi obtenu est à peu près égal à un pseudo-estimateur par régression linéaire optimal lorsqu'on dispose de variables auxiliaires avec totaux connus. Ce dernier estimateur possède de bonnes propriétés conditionnelles fondées sur le plan de sondage (voir la section 3.4). L'article de Hartley et Rao (1968) portait sur l'estimation d'un total, mais l'approche de la vraisemblance a une portée beaucoup plus vaste en échantillonnage, dont l'estimation de fonctions de distribution et de quantiles et la construction d'intervalles de confiance fondés sur des rapports de vraisemblance (voir la section 8.1). L'approche de la vraisemblance non paramétrique de Hartley-Rao a été découverte indépendamment vingt ans plus tard (Owen 1988) dans l'inférence statistique courante, sous le nom de « vraisemblance empirique », et a attiré passablement d'attention, notamment pour son application à divers problèmes d'échantillonnage. Dans un certain sens, les efforts d'intégration à la statistique courante ont donc partiellement réussi. L'ouvrage d'Owen (2002) présente une description complète de la théorie de la vraisemblance empirique et de ses applications.

3.2 Approche dépendante d'un modèle

En matière d'inférence, l'approche dépendante d'un modèle suppose que la structure de population obéit à un modèle de superpopulation spécifié. La distribution induite par le modèle hypothétique produit des inférences qui renvoient à l'échantillon donné d'unités s qui a été tiré. Ces inférences conditionnelles peuvent s'avérer plus pertinentes et plus attrayantes que les inférences établies par échantillonnage répété. Par contre, lorsque le modèle n'est pas spécifié correctement, les stratégies dépendantes d'un modèle peuvent donner de mauvais résultats dans le cas de grands échantillons; même de faibles écarts par rapport au modèle hypothétique, difficiles à déceler au moyen de méthodes de vérification de modèle, peuvent causer de graves problèmes. Par exemple, prenons le modèle par quotient souvent utilisé lorsqu'une variable auxiliaire x au total connu X est aussi mesurée dans l'échantillon :

$$y_i = \beta x_i + \varepsilon_i \,;\; i = 1, \ldots, N \qquad (2)$$

où les ε_i sont des variables aléatoires indépendantes avec moyenne nulle et variance proportionnelle à x_i. En supposant que le modèle soit valable pour l'échantillon, c'est-à-dire sans biais d'échantillonnage, le meilleur prédicteur sans biais par rapport au modèle linéaire du total Y est donné par l'estimateur par quotient $(\bar{y}/\bar{x})X$ sans égard au plan de sondage. Cet estimateur n'est pas convergent selon le plan de sondage, sauf si le plan est autopondéré, par exemple sous échantillonnage aléatoire stratifié avec répartition proportionnelle. Par conséquent, sous des plans non autopondérés, il peut donner de très mauvais résultats dans le cas de grands échantillons, même si les écarts par rapport au modèle sont faibles. Hansen et coll. (1983) ont démontré les mauvais résultats obtenus dans des conditions d'échantillonnage répété, en utilisant un plan d'échantillonnage aléatoire stratifié avec une répartition de l'échantillon presque optimale (couramment utilisé en présence de populations très asymétriques). Rao (1996) a utilisé le même plan pour démontrer les mauvais résultats obtenus dans le contexte d'un cadre conditionnel pertinent à l'approche dépendante d'un modèle (Royall et Cumberland 1981). Néanmoins, les approches dépendantes d'un modèle peuvent jouer un rôle capital dans l'estimation sur petits domaines où la taille de l'échantillon dans un petit domaine peut être infime, voire nulle (voir la section 7).

Brewer (1963) a été le premier à proposer l'approche dépendante d'un modèle dans le contexte du modèle par quotient (2). Royall (1970) et ses collaborateurs ont mené une étude systématique de cette approche. Valliant, Dorfman et Royall (2000) donnent une description complète de la théorie, dont l'estimation de la variance (conditionnelle) par rapport au modèle de l'estimateur qui varie avec s; par exemple, sous le modèle par quotient (2), la variance par rapport au modèle dépend de la moyenne d'échantillon $\bar{x}_s$. Il est intéressant de noter que l'échantillonnage équilibré au moyen de la sélection raisonnée figure dans l'approche dépendante d'un modèle dans le contexte de la protection contre la spécification incorrecte du modèle (Royall et Herson 1973).

3.3 Approche assistée par un modèle

L'approche assistée par un modèle cherche à combiner les caractéristiques positives de la méthode fondée sur le plan de sondage et de la méthode dépendante d'un modèle. Elle considère uniquement les estimateurs convergents selon le plan de sondage du total Y qui sont aussi sans biais par rapport au modèle sous le modèle « de travail » hypothétique. Par exemple, sous le modèle par quotient (2), un estimateur assisté par un modèle de Y pour un plan d'échantillonnage probabiliste spécifié est donné par l'estimateur par quotient $\hat{Y}_r = (\hat{Y}_{\mathrm{NHT}} / \hat{X}_{\mathrm{NHT}})X$ qui est convergent selon le plan de sondage sans égard au modèle hypothétique. Hansen et coll. (1983) ont utilisé cet estimateur dans leur plan d'échantillonnage stratifié pour démontrer que ses résultats étaient supérieurs à ceux de l'estimateur dépendant d'un modèle $(\bar{y}/\bar{x})X$. Pour l'estimation de la variance, l'approche assistée par un modèle utilise des estimateurs convergents pour la variance de l'estimateur par rapport au plan tout en étant exactement ou asymptotiquement sans biais par rapport au modèle pour la variance par rapport au modèle. Toutefois, les inférences sont fondées sur le plan de sondage, car le modèle est utilisé uniquement comme modèle « de travail ».

Pour l'estimateur par quotient $\hat{Y}_r$ l'estimateur de variance est donné par

$$\mathrm{Var}(\hat{Y}_r) = (X / \hat{X}_{\mathrm{NHT}})^2 \, v(e), \qquad (3)$$

où, dans la notation d'opérateur, $v(e)$ est obtenu à partir de $v(y)$ en remplaçant y_i par les résidus $e_i = y_i - (\hat{Y}_{\mathrm{NHT}} / \hat{X}_{\mathrm{NHT}})x_i$. Cet estimateur de variance est asymptotiquement équivalent à un estimateur de linéarisation courant de la variance $v(e)$, mais il reflète le fait que l'information contenue dans l'échantillon varie avec $\hat{X}_{\mathrm{NHT}}$: les valeurs élevées produisent une faible variabilité, et les valeurs faibles, une grande variabilité. Le pivot normal ainsi obtenu produit des inférences dépendantes d'un modèle qui sont valides sous le modèle hypothétique (contrairement à l'utilisation de $v(e)$ dans le pivot) tout en protégeant contre les écarts par rapport au modèle, en ce sens qu'il produit des inférences asymptotiquement valides fondées sur le plan de sondage. Il est à noter que le pivot est asymptotiquement équivalent à $\hat{Y}(\tilde{e})/[v(\tilde{e})]^{1/2}$ avec $\tilde{e}_i = y_i - (Y/X)x_i$. Si les écarts par rapport au modèle sont faibles, l'asymétrie dans les résidus $\tilde{e}_i$ est faible même si y_i et x_i sont très asymétriques, et les intervalles de confiance normaux

donnent de bons résultats. Par contre, pour des populations très asymétriques, les intervalles normaux fondés sur $\hat{Y}_{\text{NHT}}$ et son erreur-type peuvent donner de mauvais résultats sous échantillonnage répété, même pour des échantillons assez grands, car le pivot dépend de l'asymétrie des y_i. La structure de population joue donc un rôle dans les inférences fondées sur le plan de sondage, contrairement à ce qu'affirment Neyman (1934), Hansen et coll. (1983) et d'autres auteurs. Rao, Jocelyn et Hidiroglou (2003) ont considéré l'estimateur par régression linéaire simple sous échantillonnage aléatoire simple à deux phases avec seulement x observé dans la première phase. Ils ont démontré que le rendement de couverture des intervalles normaux associés pouvait être faible même pour des échantillons de deuxième phase passablement grands si le vrai modèle sous-jacent qui produisait la population s'écartait considérablement du modèle de régression linéaire (par exemple, une régression quadratique de y sur x) et si l'asymétrie de x est grande. Dans ce cas, les valeurs x de la première phase sont observées et une approche assistée par un modèle approprié utiliserait un estimateur par régression linéaire multiple avec x et $z = x^2$ comme variables auxiliaires. Il est à noter que pour l'échantillonnage à une seule phase, on ne peut mettre en œuvre un tel estimateur assisté par un modèle si l'on connaît uniquement les X, puisque l'estimateur dépend du total de population de z.

Särndal, Swenson et Wretman (1992) proposent une description complète de l'approche assistée par un modèle pour estimer le total Y d'une variable y sous le modèle de régression linéaire de travail

$$y_i = x_i'\beta + \varepsilon_i;\ i = 1, ..., N \qquad (4)$$

avec moyenne nulle, erreurs non corrélées ε_i et variance par rapport au modèle $V_m(\varepsilon_i) = \sigma^2 q_i = \sigma_i^2$ où les q_i sont des constantes connues et les vecteurs x ont des totaux connus X (les valeurs de population $x_1, ..., x_N$ ne sont pas nécessairement connues). Dans ces conditions, l'approche assistée par un modèle produit l'estimateur de régression généralisée (generalized regression ou GREG)

$$\hat{Y}_{gr} = \hat{Y}_{\text{NHT}} + \hat{B}'(X - \hat{X}_{\text{NHT}}) =: \sum_{i \in s} w_i(s) y_i, \qquad (5)$$

où

$$\hat{B} = \hat{T}^{-1}\left(\sum_s \pi_i^{-1} x_i y_i / q_i\right) \qquad (6)$$

avec $\hat{T} = \sum_s \pi_i^{-1} x_i x_i' / q_i$ est un coefficient de régression pondéré et $w_i(s) = g_i(s)\pi_i^{-1}$ avec $g_i(s) = 1 + (X - \hat{X}_{\text{NHT}})'\hat{T}^{-1} x_i / q_i$, appelé « poids g ». Il est à noter que l'estimateur GREG (5) peut également s'écrire $\sum_{i \in U} \hat{y}_i + \hat{E}_{\text{NHT}}$, où $\hat{y}_i = x_i'\hat{B}$ est le prédicteur de y_i sous le modèle de travail et $\hat{E}_{\text{NHT}}$ est l'estimateur NHT de l'erreur de prévision totale $E = \sum_{i \in U} e_i$ avec $e_i = y_i - \hat{y}_i$. Cette représentation montre le rôle du modèle de travail dans l'approche assistée par un modèle. L'estimateur GREG (5) est cohérent avec le plan de sondage et sans biais par rapport au modèle sous le modèle de travail (4). En outre, il est presque « optimal » en ce sens qu'il minimise l'erreur quadratique moyenne prévue asymptotique (espérance du modèle de l'erreur quadratique moyenne par rapport au plan) sous le modèle de travail, à condition que la probabilité d'inclusion, π_i, soit proportionnelle à l'écart-type par rapport au modèle σ_i. Toutefois, dans les enquêtes à plusieurs variables d'intérêt, la variance par rapport au modèle peut varier selon les variables. Comme on doit utiliser un plan de sondage général, tel que le plan avec probabilités d'inclussion proportionnelles à la taille, le résultat d'optimalité n'est plus valable, même si le même vecteur x_i est utilisé pour toutes les variables y_i du modèle de travail.

L'estimateur GREG devient simplement l'estimateur « par projection » $X'\hat{B} = \sum_s w_i(s) y_i$ avec $g_i(s) = X'\hat{T}^{-1} x_i / q_i$ si la variance par rapport au modèle σ_i^2 est proportionnelle à $\lambda' x_i$ pour certains λ. On obtient l'estimateur par quotient sous forme de cas particulier de l'estimateur par projection, étant donné $q_i = x_i$, d'où $g_i(s) = X / \hat{X}_{\text{HT}}$. Il est à noter que l'estimateur GREG (5) exige uniquement les totaux de population X, et pas nécessairement les valeurs de population individuelles x_i. Cette caractéristique est très utile, car les totaux de population auxiliaire sont souvent attestés à l'aide de sources externes comme les projections démographiques des dénombrements selon l'âge et le sexe. De plus, il assure la cohérence avec les totaux connus X en ce sens que $\sum_s w_i(s) x_i = X$. En raison de cette propriété, l'estimateur GREG est également un estimateur par calage.

Supposons qu'il y ait p variables d'intérêt, par exemple $y^{(1)}, ..., y^{(p)}$, et qu'on veuille utiliser l'approche assistée par un modèle pour estimer les totaux de population correspondants $Y^{(1)}, ..., Y^{(p)}$. Supposons également que le modèle de travail de $y^{(j)}$ prenne la forme (4) mais nécessite un vecteur x peut-être différent $x^{(j)}$ avec total connu $X^{(j)}$ pour chaque $j = 1, ..., p$:

$$y_i^{(j)} = x_i^{(j)'}\beta^{(j)} + \varepsilon_i^{(j)},\ i = 1, ..., N. \qquad (7)$$

Dans ce cas, les poids g dépendent de j et, à leur tour, les poids finaux $w_i(s)$ dépendent aussi de j. Dans la pratique, il est souvent souhaitable d'utiliser un seul ensemble de poids finaux pour toutes les variables p afin d'assurer la cohérence interne des chiffres lorsqu'ils sont agrégés à partir de variables différentes. On ne peut réaliser cette propriété qu'en élargissant le vecteur x dans le modèle (7) pour recevoir toutes les variables $y^{(j)}$, par exemple $\tilde{x}$ avec total connu $\tilde{X}$, puis en utilisant le modèle de travail

$$y_i^{(j)} = \tilde{x}_i' \beta^{(j)} + \varepsilon_i^{(j)},\ i = 1, \ldots, N. \qquad (8)$$

Toutefois, les coefficients de régression pondérés ainsi obtenus pourraient devenir instables à cause du risque de multicolinéarité dans l'ensemble élargi de variables auxiliaires. Par conséquent, l'estimateur GREG de $Y^{(j)}$ sous le modèle (8) est moins efficient que l'estimateur GREG sous le modèle (7). En outre, certains poids finaux ainsi obtenus, par exemple $\tilde{w}_i(s)$, risquent de ne pas satisfaire les restrictions relatives à l'étendue en prenant des valeurs inférieures à 1 (dont des valeurs négatives) ou de très grandes valeurs positives. Il est possible de résoudre ce problème en utilisant un estimateur par régression ridge généralisée de $Y^{(j)}$ qui est assisté par un modèle sous le modèle élargi (Chambers 1996; Rao et Singh 1997).

Pour l'estimation de la variance, l'approche assistée par un modèle cherche à utiliser des estimateurs de variance convergents selon le plan de sondage qui sont aussi sans biais par rapport au modèle (du moins pour les grands échantillons) en ce qui concerne la variance conditionnelle par rapport au modèle de l'estimateur GREG. Dénotant l'estimateur de variance de l'estimateur NHT de Y par $v(y)$ dans une notation d'opérateur, un estimateur de variance par linéarisation de Taylor simple satisfaisant la propriété susmentionnée est donné par $v(ge)$, où l'on obtient $v(ge)$ en remplaçant y_i par $g_i(s)e_i$ dans la formule de $v(y)$; voir Hidiroglou, Fuller et Hickman (1976) et Särndal, Swenson et Wretman (1989).

Dans l'exposé qui précède, nous avons supposé un modèle de régression linéaire de travail pour toutes les variables $y^{(j)}$. Dans la pratique, cependant, un modèle de régression linéaire n'est pas nécessairement bien adapté à certaines variables d'intérêt y, par exemple, une variable binaire. Dans ce dernier cas, la régression logistique offre un modèle de travail approprié. Un modèle de travail général qui couvre la régression logistique prend la forme $E_m(y_i) = h(x_i'\beta) = \mu_i$, où $h(.)$ pourrait être non linéaire; le modèle (5) est un cas particulier avec $h(a) = a$. Un estimateur assisté par un modèle du total sous le modèle de travail général est l'estimateur par la différence $\hat{Y}_{\text{NHT}} + \sum_U \hat{\mu}_i - \sum_s \pi_i^{-1} \hat{\mu}_i$, où $\hat{\mu}_i = h(x_i'\hat{\beta})$ et $\hat{\beta}$ est un estimateur du paramètre de modélisation β. Il se réduit à l'estimateur GREG (5) si $h(a) = a$. Cet estimateur par la différence est presque optimal si la probabilité d'inclusion π_i est proportionnelle à σ_i, où σ_i^2 dénote la variance par rapport au modèle, $V_m(y_i)$.

Les estimateurs GREG sont très appréciés par les utilisateurs parce que bon nombre d'estimateurs couramment utilisés peuvent être obtenus sous forme de cas particuliers de (5) par des spécifications appropriées de x_i et q_i. Statistique Canada a mis au point un Système généralisé d'estimation (SGE) fondé sur l'estimateur GREG.

Kott (2005) a proposé un autre paradigme de l'inférence, appelé approche fondée sur un modèle et assistée par randomisation, qui est axé sur l'inférence fondée sur un modèle et assistée par randomisation (ou échantillonnage répété). La définition de la variance prévue est inversée pour devenir la variance prévue par rapport au modèle à randomisation d'un estimateur, mais elle est identique à la variance prévue habituelle lorsque le modèle de travail est valable pour l'échantillon, comme on le suppose dans l'article. Par conséquent, les choix de l'estimateur et de l'estimateur de variance sont souvent semblables à ceux qui sont faits sous l'approche assistée par un modèle. Toutefois, Kott soutient que la motivation est plus claire et que « l'approche proposée ici pour l'estimation de la variance mène, au besoin, à un traitement logiquement cohérent de rajustements d'une population finie et d'un petit échantillon ».

3.4 Approche conditionnelle fondée sur le plan de sondage

On a également proposé une approche conditionnelle fondée sur le plan de sondage. Cette approche cherche à combiner les caractéristiques conditionnelles de l'approche dépendante d'un modèle avec les caractéristiques indépendantes de l'approche fondée sur le plan de sondage. Elle permet de restreindre l'ensemble d'échantillons de référence à un sous-ensemble « pertinent » de tous les échantillons possibles spécifiés par le plan de sondage. On obtient des inférences conditionnellement valides en ce sens que le ratio de biais conditionnel (soit le ratio du biais conditionnel à l'erreur-type conditionnelle) devient nul à mesure que la taille de l'échantillon augmente. Environ $100(1-\alpha)\%$ des intervalles de confiance réalisés dans l'échantillonnage répété à partir de l'ensemble conditionnel contiennent le total inconnu Y.

Holt et Smith (1979) fournissent des arguments convaincants en faveur de l'inférence conditionnelle fondée sur le plan, même si leur analyse est limitée à la post-stratification simple d'un échantillon aléatoire simple, auquel cas il est naturel de faire des inférences conditionnelles à la taille des strates de l'échantillon réalisé. Rao (1992, 1994) et Casady et Valliant (1993) ont étudié l'inférence conditionnelle lorsque seul le total auxiliaire X est connu d'après des sources externes. Dans ce dernier cas, la subordination à l'estimateur NHT $\hat{X}_{\text{NHT}}$ peut s'avérer raisonnable parce qu'il s'agit « à peu près » d'une statistique auxiliaire lorsque X est connu et que la différence $\hat{X}_{\text{NHT}} - X$ fournit une mesure du déséquilibre de l'échantillon réalisé. La subordination à $\hat{X}_{\text{NHT}}$ permet de calculer l'estimateur par régression linéaire « optimal », de même forme que l'estimateur GREG (5), dans lequel $\hat{B}$ donné par (6) est remplacé par la valeur optimale estimative $\hat{B}_{\text{opt}}$ du coefficient

de régression qui fait intervenir la covariance estimative de $\hat{Y}_{\text{NHT}}$ et $\hat{X}_{\text{NHT}}$ et la variance estimative de $\hat{X}_{\text{NHT}}$. Cet estimateur optimal permet d'établir des inférences conditionnellement valides fondées sur le plan de sondage et est sans biais par rapport au modèle sous le modèle de travail (4). Il s'agit également d'un estimateur par calage dépendant uniquement du total X et il peut être exprimé comme suit : $\sum_{i\in s}\tilde{w}_i(s)y_i$ avec poids $\tilde{w}_i(s) = d_i\tilde{g}_i(s)$ et le facteur de calage $\tilde{g}_i(s)$ dépendant uniquement du total X et les valeurs x de l'échantillon. Il fonctionne bien pour l'échantillonnage aléatoire stratifié (couramment utilisé dans les enquêtes-établissements). Toutefois, $\hat{B}_{\text{opt}}$ peut devenir instable dans le cas de l'échantillonnage stratifié à plusieurs degrés, sauf si la différence entre le nombre de grappes d'échantillon et le nombre de strates est passablement élevée. L'estimateur GREG n'exige pas cette dernière condition, mais il peut donner de mauvais résultats en ce qui concerne le ratio de biais conditionnel et les taux de couverture conditionnels, comme l'a montré Rao (1996). L'estimateur NHT sans biais peut être conditionnellement très mauvais, sauf si le plan assure que la mesure du déséquilibre définie plus haut est faible. Par exemple, dans le plan de sondage fondé sur la stratification x efficiente et proposé par Hansen et coll. (1983), le déséquilibre est faible et l'estimateur NHT a donné conditionnellement de bons résultats.

Tillé (1998) a proposé un estimateur NHT du total Y fondé sur des probabilités d'inclusion conditionnelles approximatives en présence de $\hat{X}_{\text{NHT}}$. Sa méthode permet également d'établir des inférences conditionnellement valides, mais l'estimateur n'est pas calé en fonction de X, contrairement à l'estimateur par régression linéaire « optimal ». Park et Fuller (2005) ont proposé une version calée de l'estimateur GREG fondée sur l'estimateur de Tillé qui donne des poids non négatifs plus souvent que l'estimateur GREG.

Je crois que les praticiens devraient accorder une plus grande attention aux aspects conditionnels de l'inférence fondée sur le plan de sondage et envisager sérieusement les nouvelles méthodes qui ont été proposées.

Kalton (2002) a donné des arguments convaincants pour favoriser des approches fondées sur le plan de sondage (et peut-être conditionnelles ou assistées par un modèle) de l'inférence en fonction des paramètres descriptifs d'une population finie. Smith (1994) a nommé « inférence procédurale » l'inférence fondée sur le plan de sondage et a soutenu qu'il s'agissait de l'approche à adopter pour les enquêtes du domaine public. Le lecteur trouvera dans Smith (1976) et Rao et Bellhouse (1990) des études des questions d'inférence dans la théorie des enquêtes par sondage.

4. Estimateurs par calage

On obtient les poids de calage $w_i(s)$ qui assurent la cohérence avec les totaux auxiliaires X spécifiés par l'utilisateur en rajustant les poids de sondage $d_i = \pi_i^{-1}$ pour satisfaire les contraintes d'étalonnage $\sum_{i\in s} w_i(s)x_i = X$. Les estimateurs qui utilisent des poids de calage sont appelés estimateurs par calage et utilisent un seul ensemble de poids $\{w_i(s)\}$ pour toutes les variables d'intérêt. Nous avons mentionné dans la section 3.4 que l'estimateur GREG assisté par un modèle était un estimateur par calage, mais un estimateur par calage n'est pas nécessairement assisté par un modèle, en ce sens qu'il risque d'être biaisé par rapport au modèle sous un modèle de travail (4), sauf si les variables x du modèle coïncident exactement avec les variables correspondant aux totaux spécifiés par l'utilisateur. Par exemple, supposons que le modèle de travail suggéré par les données soit un modèle quadratique dans une variable scalaire x alors que le total spécifié par l'utilisateur est uniquement son total X. L'estimateur par calage ainsi obtenu peut donner de mauvais résultats même dans des échantillons assez grands, comme nous l'avons mentionné dans la section 3.3, contrairement à l'estimateur GREG assisté par un modèle fondé sur le modèle quadratique de travail qui nécessite le total de population des variables quadratiques x_i^2 en plus de X.

Dans la pratique, on utilise abondamment la post-stratification pour assurer la cohérence avec les valeurs connues de la cellule correspondant à une variable de post-stratification, par exemple des valeurs dans différents groupes d'âge vérifiées d'après des sources externes comme des projections démographiques. L'estimateur post-stratifié ainsi obtenu est un estimateur par calage. On a également utilisé dans la pratique des estimateurs par calage qui assurent la cohérence avec les valeurs marginales connues de deux ou plusieurs variables de post-stratification, notamment les estimateurs de la méthode itérative du quotient, qu'on obtient par étalonnage répété des valeurs marginales jusqu'à ce que la convergence soit approximativement réalisée, habituellement en quatre itérations ou moins. Les poids obtenus par la méthode itérative du quotient $w_i(s)$ sont toujours positifs. Dans le cadre du Recensement du Canada, Statistique Canada a déjà utilisé les estimateurs de la méthode itérative du quotient pour assurer la cohérence des estimateurs de données-échantillon (2B) avec les valeurs connues des données intégrales (2A). Toujours dans le contexte du Recensement du Canada, Brackstone et Rao (1979) ont étudié l'efficience des estimateurs de la méthode itérative du quotient et ont aussi calculé des estimateurs de variance par linéarisation de Taylor lorsque le nombre d'itérations était de quatre ou moins. On a également employé les estimateurs de la méthode itérative du quotient dans la Current

Population Survey (CPS) des États-Unis. Il convient de noter que la méthode de rajustement des valeurs de la cellule en fonction des valeurs marginales données dans un tableau à double entrée a d'abord été proposée dans l'article marquant de Deming et Stephan (1940).

Des approches unifiées du calage, fondées sur la minimisation d'une mesure appropriée de la distance entre les poids de calage et les poids de sondage sous réserve des contraintes d'étalonnage, ont attiré l'attention des utilisateurs en raison de leur capacité de recevoir un nombre arbitraire de contraintes d'étalonnage spécifiées par l'utilisateur, par exemple, le calage en fonction des valeurs marginales de plusieurs variables de post-stratification. Des logiciels de calage sont également disponibles, dont le SGE (Statistique Canada), LIN WEIGHT (Bureau national de la statistique des Pays-Bas), CALMAR (INSEE, France) et CLAN97 (Bureau de la statistique de Suède).

Une distance de chi carré, $\sum_{i\in s} q_i (d_i - w_i)^2 / d_i$, permet de calculer l'estimateur GREG (5), où le vecteur x – correspond aux contraintes d'étalonnage spécifiées par l'utilisateur et $w_i(s)$ est dénoté w_i par souci de simplicité (Huang et Fuller 1978; Deville et Särndal 1992). Toutefois, les poids de calage ainsi obtenus ne satisfont pas nécessairement les restrictions relatives à l'étendue souhaitable; par exemple, certains poids peuvent être négatifs ou trop grands, surtout lorsque le nombre de contraintes est élevé et que la variabilité des poids de sondage est élevée. Huang et Fuller (1978) ont proposé une mesure de distance de chi carré modifiée à l'échelle et obtenu les poids de calage au moyen d'une solution itérative qui satisfait les contraintes d'étalonnage à chaque itération. Toutefois, il n'existe peut-être pas de solution qui satisfait à la fois les contraintes d'étalonnage et les contraintes relatives à l'étendue. Une autre méthode, appelée minimisation par rétrécissement (Singh et Mohl 1996), se heurte à la même difficulté. On a également proposé des méthodes de programmation quadratique qui minimisent la distance de chi carré sous réserve des contraintes d'étalonnage et des contraintes relatives à l'étendue (Hussain 1969), mais l'ensemble de solutions réalisables satisfaisant les deux types de contrainte peut être vide. D'autres méthodes proposées consistent à modifier la fonction de distance (Deville et Särndal 1992) ou à abandonner certaines contraintes d'étalonnage (Bankier, Rathwell et Majkowski 1992). Par exemple, une distance d'information de forme $\sum_{i\in s} q_i \{ w_i \log(w_i / d_i) - w_i + d_i \}$ donne des estimateurs de la méthode itérative du quotient avec poids non négatifs w_i, mais certains poids peuvent être beaucoup trop grands. On a également proposé des poids « ridge » obtenus en minimisant une distance de chi carré pénalisée (Chambers 1996), mais rien ne garantit qu'ils satisfont les contraintes d'étalonnage ou les contraintes relatives à l'étendue, quoique les poids soient plus stables que les poids GREG. Rao et Singh (1997) ont proposé une méthode itérative de « rétrécissement ridge » qui assure la convergence pour un nombre spécifié d'itérations en utilisant une spécification de tolérance intégrée pour assouplir certaines contraintes d'étalonnage tout en satisfaisant les contraintes relatives à l'étendue. Chen, Sitter et Wu (2002) ont proposé une méthode semblable.

On a utilisé les poids de calage GREG dans l'Enquête sur la population active du Canada qui, tout récemment, a fait appel à des estimateurs composites qui utilisent l'information des mois antérieurs sur l'échantillon, comme nous l'avons mentionné dans la section 2 (Fuller et Rao 2001; Gambino, Kennedy et Singh 2001; Singh, Kennedy et Wu 2001). On a également utilisé des estimateurs par calage de type GREG pour intégrer deux ou plusieurs enquêtes indépendantes portant sur la même population. Ces estimateurs assurent la cohérence entre les enquêtes, en ce sens que les estimateurs de variables communes aux deux enquêtes sont identiques, ainsi que l'étalonnage en fonction de totaux de population connus (Renssen et Nieuwenbroek 1997; Singh et Wu 1996; Merkouris 2004). Pour le Recensement du Canada de 2001, Bankier (2003) a étudié des poids de calage correspondant à l'estimateur par régression linéaire « optimal » (section 3.3) sous échantillonnage aléatoire stratifié. Il a montré que la méthode de calage « optimale » donnait de meilleurs résultats que l'estimateur par calage GREG, utilisé lors du recensement précédent, dans la mesure où elle permettait de conserver plus de contraintes d'étalonnage tout en permettant aux poids de calage d'être au moins un. On peut obtenir le poids de calage « optimal » à l'aide du logiciel SGE en précisant dans les contraintes d'étalonnage la taille connue des strates et en définissant comme il se doit la constante de réglage q_i. Il est à noter que l'estimateur par calage « optimal » possède également des propriétés conditionnelles souhaitables par rapport au plan (section 3.4). Pour la pondération des données du Recensement du Canada de 2001, la méthode de la régression linéaire « optimale » a remplacé celle de l'estimateur GREG par projection (utilisée lors du Recensement de 1996).

Demnati et Rao (2004) ont calculé des estimateurs de variance par linéarisation de Taylor pour une classe générale d'estimateurs par calage avec poids $w_i = d_i F(x_i' \hat{\lambda})$, où l'on détermine le multiplicateur de LaGrange $\hat{\lambda}$ en résolvant les contraintes de calage. Le choix $F(a) = 1 + a$ donne des poids GREG et $F(a) = e^a$ permet de calculer des poids obtenus par la méthode itérative du quotient. Dans le cas particulier des poids GREG, l'estimateur de variance se réduit à $v(ge)$ donné dans la section 3.3.

Le lecteur trouvera dans l'article de Fuller (2002), récipiendaire du prix Waksberg, un aperçu et une évaluation très éloquents de l'estimation par régression dans l'échantillonnage d'enquête, y compris l'estimation par calage.

5. Échantillonnage avec probabilités inégales sans remise

Nous avons mentionné dans la section 2 que l'échantillonnage PPT d'UPÉ à l'intérieur de strates dans les enquêtes à grande échelle était motivé par des considérations pratiques, soit la volonté de répartir des charges de travail à peu près égales. L'échantillonnage PPT permet également de réduire considérablement la variance en neutralisant la variabilité découlant de la taille inégale des UPÉ sans vraiment stratifier par taille. Les UPÉ sont habituellement échantillonnées sans remise, de manière que la probabilité d'inclusion des UPÉ, π_i, soit proportionnelle à la mesure de la taille des UPÉ x_i. Par exemple, l'échantillonnage PPT systématique, avec ou sans randomisation initiale des étiquettes UPÉ, est un plan avec probabilité d'inclusion proportionnelle à la taille (PIPT) (appelé aussi plan πPT) utilisé dans un grand nombre d'enquêtes complexes, dont l'EPA du Canada. L'estimateur d'un total associé à un plan PIPT est l'estimateur NHT.

L'élaboration de stratégies appropriées (PIPT, NHT) soulève des problèmes sur le plan théorique, dont l'évaluation de probabilités d'inclusion conjointes exactes, π_{ij}, ou des approximations exactes de π_{ij} nécessitant uniquement les π_i individuels, qui sont nécessaires pour obtenir un estimateur de variance sans biais ou presque sans biais. J'ai étudié ce dernier problème dans la thèse de doctorat que j'ai présentée en 1961 à la Iowa State University. D'éminents statisticiens-mathématiciens ont publié depuis plusieurs solutions nécessitant des outils théoriques perfectionnés. Toutefois, ces travaux théoriques sont souvent qualifiés de « théorie sans application » puisque, dans la pratique, il est courant de traiter les UPÉ comme si elles étaient échantillonnées avec remise, d'où une grande simplification. L'estimateur de variance est obtenu simplement à partir des totaux estimatifs d'UPÉ; cette hypothèse est d'ailleurs à la base des méthodes de rééchantillonnage (section 6). Cet estimateur de variance peut entraîner une surestimation substantielle, sauf si la fraction d'échantillonnage des UPÉ globales est faible, ce qui peut être vrai dans bon nombre d'enquêtes à grande échelle. Dans les paragraphes qui suivent, je tenterai de démontrer que les travaux théoriques portant sur certaines stratégies (PIPT, NHT) et sur des plans de sondage sans PIPT ont une grande applicabilité dans la pratique.

J'aborderai d'abord certaines stratégies (PIPT, NHT). En Suède et dans d'autres pays européens, on utilise souvent l'échantillonnage stratifié à un seul degré en raison de la disponibilité de listes et les plans PIPT sont des options attrayantes, mais les fractions d'échantillonnage sont souvent grandes. Par exemple, Rosén (1991) mentionne que le baromètre de la population active du Bureau de la statistique de Suède échantillonne une centaine de populations différentes en utilisant l'échantillonnage PPT systématique et que les taux d'échantillonnage peuvent dépasser 50 %. Aires et Rosén (2005) ont étudié l'échantillonnage πPT de Pareto pour les enquêtes suédoises. Cette méthode possède des propriétés attrayantes, dont la taille fixe de l'échantillon, l'échantillonnage simple, une bonne précision d'estimation, et une estimation convergente de la variance sans égard aux taux d'échantillonnage. En outre, elle permet de coordonner les échantillons au moyen de nombres aléatoires permanents (NAP), comme dans l'échantillonnage de Poisson, mais cette dernière méthode produit des échantillons de taille variable. En raison de ces mérites, on a mis en œuvre l'échantillonnage πPT de Pareto dans un certain nombre d'enquêtes du Bureau de la statistique de Suède, notamment dans les enquêtes sur l'indice des prix. Ohlsson (1995) a décrit les techniques des NAP qui sont couramment utilisées dans la pratique.

La méthode de Rao-Sampford (voir Brewer et Hanif 1983, page 28) produit des plans PIPT exacts et des estimateurs de variance non négatifs sans biais pour des échantillons de taille fixe arbitraire. Elle a été mise en œuvre dans la nouvelle version du SAS. Stehman et Overton (1994) notent que la structure de la probabilité variable se manifeste naturellement dans les enquêtes environnementales au lieu d'être sélectionnée uniquement pour l'efficience accrue et que les π_i sont connus uniquement pour les unités i de l'échantillon s. En traitant le plan de sondage selon la méthode d'échantillonnage systématique aléatoire avec PPT, Stehman et Overton ont obtenu des approximations des π_{ij} qui dépendent uniquement des π_i, $i \in s$, contrairement aux approximations initiales de Hartley et Rao (1962) qui nécessitent la somme des carrés de tous les π_i de la population. Dans les applications de Stehman et Overton, les taux d'échantillonnage sont assez substantiels pour justifier l'évaluation des probabilités d'inclusion conjointes.

Je vais maintenant aborder les plans sans PIPT utilisant des estimateurs différents de l'estimateur NHT qui assure une variance nulle lorsque y est exactement proportionnel à x. La méthode des groupes aléatoires de Rao, Hartley et Cochran (1962) permet de calculer un estimateur de variance non négatif simple pour n'importe quelle taille fixe de l'échantillon; pourtant, elle se compare favorablement aux stratégies (PIPT, NHT) sur le plan de l'efficience et elle est toujours plus efficiente que la stratégie PPT avec remise. Schabenberger et Grégoire (1994) ont constaté que les stratégies (PIPT, NHT) n'avaient pas trouvé beaucoup d'applications en foresterie à cause de la difficulté de mise en œuvre et ont recommandé la stratégie de Rao-Hartley-Cochran en raison de sa remarquable simplicité et de son efficience. Il est intéressant de constater que cette stratégie a été utilisée dans l'EPA du Canada parce qu'elle permettait

d'adopter de nouvelles mesures de la taille en utilisant la méthode de Keyfitz à l'intérieur de chaque groupe aléatoire. Par contre, les stratégies (PIPT, NHT) ne conviennent pas tellement à cette fin (Fellegi 1966). Je crois savoir qu'on utilise souvent la stratégie de Rao-Hartley-Cochran en contrôle par sondage et dans d'autres applications comptables.

Murthy (1957) a utilisé un plan sans PIPT fondé sur le tirage d'unités successives avec probabilités p_i, $p_j/(1-p_i)$, $p_k/(1-p_i-p_j)$ et ainsi de suite, et l'estimateur suivant :

$$\hat{Y}_M = \sum_{i \in s} y_i \frac{p(s|i)}{p(s)}, \qquad (9)$$

où $p(s|i)$ est la probabilité conditionnelle d'obtenir l'échantillon s lorsque l'unité i a été sélectionnée en premier. Il a également proposé un estimateur de variance non négatif nécessitant les probabilités conditionnelles, $p(s|i, j)$, d'obtenir s lorsque i et j sont sélectionnés dans les deux premiers tirages. Pendant plusieurs années, les praticiens ont accordé peu d'attention à cette méthode à cause de la complexité des calculs mais, plus récemment, on l'a appliquée dans des domaines inattendus, dont la découverte de pétrole (Andreatta et Kaufmann 1986) et l'échantillonnage séquentiel, dont l'échantillonnage inverse et certains schémas d'échantillonnage adaptable (Salehi et Seber 1997). Il convient de noter qu'au cours des dernières années, on s'est beaucoup intéressé à l'échantillonnage adaptable puisqu'il s'agit d'une méthode d'échantillonnage efficiente pour estimer des totaux ou des moyennes de populations rares (Thompson et Seber 1996). Dans son application à la découverte de pétrole, le schéma d'échantillonnage successif est une caractérisation de la découverte et l'ordre dans lequel les champs de pétrole sont découverts est déterminé par l'échantillonnage proportionnel à la taille des champs et sans remise, selon un vieux principe de l'industrie : « en moyenne, on trouve d'abord les grands champs ». Ici, $p_i = y_i / Y$ et la réserve de pétrole totale Y est présumée connue d'après des critères géologiques. Dans cette application, les géologues s'intéressent à la distribution par taille de tous les champs du bassin et, après l'exploration partielle d'un bassin, l'échantillon est composé de grandeurs y_i de dépôts découverts. On peut estimer la fonction de distribution par taille $F(a)$ en utilisant l'estimateur de Murthy (9) dans lequel y_i est remplacé par la variable indicatrice $I(y_i \le a)$. Le calcul de $p(s|i)$ et $p(s)$, toutefois, est très complexe, même pour des échantillons de taille moyenne. Afin de surmonter cette difficulté de calcul, Andreatta et Kaufman (1986) ont utilisé des représentations intégrales de ces quantités pour formuler des expressions asymptotiques de l'estimateur de Murthy, dont les premiers termes sont aisés à calculer. De même, ils obtiennent des approximations calculables de l'estimateur de variance de Murthy. Il est à noter qu'on ne peut employer ici l'estimateur NHT de $F(a)$, car les probabilités d'inclusion sont des fonctions de toutes les valeurs y de la population.

L'exposé qui précède vise à démontrer qu'une théorie donnée peut avoir des applications dans divers secteurs pratiques même si elle n'est pas nécessaire dans une situation donnée, comme les enquêtes à grande échelle avec fractions d'échantillonnage du premier degré négligeables. Il montre également que les plans d'échantillonnage avec probabilités inégales jouent un rôle essentiel dans l'échantillonnage d'enquête, malgré l'affirmation de Särndal (1996) selon laquelle des plans simples, comme l'ÉAS stratifié et l'échantillonnage stratifié de Bernoulli, ainsi que les estimateurs GREG, devraient remplacer les stratégies fondées sur l'échantillonnage avec probabilités inégales sans remise.

6. Analyse des données d'enquête et des méthodes de rééchantillonnage

Les méthodes-types d'analyse des données sont généralement fondées sur l'hypothèse de l'échantillonnage aléatoire simple, quoique certains progiciels tiennent compte des poids d'échantillonnage et fournissent des estimations ponctuelles correctes. Toutefois, l'application de méthodes-types aux données d'enquête, abstraction faite de l'effet du plan de sondage dû à la mise en grappes et aux probabilités de sélection inégales, risque de produire des inférences erronées, même pour de grands échantillons. En particulier, les erreurs-types des estimations de paramètres et des intervalles de confiance associés peuvent être lourdement sous-estimées, les taux d'erreur de type I des tests d'hypothèses peuvent être beaucoup plus élevés que les niveaux nominaux et les diagnostics de modèles-types, comme l'analyse des résidus pour déceler les écarts par rapport au modèle, sont aussi influencés. Kish et Frankel (1974) et d'autres auteurs se sont penchés sur certains de ces problèmes et ont souligné la nécessité de nouvelles méthodes qui tiennent suffisamment compte de la complexité des données provenant d'enquêtes à grande échelle. Fuller (1975) a mis au point des méthodes asymptotiquement valides d'analyse par régression linéaire, fondées sur des estimateurs de variance par linéarisation de Taylor. Au cours des vingt dernières années, on a fait des progrès rapides en mettant au point des méthodes appropriées. Les méthodes de rééchantillonnage jouent un rôle capital dans la mise au point de méthodes qui tiennent compte du plan d'enquête dans l'analyse des données. On a simplement besoin d'un fichier de données contenant les données observées, des poids d'échantillonnage finaux et des poids finaux correspondant à chaque pseudo-répétition produit par la méthode de rééchantillonnage. On peut alors utiliser des progiciels qui tiennent compte des poids d'échantillonnage dans l'estimation

ponctuelle des paramètres d'intérêt pour calculer les bons estimateurs et les erreurs-types, comme nous le démontrons ci-dessous. Les méthodes d'inférence par rééchantillonnage ont donc attiré l'attention des utilisateurs, qui peuvent très facilement effectuer les analyses eux-mêmes à l'aide de progiciels standard. Toutefois, la mise en circulation de fichiers de données à grande diffusion avec poids de rééchantillonnage risque d'entraîner des problèmes de confidentialité, comme l'identification des grappes à partir des poids de rééchantillonnage. Les théoriciens ont d'ailleurs un défi à relever : celui de mettre au point des méthodes appropriées qui préservent la confidentialité des données. Lu, Brick et Sitter (2004) ont proposé de regrouper les strates et de former des pseudo-répétitions en utilisant les strates combinées pour l'estimation de la variance, limitant ainsi le risque d'identification des grappes à partir du fichier de données à grande diffusion ainsi obtenu. Le groupement des strates ou des UPÉ à l'intérieur des strates simplifie l'estimation de la variance en réduisant le nombre de pseudo-répétitions utilisés, comparativement à la méthode jackknife avec suppression d'une grappe, qu'on utilise couramment et que nous abordons ci-dessous. Une méthode d'échantillonnage inverse servant à défaire la structure complexe des données d'enquête tout en offrant une protection contre la révélation des étiquettes de grappe (Hinkins, Oh et Scheuren 1997; Rao, Scott et Benhin 2003) semble prometteuse, mais il reste beaucoup de travail à accomplir sur les méthodes d'échantillonnage inverse avant qu'elle n'intéresse l'utilisateur.

Rao et Scott (1981, 1984) ont mené une étude systématique de l'effet du plan de sondage sur le test chi carré et le test du rapport des vraisemblances, tests standardisés associés à un tableau multiple de comptes estimatifs ou de proportions. Ils ont montré que la variable à tester était asymptotiquement distribuée sous forme de somme pondérée de variables χ_1^2 indépendantes, les poids étant les valeurs propres d'une matrice d'« effets généralisés du plan de sondage ». Ce résultat général montre que le plan d'enquête peut avoir un effet important sur le taux d'erreur de type I. Rao et Scott ont proposé des corrections simples du premier ordre aux statistiques chi carré standardisées, qu'on peut calculer à partir de tableaux publiés comprenant des estimations des effets du plan de sondage pour les cellules d'estimations et leurs totaux marginaux, ce qui facilite les analyses secondaires à partir de tableaux publiés. Ils ont également calculé des corrections du deuxième ordre qui sont plus exactes, mais qui nécessitent la connaissance d'une matrice complète des covariances estimatives des cellules d'estimations, comme dans le cas des tests de Wald, bien connus. Toutefois, les tests de Wald peuvent devenir très instables lorsque le nombre de cellules d'un tableau multiple augmente et que le nombre de grappes d'échantillon diminue, ce qui entraîne des taux d'erreur de type I démesurément élevés par rapport aux niveaux nominaux, contrairement aux corrections du deuxième ordre de Rao-Scott (Thomas et Rao 1987). Les corrections du premier et du deuxième ordre sont maintenant appelées corrections de Rao-Scott et constituent des options par défaut dans la nouvelle version du SAS. Roberts, Rao et Kumar (1987) ont mis au point des corrections du type Rao-Scott pour les tests d'analyse de régression logistique des proportions estimatives des cellules associées à une variable de réponse binaire. Ils ont appliqué les méthodes à un tableau à double entrée de taux d'emploi provenant de l'EPA du Canada de 1977, obtenus en recoupant des groupes d'âge et de niveau de scolarité. Bellhouse et Rao (2002) ont étendu les travaux de Roberts et coll. à l'analyse des moyennes de domaine à l'aide de modèles linéaires généralisés. Ils ont appliqué les méthodes aux moyennes de domaine provenant d'une enquête sur la fécondité menée au Fidji, recoupées par niveau de scolarité et par nombre d'années depuis le premier mariage de la femme, une moyenne de domaine étant le nombre moyen d'enfants nés de femmes de race indienne appartenant au domaine.

Dans le contexte des enquêtes à grande échelle utilisant des plans d'échantillonnage stratifié à plusieurs degrés, les méthodes de rééchantillonnage ont fait l'objet de nombreuses études. Pour les besoins de l'inférence, les UPÉ de l'échantillon sont traitées comme si elles étaient tirées avec remise à l'intérieur des strates. Les variances s'en trouvent surestimées, mais cette surestimation est faible si la fraction d'échantillonnage globale des UPÉ est négligeable. Soit $\hat{\theta}$ l'estimateur pondéré d'un paramètre « de recensement » d'intérêt, calculé d'après les poids finaux w_i, et soient les poids correspondant à chaque pseudo-répétition r produits par la méthode de rééchantillonnage dénotés par $w_i^{(r)}$. L'estimateur fondé sur les pseudo-poids de rééchantillonnage $w_i^{(r)}$ est dénoté $\hat{\theta}^{(r)}$ pour chaque $r = 1, ..., R$. Un estimateur de variance par rééchantillonnage de $\hat{\theta}$ prend alors la forme

$$v(\hat{\theta}) = \sum_{r=1}^{R} c_r (\hat{\theta}^{(r)} - \hat{\theta})(\hat{\theta}^{(r)} - \hat{\theta})' \qquad (10)$$

pour les coefficients spécifiés c_r dans (10) déterminés par la méthode de rééchantillonnage.

Les méthodes de rééchantillonnage couramment utilisées comprennent : a) le jackknife avec suppression d'une grappe (ou d'une UPÉ), b) la répétition compensée (*balanced repeated replicate* ou BRR), notamment pour $n_h = 2$ UPÉ dans chaque strate h, et c) le bootstrap de Rao et Wu (1988). On obtient les pseudo-répétitions jackknife en supprimant tour à tour chaque grappe d'échantillon $r = (hj)$, et les poids de sondage jackknife $d_i^{(r)}$ prennent la valeur 0 si l'unité d'échantillonnage i est dans la grappe

supprimée, $n_h d_i/(n_h-1)$ si i n'est pas dans la grappe supprimée mais dans la même strate, et restent inchangés si i est dans une strate différente. Les poids de sondage jackknife sont alors rajustés pour la non-réponse totale et la post-stratification, ce qui donne les poids jackknife finaux $w_i^{(r)}$. L'estimateur jackknife de la variance est donné par (10) avec $c_r=(n_h-1)/n_h$ pour $r=(hj)$. La méthode du jackknife avec suppression d'une grappe peut présenter deux inconvénients : 1) lorsque le nombre total d'UPÉ échantillonnées, $n=\sum n_h$, est très élevé, R est aussi très élevé parce que $R=n$; 2) on ignore si l'estimateur jackknife de la variance avec suppression d'une grappe est convergent selon le plan de sondage dans le cas d'estimateurs non lisses $\hat{\theta}$, par exemple, l'estimateur pondéré de la médiane. Pour l'échantillonnage aléatoire simple, on sait que le jackknife est non convergent pour la médiane ou d'autres quantiles. Il serait stimulant (sur le plan théorique) et pertinent (sur le plan pratique) de trouver les conditions de convergence de l'estimateur jackknife de la variance avec suppression d'une grappe d'un estimateur non lisse $\hat{\theta}$.

La méthode BRR peut convenir à l'estimateur non lisse $\hat{\theta}$, mais elle ne s'applique aisément qu'à un cas particulier important, celui de $n_h=2$ UPÉ par strate. On peut construire un ensemble minimal de demi-échantillons équilibrés à partir d'une matrice Hadamard $R\times R$ en sélectionnant H colonnes, à l'exclusion de la colonne des $+1$, où $H+1\le R\le H+4$ (McCarthy 1969). Les poids de sondage BRR $d_i^{(r)}$ égalent $2d_i$ ou 0 selon que i se trouve ou non dans le demi-échantillon. Contrairement à la méthode BRR, une méthode BRR modifiée, due à Bob Fay, utilise toutes les unités échantillonnées lors de chaque répétition en définissant les poids de rééchantillonnage comme suit : $d_i^{(r)}(\varepsilon)=(1+\varepsilon)d_i$ ou $(1-\varepsilon)d_i$ selon que i se trouve ou non dans le demi-échantillon, où $0<\varepsilon<1$; un bon choix de ε est $1/2$. Les poids BRR modifiés sont alors rajustés pour la non-réponse et la post-stratification, ce qui donne les poids finaux $w_i^{(r)}(\varepsilon)$ et l'estimateur $\hat{\theta}^{(r)}(\varepsilon)$. L'estimateur BRR modifié de la variance est donné par (10) divisé par ε^2, $\hat{\theta}^{(r)}$ étant remplacé par $\hat{\theta}^{(r)}(\varepsilon)$ (voir Rao et Shao (1999)). L'estimateur BRR modifié est particulièrement utile dans le cas d'une réimputation indépendante de réponses manquantes lors de chaque répétition, car il peut utiliser les donneurs de l'échantillon complet pour réaliser l'imputation, contrairement à l'estimateur BRR, qui utilise uniquement les donneurs du demi-échantillon.

Contrairement à la méthode BRR, le bootstrap de Rao-Wu est valide pour les $n_h(\ge 2)$ arbitraires ainsi que pour les estimateurs non lisses $\hat{\theta}$. On construit chaque répétition bootstrap en tirant un échantillon aléatoire simple d'UPÉ de taille n_h-1 à partir des grappes d'échantillon n_h, indépendamment d'une strate à l'autre. Les poids de sondage bootstrap $d_i^{(r)}$ sont donnés par $[n_h/(n_h-1)]m_{hi}^{(r)}d_i$ si i est dans la strate h et la répétition r, où $m_{hi}^{(r)}$ est le nombre de fois que l'UPÉ échantillonnée (hi) est sélectionnée, $\sum_i m_{hi}^{(r)}=n_h-1$. Les poids $d_i^{(r)}$ sont alors rajustés pour la non-réponse totale et la post-stratification, ce qui donne les poids bootstrap finaux et l'estimateur $\hat{\theta}^{(r)}$. Souvent, on utilise $R=500$ répétitions dans l'estimateur bootstrap de la variance (10). Plusieurs enquêtes récentes de Statistique Canada ont adopté la méthode bootstrap d'estimation de la variance en raison de sa souplesse dans le choix de R et de sa grande applicabilité. Les utilisateurs des fichiers de microdonnées d'enquête de Statistique Canada semblent très satisfaits de la méthode bootstrap d'analyse des données.

Les premiers travaux sur le jackknife et le BRR étaient en grande partie empiriques (cf. Kish et Frankel 1974). Krewski et Rao (1981) ont élaboré un cadre asymptotique formel approprié pour l'échantillonnage stratifié à plusieurs degrés et ont établi la convergence selon le plan de sondage des estimateurs jackknife et BRR de la variance lorsque $\hat{\theta}$ peut être exprimé comme une fonction lisse des moyennes estimatives. Plusieurs ajouts à ces travaux de base ont été signalés dans la documentation récente, comme en témoigne l'ouvrage de Shao et Tu (1995, chapitre 6). Le soutien théorique des méthodes de rééchantillonnage est essentiel pour leur utilisation dans la pratique.

Dans l'exposé qui précède, $\hat{\theta}$ dénote l'estimateur d'un paramètre « de recensement ». Ordinairement, le paramètre de recensement θ_C est motivé par un modèle sous-jacent de superpopulation et le recensement est considéré comme un échantillon produit par le modèle, ce qui donne des équations d'estimation de recensement dont la solution est θ_C. Les fonctions d'estimation de recensement $U_C(\theta)$ sont simplement des totaux de population des fonctions $u_i(\theta)$ avec espérance nulle sous le modèle hypothétique, et les équations d'estimation de recensement sont données par $U_C(\theta)=0$ (Godambe et Thompson 1986). Kish et Frankel (1974) ont soutenu que le paramètre de recensement est valable même si le modèle n'est pas correctement spécifié. Par exemple, dans le cas de la régression linéaire, le coefficient de régression de recensement pourrait expliquer dans quelle mesure la relation entre la variable réponse et les variables indépendantes est prise en compte par un modèle de régression linéaire. Comme les fonctions d'estimation de recensement sont simplement des totaux de population, on obtient les estimateurs pondérés $\hat{U}(\theta)$ à partir de l'échantillon complet et $\hat{U}^{(r)}(\theta)$ à partir de chaque pseudo-répétition. Les solutions des équations d'estimation correspondantes $\hat{U}(\theta)=0$ et $\hat{U}^{(r)}(\theta)=0$ donnent respectivement $\hat{\theta}$ et $\hat{\theta}^{(r)}$. Il est à noter que les estimateurs de variance par rééchantillonnage ont pour objet d'estimer la variance de $\hat{\theta}$ comme un estimateur des paramètres de recensement, mais non des paramètres de modélisation.

Dans certaines conditions, on peut faire abstraction de la différence mais, en général, on est en présence d'une situation d'échantillonnage à deux phases où le recensement est l'échantillon de première phase tiré de la superpopulation et l'échantillon est un échantillon probabiliste tiré de la population de recensement. Récemment, on a mené des travaux utiles sur l'estimation de la variance à deux phases lorsque les paramètres de modélisation sont les paramètres cibles (Graubard et Korn 2002; Rubin-Bleuer et Schiopu Kratina 2005), mais il faudrait approfondir ces travaux pour surmonter la difficulté de spécifier la structure de covariance des erreurs de modèle.

Le bootstrap présente une difficulté : la solution $\hat{\theta}^{(r)}$ n'existe pas nécessairement pour certaines répétitions bootstrap r (Binder, Kovacevic et Roberts 2004). Rao et Tausi (2004) ont utilisé la méthode du bootstrap avec fonction d'estimation, qui évite la difficulté. Selon cette méthode, on résout $\hat{U}(\theta) = \hat{U}^{(r)}(\hat{\theta})$ pour θ en utilisant une seule étape de l'itération de Newton-Raphson avec $\hat{\theta}$ comme valeur de départ. On utilise alors dans (10) l'estimateur $\tilde{\theta}^{(r)}$ ainsi obtenu pour calculer l'estimateur bootstrap avec fonction d'estimation de la variance de $\hat{\theta}$ qu'on peut facilement mettre en œuvre à partir du fichier de données qui fournit les poids de rééchantillonnage, en modifiant légèrement un progiciel qui tient compte des poids d'échantillonnage. Il est intéressant de noter que l'estimateur bootstrap avec fonction d'estimation de la variance équivaut à un estimateur de sandwich par linéarisation de Taylor de la variance qui utilise l'estimateur bootstrap de la variance de $\hat{U}(\theta)$ et l'inverse de la matrice d'information observée (dérivée de $-\hat{U}(\theta)$), tous deux évalués à $\theta = \hat{\theta}$ (Binder et coll. 2004).

Contrairement aux méthodes de rééchantillonnage, les méthodes de linéarisation de Taylor produisent des estimateurs de variance asymptotiquement valides pour les plans d'échantillonnage généraux, mais elles nécessitent une formule distincte pour chaque estimateur $\hat{\theta}$. Binder (1983), Rao, Yung et Hidiroglou (2002) et Demnati et Rao (2004) ont fourni des formules unifiées d'estimation de variance par linéarisation pour des estimateurs définis comme des solutions aux équations d'estimation.

Pfeffermann (1993) a étudié le rôle des poids de sondage dans l'analyse des données d'enquête. Si le modèle de population est valable pour l'échantillon (c'est-à-dire s'il est sans biais d'échantillonnage), les estimateurs non pondérés fondés sur un modèle sont alors plus efficaces que les estimateurs pondérés et donnent des inférences valides, notamment pour des données où la taille des échantillons est faible et la variation des poids est élevée. Toutefois, pour les données ordinaires provenant d'enquêtes à grande échelle, le plan d'enquête est informatif et le modèle de population n'est pas nécessairement valable pour l'échantillon. Par conséquent, les estimateurs fondés sur un modèle peuvent être fortement biaisés et les inférences risquent d'être erronées. Pfeffermann et ses collègues ont proposé une nouvelle approche de l'inférence sous échantillonnage informatif (voir Pfeffermann et Sverchkov 2003), qui semble donner des inférences plus efficientes que l'approche pondérée et mérite certainement l'attention des utilisateurs de données d'enquête. Toutefois, il reste beaucoup de travail à accomplir, surtout en ce qui concerne le traitement de données fondées sur l'échantillonnage à plusieurs degrés.

Skinner, Holt et Smith (1989), Chambers et Skinner (2003) et Lehtonen et Pahkinen (2004) donnent d'excellentes descriptions des méthodes d'analyse de données d'enquête complexes.

7. Estimation sur petits domaines

Dans les sections précédentes, nous avons abordé les méthodes traditionnelles qui utilisent des estimateurs directs par domaine fondés sur des observations d'échantillons spécifiques aux domaines et sur des données auxiliaires sur la population. Toutefois, ces méthodes ne donnent pas nécessairement des inférences fiables lorsque la taille des échantillons du domaine est infime, voire nulle pour certains domaines. Dans la documentation, les domaines ou sous-populations dont la taille est infime ou nulle sont appelés petits domaines. Au cours des dernières années, la demande de statistiques fiables sur les petits domaines a grandement augmenté en raison du recours croissant à la statistique des petits domaines dans la formulation des politiques et des programmes, la répartition des fonds et la planification régionale. Manifestement, il est rarement possible d'obtenir des échantillons dont la taille globale est assez grande pour soutenir des estimations directes fiables pour tous les domaines d'intérêt. De plus, dans la pratique, il n'est pas possible de prévoir toutes les utilisations des données d'enquête et « le client exige toujours plus qu'il n'est spécifié à l'étape de l'élaboration du plan de sondage » (Fuller 1999, page 344). Pour faire des estimations sur petits domaines avec un niveau suffisant de précision, il faut souvent utiliser des estimateurs « indirects » qui empruntent de l'information à des domaines connexes par le biais de données auxiliaires, comme celles du recensement et les données administratives courantes, pour accroître la taille « effective » des échantillons à l'intérieur des petits domaines.

Aujourd'hui, on s'entend à reconnaître que des modèles explicites liant les petits domaines par le biais de données auxiliaires et tenant compte de la variation résiduelle entre domaines par le biais des effets aléatoires des petits domaines sont nécessaires pour calculer des estimateurs indirects. Le succès des méthodes fondées sur un modèle dépend fortement de la disponibilité de données auxiliaires

fiables et d'une validation complète des modèles au moyen d'évaluations internes et externes. Bon nombre de méthodes axées sur les effets aléatoires et utilisées dans la théorie statistique courante sont pertinentes à l'estimation sur petits domaines, dont la méthode du meilleur prédicteur empirique (ou méthode de Bayes), celle du meilleur prédicteur linéaire sans biais empirique et celle du modèle hiérarchique bayésien fondé sur des lois de distribution a priori des paramètres de modélisation. Rao (2003) donne une description complète de ces méthodes. La pertinence (sur le plan pratique) et l'intérêt (sur le plan théorique) de l'estimation sur petits domaines ont attiré l'attention de nombreux chercheurs, d'où la réalisation de progrès importants dans l'estimation ponctuelle et celle de l'erreur quadratique moyenne. Dans le monde entier, les « nouvelles » méthodes ont été appliquées avec succès à divers problèmes liés aux petits domaines. Aux États-Unis, on a utilisé récemment des méthodes fondées sur un modèle pour produire des estimations par comté et par district scolaire relativement aux enfants pauvres d'âge scolaire. Chaque année, le département de l'Éducation des États-Unis accorde aux comtés des fonds de plus de sept milliards de dollars sur la base d'estimations par comté fondées sur un modèle. Les fonds alloués soutiennent des programmes d'éducation compensatoire pour répondre aux besoins des enfants défavorisés sur le plan scolaire. Le lecteur trouvera dans Rao (2003, exemple 7.1.2) des renseignements sur cette application. Au Royaume-Uni, le Office of National Statistics a mis sur pied un projet d'estimation sur petits domaines pour établir des estimations fondées sur un modèle au niveau des sections électorales (quelque 2 000 ménages). Schaible (1996) décrit la pratique et les méthodes d'estimation des programmes statistiques fédéraux des États-Unis qui utilisent des estimateurs indirects pour produire des estimations publiées. Singh, Gambino et Mantel (1994) et Brackstone (2002) traitent de certains aspects pratiques et stratégiques de la statistique des petits domaines.

L'estimation sur petits domaines constitue un exemple frappant de l'interaction entre la théorie et la pratique. Les progrès de la théorie sont impressionnants, mais bon nombre de questions d'ordre pratique nécessitent une plus grande attention de la part des théoriciens, notamment les suivantes : a) des estimateurs d'étalonnage fondés sur un modèle pour concorder avec des estimateurs directs fiables au niveau des grands domaines; b) l'établissement et la validation de modèles de liaison appropriés et l'étude de questions comme les erreurs dans les variables, la spécification incorrecte du modèle de liaison et les variables omises; c) la mise au point de méthodes qui satisfont plusieurs objectifs : de bonnes estimations spécifiques au domaine, de bons rangs et un bon histogramme des petits domaines.

8. Certains aspects théoriques méritant l'attention des praticiens et vice versa

Dans la présente section, j'aborde brièvement quelques exemples d'aspects théoriques importants qui existent mais qui sont peu utilisés dans la pratique.

8.1 Inférence par la vraisemblance empirique

La théorie traditionnelle de l'échantillonnage portait dans une large mesure sur l'estimation ponctuelle et les erreurs-types associées, faisant appel à des approximations normales pour déterminer des intervalles de confiance à l'égard des paramètres d'intérêt. En statistique courante, l'approche de la vraisemblance empirique (VE) (Owen 1988) a beaucoup attiré l'attention en raison de plusieurs propriétés souhaitables. Elle offre une vraisemblance non paramétrique, ce qui donne des intervalles de confiance de VE semblables aux intervalles de vraisemblance paramétrique. La forme et l'orientation des intervalles de VE sont entièrement déterminées par les données; les intervalles préservent l'étendue tout en respectant la transformation et, contrairement aux intervalles symétriques de la théorie normale, ils sont particulièrement utiles puisqu'ils donnent des taux d'erreur équilibrés de la queue. Comme je l'ai mentionné dans la section 3.1, Hartley et Rao (1968) ont été les premiers à proposer l'approche de la VE dans le contexte des enquêtes par sondage, mais leur démarche était axée sur des questions d'inférence liées à l'estimation ponctuelle. Chen, Chen et Rao (2003) ont obtenu des intervalles de VE sur la moyenne de population sous échantillonnage aléatoire simple et sous échantillonnage aléatoire stratifié pour des populations contenant bien des zéros. On trouve ces populations dans le contrôle par sondage, où y dénote le montant d'argent dû à l'État et la moyenne arithmétique $\overline{Y}$ correspond au montant moyen des créances excessives. Des travaux antérieurs sur le contrôle par sondage ont utilisé des intervalles de vraisemblance paramétrique fondés sur des distributions de mélanges paramétriques pour la variable y. Ces intervalles donnent de meilleurs résultats que les intervalles-types de la théorie normale, mais les intervalles de VE donnent de meilleurs résultats en présence d'écarts par rapport au modèle hypothétique de mélanges, en donnant un taux de non-couverture inférieur à la borne inférieure plus proche du taux d'erreur nominal, ainsi qu'une borne inférieure plus grande. Pour les plans généraux, Wu et Rao (2004) ont utilisé une pseudo-vraisemblance empirique (Chen et Sitter 1999) pour obtenir des intervalles de pseudo-VE rajustés sur la moyenne arithmétique et la fonction de distribution qui tiennent compte des caractéristiques du plan, et ils ont montré que les intervalles donnaient des taux d'erreur de la queue plus équilibrés que dans le cas des intervalles de la théorie normale. La méthode VE offre

également une approche systématique de l'estimation par calage et de l'intégration des enquêtes. Le lecteur est invité à consulter les articles de Rao (2004) et Wu et Rao (2005).

Il reste encore à perfectionner ces notions, notamment en ce qui concerne la pseudo-vraisemblance empirique, mais la théorie de la VE dans le contexte des enquêtes mérite l'attention des praticiens.

8.2 Analyses exploratoires des données d'enquête

Dans la section 6, nous avons abordé les méthodes d'analyse confirmative de données d'enquête tenant compte du plan de sondage, comme l'estimation ponctuelle des paramètres de modélisation (ou de recensement) et des erreurs-types associées, ainsi que les tests formels d'hypothèses. Les graphiques et les analyses exploratoires des données d'enquête sont aussi très utiles. Ces méthodes ont fait l'objet d'une foule d'études dans la documentation courante. Encore récemment, certains ajouts à ces méthodes modernes ont été signalés dans la documentation sur les enquêtes et ils méritent l'attention des praticiens. J'en aborde brièvement un certain nombre. Premièrement, on utilise couramment des estimations non paramétriques de densité du noyau pour présenter la forme d'un ensemble de données sans recourir à des modèles paramétriques. On peut aussi les utiliser pour comparer différentes sous-populations.

Bellhouse et Stafford (1999) ont proposé des estimateurs de densité du noyau qui tiennent compte du plan d'enquête, en ont étudié les propriétés et ont appliqué les méthodes aux données de l'Enquête sur la santé en Ontario. Buskirk et Lohr (2005) ont étudié les propriétés asymptotiques et les propriétés de population finie des estimateurs de densité du noyau et ont obtenu des bandes de confiance. Ils ont appliqué les méthodes aux données de deux enquêtes américaines, la National Crime Victimization Survey et la National Health and Nutrition Examination Survey.

Deuxièmement, Bellhouse et Stafford (2001) ont mis au point des méthodes de régression polynomiale locale qui tiennent compte du plan de sondage et qu'on peut utiliser pour étudier la relation entre une variable réponse et des variables prédictives sans faire d'hypothèses audacieuses au sujet d'un modèle paramétrique. Les graphiques ainsi obtenus sont utiles pour comprendre les relations ainsi que pour comparer différentes sous-populations. À l'aide des données de l'Enquête sur la santé en Ontario, les auteurs ont illustré la régression polynomiale locale en montrant, par exemple, la relation entre l'indice de masse corporelle des femmes et leur âge. Bellhouse, Chipman et Stafford (2004) ont étudié des modèles additifs de données d'enquête au moyen de la méthode des moindres carrés pénalisée pour traiter plus d'une variable prédictive, et ont illustré les méthodes à l'aide des données de l'Enquête sur la santé en Ontario. Cette approche offre de nombreux avantages en ce qui concerne les graphiques, l'estimation, les tests et la sélection de paramètres « de lissage » pour ajuster les modèles.

8.3 Erreurs de mesure

Habituellement, on suppose que les erreurs de mesure sont additives et que leur moyenne est nulle. Par conséquent, les estimateurs habituels du total et des moyennes restent sans biais ou convergents. Toutefois, cette caractéristique positive n'est pas nécessairement valable pour des paramètres plus complexes comme la fonction de distribution, les quantiles et les coefficients de régression. Dans ce dernier cas, les estimateurs habituels sont biaisés, même pour de grands échantillons, et peuvent donc produire des inférences erronées (Fuller 1995). Il est possible d'obtenir des estimateurs corrigés pour le biais si l'on dispose d'estimations des variances de l'erreur de mesure. On peut obtenir ces dernières en affectant des ressources, à l'étape de l'élaboration du plan de sondage, pour faire des observations répétées sur un sous-échantillon. Fuller (1975, 1995) a préconisé l'utilisation de méthodes appropriées en présence d'erreurs de mesure, et les méthodes corrigées pour le biais méritent l'attention des praticiens.

Hartley et Rao (1978) et Hartley et Biemer (1978) ont établi des conditions d'affectation des intervieweurs et des codeurs qui permettent d'estimer les variances d'échantillonnage et de réponse pour la moyenne arithmétique ou le total à partir d'enquêtes courantes. Malheureusement, le plan de sondage des enquêtes d'aujourd'hui satisfait rarement ces conditions et, même si c'était le cas, on dispose rarement de l'information requise sur les affectations des intervieweurs et des codeurs à l'étape de l'estimation.

On utilise souvent les composantes linéaires des modèles de variance pour estimer la variabilité des intervieweurs. Ces modèles sont appropriés pour la réponse continue, mais pas pour les réponses binaires. L'approche du modèle linéaire pour les réponses binaires peut entraîner une sous-estimation des corrélations intra-intervieweurs. Scott et Davis (2001) ont proposé des modèles hiérarchiques pour les réponses binaires afin d'estimer la variabilité due aux intervieweurs. Comme les réponses sont souvent binaires dans bon nombre d'enquêtes, les praticiens doivent prêter attention à ces modèles pour effectuer des analyses pertinentes des données d'enquête avec réponses binaires.

8.4 Imputation des données d'enquête manquantes

Dans la pratique, on utilise couramment l'imputation pour remplacer des éléments manquants. On s'assure ainsi que les résultats d'analyses différentes de l'ensemble de données complété sont cohérents entre eux en utilisant le même poids d'échantillonnage pour tous les éléments. Bon nombre d'organismes statistiques utilisent des méthodes

d'imputation marginale comme celles du ratio, du plus proche voisin et du donneur aléatoire à l'intérieur des classes d'imputation. Malheureusement, on traite souvent les valeurs imputées comme s'il s'agissait de valeurs vraies, puis on calcule des estimations et des estimations de la variance. Les estimations ponctuelles imputées de paramètres marginaux sont généralement valides en présence d'un mécanisme de réponse ou d'un modèle d'imputation hypothétique. Mais les estimateurs « naïfs » de la variance peuvent produire des inférences erronées, même pour de grands échantillons, notamment une forte sous-estimation de la variance de l'estimateur imputé, faute de prendre en compte la variabilité additionnelle due à l'estimation des valeurs manquantes. Les partisans de l'imputation multiple de Rubin (1987) soutiennent que l'estimateur de variance à imputation multiple peut régler ce problème parce qu'une somme des carrés entre estimateurs imputés est ajoutée à la moyenne des estimateurs naïfs de la variance obtenus au moyen des imputations multiples. Malheureusement, les estimateurs de variance à imputation multiple comportent certaines difficultés, comme en font état Kott (1995), Fay (1996), Binder et Sun (1996), Wang et Robins (1998), Kim, Brick, Fuller et Kalton (2004) et d'autres auteurs. En outre, on préfère souvent l'imputation simple pour des raisons d'efficacité opérationnelle et de rentabilité. Au cours des dernières années, on a fait des progrès impressionnants en réalisant des inférences efficientes et asymptotiquement valides à partir d'ensembles de données imputées une seule fois. Le lecteur est invité à consulter les articles de Shao (2002) et Rao (2000, 2005) sur les méthodes d'estimation de la variance au moyen de l'imputation simple. Kim et Fuller (2004) ont étudié l'imputation partielle en utilisant plus d'une valeur imputée au hasard et ont montré que cette méthode donnait également des inférences asymptotiquement valides; voir aussi Kalton et Kish (1984) et Fay (1996). L'imputation partielle offre l'avantage de réduire la variance due à l'imputation par rapport à l'imputation unique utilisant une seule valeur imputée au hasard. Les méthodes d'estimation de la variance susmentionnées méritent l'attention des praticiens.

8.5 Enquêtes à bases multiples

Les enquêtes à bases multiples emploient deux ou plusieurs bases chevauchantes pour couvrir entièrement la population cible. Hartley (1962) a étudié le cas particulier d'une base complète B, d'une base incomplète A et d'un échantillonnage aléatoire simple mené indépendamment dans les deux bases. Il a montré que par rapport à l'estimateur à base unique complète, un estimateur à double base « optimal » pouvait donner lieu à d'importants gains d'efficience pour le même coût, à condition que le coût par unité pour la base A soit nettement inférieur au coût par unité pour la base B. Les enquêtes à bases multiples conviennent particulièrement à l'échantillonnage de populations rares ou difficiles à joindre, comme les populations de sans-abri et de personnes atteintes du SIDA, lorsque des listes incomplètes contiennent de fortes proportions de personnes appartenant à la population cible. Dans un article marquant, Hartley (1974) a calculé des estimateurs à double base « optimaux » pour des plans d'échantillonnage généraux et des unités d'observation pouvant être différentes dans les deux bases. Fuller et Burmeister (1972) ont proposé des estimateurs « optimaux » améliorés. Toutefois, les estimateurs optimaux utilisent des ensembles de poids différents pour chaque élément y, ce qui n'est pas souhaitable dans la pratique. Skinner et Rao (1996) ont calculé, pour les enquêtes à double base, des estimateurs du pseudo-maximum de vraisemblance (PMV) qui utilisent le même ensemble de poids pour tous les éléments y, comme dans le cas des estimateurs « à base unique » (Kalton et Anderson 1986), et qui maintiennent l'efficience. Lohr et Rao (2005) ont formulé une théorie unifiée des conditions des enquêtes à bases multiples en prolongeant les estimateurs optimal, PMV et à base unique. Lohr et Rao (2000, 2005) ont obtenu des estimateurs de variance jackknife asymptotiquement valides. Ces résultats généraux méritent l'attention des praticiens lorsqu'on travaille avec deux ou plusieurs bases. Les enquêtes téléphoniques à double base (téléphones cellulaires et téléphones fixes) nécessitent l'attention des théoriciens, car on ignore comment pondérer dans le cas de l'enquête menée par téléphone cellulaire : certaines familles partagent un téléphone cellulaire, d'autres en possèdent un pour chaque personne.

8.6 Échantillonnage indirect

On peut utiliser la méthode de l'échantillonnage indirect lorsqu'on ne dispose pas de la base d'une population cible U^B mais qu'on emploie la base d'une autre population U^A, liée à U^B, pour tirer un échantillon probabiliste. On utilise les liens entre les deux populations pour établir des poids appropriés qui peuvent donner des estimateurs sans biais et des estimateurs de variance. Lavallée (2002) a mis au point une méthode unifiée, appelée méthode généralisée du partage des poids (MGPP), inspirée de plusieurs méthodes connues : la méthode du partage des poids d'Ernst (1989) pour l'estimation transversale à partir d'enquêtes-ménages longitudinales, l'échantillonnage par réseau et l'estimation de la multiplicité (Sirken 1970), ainsi que l'échantillonnage en grappes adaptatif (Thompson et Seber 1996). La théorie de Rao (1968) sur l'échantillonnage à partir d'une base contenant une quantité inconnue de doubles comptes peut être considérée comme un cas particulier de la MGPP. On peut aussi employer la MGPP pour travailler avec des bases multiples; les estimateurs ainsi

obtenus sont simples, mais pas nécessairement efficients par rapport aux estimateurs optimaux de Hartley (1974) ou aux estimateurs du PMV. La méthode MGPP a une grande applicabilité et mérite l'attention des praticiens.

9. Conclusion

L'apport de Joe Waksberg à la théorie et aux méthodes des enquêtes par sondage reflète bien l'interaction entre la théorie et la pratique. Dans le cadre de son travail au Census Bureau des États-Unis, puis à Westat, il a fait face à de réels problèmes d'ordre pratique et a souvent trouvé des solutions théoriques judicieuses. Par exemple, dans un article marquant (Waksberg 1978), il a décrit une ingénieuse méthode de composition aléatoire (CA) qui réduit considérablement les coûts d'enquête par rapport à la composition de numéros entièrement au hasard. Il a présenté des arguments théoriques solides pour en démontrer l'efficience. L'utilisation généralisée des enquêtes par CA est due pour une bonne part à l'argumentation théorique de Waksberg (1978) et à des perfectionnements ultérieurs. Joe Waksberg est un spécialiste de l'échantillonnage d'enquête que j'admire énormément et je suis très honoré d'avoir reçu le prix Waksberg 2005 pour les techniques d'enquête.

Remerciements

Je tiens à remercier David Bellhouse, Wayne Fuller, Jack Gambino, Graham Kalton, Fritz Scheuren et Sharon Lohr, dont les observations et les suggestions m'ont été très utiles.

Bibliographie

Aires, N., et Rosén, B. (2005). On inclusion probabilities and relative estimator bias for Pareto πps sampling. *Journal of Statistical Planning and Inference*, 128, 543-567.

Andreatta, G., et Kaufmann, G.M. (1986). Estimation of finite population properties when sampling is without replacement and proportional to magnitude. *Journal of the American Statistical Association*, 81, 657-666.

Bankier, M.D. (1988). Power allocations: determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.

Bankier, M.D. (2003). 2001 Canadian Census weighting: switch from projection GREG to pseudo-optimal regression estimation. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, Rapport technique no. 386, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa.

Bankier, M.D., Rathwell, S. et Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. Document de travail, direction de la méthodologie, division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa.

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part I. Dans *Foundations of Statistical Inference* (Éds. V.P. Godambe et D.A. Sprott), Toronto: Holt, Rinehart and Winston, 203-242.

Bellhouse, D.R., et Rao, J.N.K. (2002). Analysis of domain means in complex surveys. *Journal of Statistical Planning and Inference*, 102, 47-58.

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

Bellhouse, D.R., et Stafford, J.E. (2001). Régression polynomiale locale dans le cas des enquêtes complexes. *Techniques d'enquête*, 27, 219-226.

Bellhouse, D.R., Chipman, H.A. et Stafford, J.E. (2004). Additive models for survey data via penalized least squares. Rapport technique.

Binder, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Binder, D.A., et Sun, W. (1996). Frequency valid multiple imputation for surveys with a complex design. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 281-286.

Binder, D.A., Kovacevic, M. et Roberts, G. (2004). Design-based methods for survey data: Alternative uses of estimating functions. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute*, 22, Supplement to Liv. 1, 6-62.

Brackstone, G. (2002). Stratégies et approches relatives aux statistiques régionales. *Techniques d'enquête*, 28, 125-133.

Brackstone, G., et Rao, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā*, Series C, 42, 97-114.

Brewer, K.R.W. (1963). Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.

Brewer, K.R.W., et Hanif, M. (1983). *Sampling With Unequal Probabilities*. New York: Springer-Verlag.

Buskirk, T.D., et Lohr, S.L. (2005). Asymptotic properties of kernel density estimation with complex survey data. *Journal of Statistical Planning and Inference*, 128, 165-190.

Casady, R.J., et Valliant, R. (1993). Propriétés conditionnelles des estimateurs de stratification a posteriori selon la théorie normale. *Techniques d'enquête*, 19, 193-203.

Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

Chambers, R.L., et Skinner, C.J. (Éds.) (2003). *Analysis of Survey Data*. Chichester: Wiley.

Chen, J., et Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 12, 1223-1239.

Chen, J., Chen, S.Y. et Rao, J.N.K. (2003). Empirical likelihood confidence intervals for the mean of a population containing many zero values. *The Canadian Journal of Statistics*, 31, 53-68.

Chen, J., Sitter, R.R. et Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89, 230-237.

Cochran, W.G. (1939). The use of analysis of variance in enumeration by sampling. *Journal of the American Statistical Association*, 34, 492-510.

Cochran, W.G. (1940). The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce. *Journal of Agricultural Science*, 30, 262-275.

Cochran, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 37, 191-212.

Cochran, W.G. (1946). Relative accuracy of systematic and stratified random samples from a certain class of populations. *Annals of Mathematical Statistics*, 17, 164-177.

Cochran, W.G. (1953). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

Dalenius, T. (1957). *Sampling in Sweden*. Stockholm: Almquist and Wicksell.

Dalenius, T., et Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.

Deming, W.E. (1950). *Some Theory of Sampling*. New York: John Wiley & Sons, Inc.

Deming, W.E. (1960). *Sample Design in Business Research*. New York: John Wiley & Sons, Inc.

Deming, W.E., et Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected margins are known. *The Annals of Mathematical Statistics*, 11, 427-444.

Demnati, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 17-27.

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

Durbin, J. (1968). Sampling theory for estimates based on fewer individuals than the number selected. *Bulletin of the International Statistical Institute*, 36, No. 3, 113-119.

Ericson, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society*, Series B, 31, 195-224.

Ernst, L.R. (1989). Weighting issues for longitudinal household and family estimates. Dans *Panel Surveys* (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh), New York: John Wiley & Sons, Inc., 135-169.

Ernst, L.R. (1999). The maximization and minimization of sample overlap problem: A half century of results. *Bulletin of the International Statistical Institute*, Vol. LVII, Book 2, 293-296.

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

Fellegi, I.P. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59, 1016-1041.

Fellegi, I.P. (1966). Changing the probabilities of selection when two units are selected with PPS sampling without replacement. *Proceedings of the Social Statistics Section, American Statistical Association*, Washington DC, 434-442.

Fellegi, I.P. (1981). Should the census counts be adjusted for allocation purposes? – Equity considerations. Dans *Current Topics in Survey Sampling* (Éds. D. Krewski, R. Platek et J.N.K. Rao). New York: Academic Press, 47-76.

Francisco, C.A., et Fuller, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.

Fuller, W.A. (1975). Regression analysis for sample survey. *Sankhyā*, Series C, 37, 117-132.

Fuller, W.A. (1995). Estimation in the presence of measurement error. *Revue Internationale de Statistique*, 63, 121-147.

Fuller, W.A. (1999). Environmental surveys over time, *Journal of Agricultural, Biological and Environmental Statistics*, 4, 331-345.

Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.

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

Fuller, W.A., et Rao, J.N.K. (2001). Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada. *Techniques d'enquête* 27, 49-56.

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

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society*, Series B, 17, 269-278.

Godambe, V.P. (1966). A new approach to sampling from finite populations. *Journal of the Royal Statistical Society*, Series B, 28, 310-328.

Godambe, V.P., et Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationship and estimation. *Revue Internationale de Statistique*, 54, 127-138.

Graubard, B.I., et Korn, E.L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, 17, 73-96.

Hacking, I. (1975). *The Emergence of Probability*. Cambridge: Cambridge University Press.

Hajék, J. (1971). Comments on a paper by Basu, D. Dans *Foundations of Statistical Inference* (Éds. V.P. Godambe et D.A. Sprott), Toronto: Holt, Rinehart and Winston,

Hansen, M.H., et Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

Hansen, M.H., Dalenius, T. et Tepping, B.J. (1985). The development of sample surveys of finite populations. Chapter 13 in *A Celebration of Statistics*. The ISI Centenary Volume, Berlin: Springer-Verlag.

Hansen, M.H., Hurwitz, W.N. et Bershad, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 38, 359-374.

Hansen, M.H., Hurwitz, W.N. et Madow, W.G. (1953). *Sample Survey Methods and Theory*, Vols. I et II. New York: John Wiley & Sons, Inc.

Hansen, M.H., Madow, W.G. et Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

Hansen, M.H., Hurwitz, W.N., Marks, E.S. et Mauldin, W.P. (1951). Response errors in surveys. *Journal of the American Statistical Association*, 46, 147-190

Hansen, M.H., Hurwitz, W.N., Nisseslson, H. et Steinberg, J. (1955). The redesign of the census current population survey. *Journal of the American Statistical Association*, 50, 701-719.

Hartley, H.O. (1959). Analytical studies of survey data. In Volume in Honour of Corrado Gini, Instituto di Statistica, Rome, 1-32.

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

Hartley, H.O. (1974). Mutiple frame methdology and selected applications. *Sankhyā*, Series C, 36, 99-118.

Hartley, H.O., et Biemer, P. (1978). The estimation of nonsampling variances in current surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 257-262.

Hartley, H.O., et Rao, J.N.K. (1962). Sampling with unequal probability and without replacement. *The Annals of Mathematical Statistics*, 33, 350-374.

Hartley, H.O., et Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

Hartley, H.O., et Rao, J.N.K. (1978). The estimation of nonsampling variance components in sample surveys. Dans *Survey Measurement* (Éd. N.K. Namboodiri), New York: Academic Press, 35-43.

Hidiroglou, M.A., Fuller, W.A. et Hickman, R.D. (1976). SUPER CARP, Statistical Laboratory, Iowa State University, Ames, Iowa, États-Unis.

Hinkins, S., Oh, H.L. et Scheuren, F. (1997). Algorithmes de plan de sondage inverses. *Techniques d'enquête*, 23, 13-24.

Holt, D., et Smith, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society*, Series A, 142, 33-46.

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.

Huang, E.T., et Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section*, American Statistical Association, 300-305.

Hubback, J.A. (1927). Sampling for rice yield in Bihar and Orissa. Imperial Agricultural Research Institute, Pusa, Bulletin No. 166 (représenté dans *Sankhyā*, 1946, vol. 7, 281-294).

Hussain, M. (1969). Construction of regression weights for estimation in sample surveys. Thèse de maîtrise non-publiée, Iowa State University, Ames, Iowa.

Jessen, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin*, No. 304.

Kalton, G. (2002). Models in the practice of survey sampling (revisited). *Journal of Official Statistics*, 18, 129-154.

Kalton, G., et Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society*, Series A, 149, 65-82.

Kalton, G., et Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics*, A13, 1919-1939.

Keyfitz, N. (1951). Sampling with probabilities proportional to size: adjustment for changing in the probabilities. *Journal of the American Statistical Association*, 46, 105-109.

Kiaer, A. (1897). The representative method of statistical surveys (1976 English translation of the original Norwegian), Oslo. Central Bureau of Statistics of Norway.

Kim, J., et Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.

Kim, J.K., Brick, J.M., Fuller, W.A. et Kalton, G. (2004). On the bias of the multiple imputation variance estimator in survey sampling. Rapport technique.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Kish, L. (1995). The hundred year's wars of survey sampling. *Statistics in Transition*, 2, 813-830.

Kish, L., et Scott, A.J. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

Kish, L., et Frankel, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, Series B, 36, 1-37.

Kott, P.S. (1995). A paradox of multiple imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 384-389.

Kott, P.S. (2005). Randomized-assisted model-based survey sampling. *Journal of Statistical Planning and Inference*, 129, 263-277.

Krewski, D., et Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

Kruskal, W.H., et Mosteller, F. (1980). Representative sampling IV: The history of the concept in Statistics, 1895-1939. *Revue Internationale de Statistique*, 48, 169-195.

Laplace, P.S. (1820). A philosophical essay on probabilities. English translation, Dover, 1951.

Lavallée, P. (2002). *Le Sondage indirect, ou la Méthode généralisée du partage des poids.* Éditions de l'Université de Burxelles, Belgique, Éditions Ellipse, France.

Lavallée, P., et Hidiroglou, M. (1988). Sur la stratification de populations asymétriques. *Techniques d'enquête*, 14, 35-45.

Lehtonen, R., et Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*. Chichester: Wiley.

Lindley, D.V. (1996). Letter to the editor. *American Statistician*, 50, 197.

Lohr, S.L. (1999). *Sampling*: *Design and Analysis*. Pacific Grove: Duxbury.

Lohr, S.L., et Rao, J.N.K. (2000). Inference in dual frame surveys. *Journal of the American Statistical Association*, 95, 2710280.

Lohr, S.L., et Rao, J.N.K. (2005). Multiple frame surveys: point estimation and inference. *Journal of the American Statistical Association* (en révision).

Lu, W.W., Brick, M. et Sitter, R.R. (2004). Algorithms for constructing combined strata grouped jackknife and balanced repeated replication with domains. Rapport technique, Westat, Rockville, Maryland.

Mach, L, Reiss, P.T. et Schiopu-Kratina, I. (2005). The use of the transportation problem in co-ordinating the selection of samples for business surveys. Rapport technique HSMD-2005-006E, Statistique Canada, Ottawa.

Madow, W.G., et Madow, L.L. (1944). On the theory of systematic sampling. *Annals of Mathematical Statistics*, 15, 1-24.

Mahalanobis, P.C. (1944). On large scale sample surveys. *Philosophical Transactions of the Royal Society*, London, Series B, 231, 329-451.

Mahalanobis, P.C. (1946a). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.

Mahalanobis, P.C. (1946b). Sample surveys of crop yields in India. *Sankhyā*, 7, 269-280.

McCarthy, P.J. (1969). Pseudo-replication: Half samples. *Review of the International Statistical Institute*, 37, 239-264.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99, 1131-1139.

Murthy, M.N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyā*, 18, 379-390.

Murthy, M.N. (1964). On Mahalanobis' contributions to the development of sample survey theory and methods. Dans *Contributions to Statistics*: Présenté au professeur P.C. Mahalanobis à l'occasion de son 70ième anniversaire, Calcutta, Statistical Publishing Society: 283-316.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Aricultural Statistics*, 3, 169-174.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.

Ohlsson, E. (1995). Coordination of samples using permanent random members. Dans *Business Survey Methods* (Éds. B.G. Cox, D.A. Binder, N. Chinnappa, A. Christianson, M.J. Colledge et P.S. Kott), New York: John Wiley & Sons, Inc., 153-169.

O'Muircheartaigh, C.A., et Wong, S.T. (1981). The impact of sampling theory on survey sampling practice: A review. *Bulletin of the International Statistical Institute*, Article, 49, No. 1, 465-493.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

Owen, A.B. (2002). *Empirical Likelihood*. New York: Chapman & Hall/CRC.

Park, M., et Fuller, W.A. (2005). Vers des poids de régression non négatifs pour les échantillons d'enquête. *Techniques d'enquête*, 31, 93-101.

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

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Revue Internationale de Statistique*, 61, 317-337.

Pfeffermann, D., et Sverchkov, M. (2003). Fitting generalized linear models under informative sampling. Dans *Analysis of Survey Data* (Éds. R.L. Chambers et C.J. Skinner), Chichester: Wiley, 175-195.

Raj, D. (1956). On the method of overlapping maps in sample surveys. *Sankhyā*, 17, 89-98.

Rao, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā*, Series A, 28, 47-60.

Rao, J.N.K. (1968). Some nonresponse sampling theory when the frame contains an unknown amount of duplication. *Journal of the American Statistical Association*, 63, 87-90.

Rao, J.N.K. (1992). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Proceedings of the workshop on uses of auxiliary information in surveys*, Bureau de la statistique de Suède.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

Rao, J.N.K. (1996). Developments in sample survey theory: An appraisal. *The Canadian Journal of Statistics*, 25, 1-21.

Rao, J.N.K. (2000). Variance estimation in the presence of imputation for missing data. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association, 599-608.

Rao, J.N.K. (2003). *Small Area Estimation*. Hoboken: Wiley.

Rao, J.N.K. (2004). Empirical likelihood methods for sample survey data: An overview. *Proceedings of the Survey Methods Section*, SSC Annual Meeting, sous presse.

Rao, J.N.K. (2005). Re-sampling variance estimation with imputed survey data: overview. *Bulletin of the International Statistical Institute.*

Rao, J.N.K., et Bellhouse, D.R. (1990). Genèse et évolution des fondements théoriques de l'estimation et de l'analyse fondées sur les sondages. *Techniques d'enquête*, 16, 3-26.

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

Rao, J.N.K., et Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

Rao, J.N.K., et Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12, 46-60.

Rao, J.N.K., et Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.

Rao, J.N.K., et Singh, A.C. (1997). A ridge shrinkage method for range restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 57-64.

Rao, J.N.K., et Singh, M.P. (1973). On the choice of estimators in survey sampling. *Australian Journal of Statistics*, 15, 95-104.

Rao, J.N.K., et Tausi, M. (2004). Estimating function jackknife variance estimators under stratified multistage sampling. *Communications in Statistics – Theory and Methods*, 33, 2087-2095.

Rao, J.N.K., et Wu, C.F.J. (1987). Methods for standard errors and confidence intervals from sample survey data: Some recent work. *Bulletin of the International Statistical Institute.*

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Rao, J.N.K., Hartley, H.O. et Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, Series B, 24, 482-491.

Rao, J.N.K., Jocelyn, W. et Hidiroglou, M.A. (2003). Confidence interval coverage properties for regression estimators in uni-phase and two-phase sampling. *Journal of Official Statistics*, 19.

Rao, J.N.K., Scott, A.J. et Benhin, E. (2003). Défaire les structures des données d'enquête complexes : Théorie élémentaire et applications de l'échantillonnage inverse. *Techniques d'enquête*, 29, 119-131.

Rao, J.N.K., Yung, W. et Hidiroglou, M. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā*, Series A, 64, 364-378.

Renssen, R.H., et Nieuwenbroek, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-375.

Rivest, L.-P. (2002). Une généralisation de l'algorithme de Lavallée et Hidiroglou pour la stratification dans les enquêtes auprès des entreprises. *Techniques d'enquête*, 28, 207-214.

Roberts, G., Rao, J.N.K. et Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

Rosén, B. (1991). Variance estimation for systematic pps-sampling. Rapport technique, Bureau de la statistique de Suède.

Royall, R.M. (1968). An old approach to finite population sampling theory. *Journal of the American Statistical Association*, 63, 1269-1279.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

Royall, R.M., et Cumberland, W.G. (1981). An empirical study of the ratio estimate and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-88.

Royall, R.M., et Herson, J.H. (1973). Robust estimation in finite populations, I et II. *Journal of the American Statistical Association*, 68, 880-889 et 890-893.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons, Inc.

Rubin-Bleuer, S., et Schiopu-Kratina, I. (2005). On the two-phase framework for joint model and design-based inference. *Annals of Statistics*, (à paraître).

Salehi, M., et Seber, G.A.F. (1997). Adaptive cluster sampling with networks selected without replacements, *Biometrika*, 84, 209-219.

Särndal, C.-E. (1996). Efficient estimators with variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.

Särndal, C.-E., Swenson, B. et Wretman, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

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

Schabenberger, O., et Gregoire, T.G. (1994). Solutions de remplacement pour les plans πpt authentiques : Une étude comparative. *Techniques d'enquête*, 20, 193-200.

Schaible, W.L. (Ed.) (1996). Indirect Estimation in U.S. Federal Programs. New York: Springer

Scott, A., et Davis, P. (2001). Estimating interviewer effects for survey responses. *Proceedings of Statistics Canada Symposium 2001*.

Shao, J. (2002). Resampling methods for variance estimation in complex surveys with a complex design. Dans *Survey Non-response* (Éds. R.M. Groves, D.A. Dillman, J.L. Eltinge et R.J.A. Little), New York: John Wiley & Sons, Inc., 303-314.

Shao, J., et Tu, D. (1995). *The Jackknife and the Bootstrap*. New York: Springer Verlag.

Singh, A.C., Kennedy, B. et Wu, S. (2001). Estimation composite par régression pour l'Enquête sur la population active du Canada avec plan de sondage à renouvellement de panel. *Techniques d'enquête*, 27, 35-48.

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

Singh, A.C., et Wu, S. (1996). Estimation for multiframe complex surveys by modified regression. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 69-77.

Singh, M.P., Gambino, J. et Mantel, H.J. (1994). Les petites régions : Problèmes et solutions. *Techniques d'enquête*, 20, 3-15.

Sirken, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65, 257-266.

Sitter, R.R., et Wu, C. (2001). A note on Woodruff confidence interval for quantiles. *Statistics & Probability Letters*, 55, 353-358.

Skinner, C.J., et Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

Skinner, C.J., Holt, D. et Smith, T.M.F. (Éds.) (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons, Inc.

Smith, T.M.F. (1976). The foundations of survey sampling: A review. *Journal of the Royal Statistical Society*, Series A, 139, 183-204.

Smith, T.M.F. (1994). Sample surveys 1975-1990; an age of reconciliation? *Revue Internationale de Statistique*, 62, 5-34.

Stehman, S.V., et Overton, W.S. (1994). Comparison of variance estimators of the Horvitz Thompson estimator for randomized variable probability systematic sampling. *Journal of the American Statistical Association*, 89, 30-43.

Sukhatme, P.V. (1947). The problem of plot size in large-scale yield surveys. *Journal of the American Statistical Association*, 42, 297-310.

Sukhatme, P.V. (1954). *Sampling Theory of Surveys, with Applications.* Ames: Iowa State College Press.

Sukhatme, P.V., et Panse, V.G. (1951). Crop surveys in India – II. *Journal of the Indian Society of Agricultural Statistics*, 3, 97-168.

Sukhatme, P.V., et Seth, G.R. (1952). Non-sampling errors in surveys. *Journal of the Indian Society of Agricultural Statistics*, 4, 5-41.

Thomas, D.R., et Rao, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82, 630-636.

Thompson, S.K., et Seber, G.A.F. (1996). *Adaptive Sampling*. New York: John Wiley & Sons, Inc.

Tillé, Y. (1998). Estimation in surveys using conditional inclusion probabilities: simple random sampling. *Revue Internationale de Statistique*, 66, 303-322.

Tschuprow, A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron*, 2, 461-493, 646-683.

Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference*: *A Prediction Approach*. New York: John Wiley & Sons, Inc.

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

Waksberg, J. (1998). The Hansen era: Statistical research and its implementation at the U.S. Census Bureau. *Journal of Official Statistics*, 14, 119-135.

Wang, N., et Robins, J.M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85, 935-948.

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

Wu, C., et Rao, J.N.K. (2004). Empirical likelihood ratio confidence intervals for complex surveys. Soumis pour publication.

Wu, C., et Rao, J.N.K. (2005). Empirical likelihood approach to calibration using survey data. Article présenté à la réunion 2005 International Statistical Institute, Sydney, Australie.

Yates, F. (1949). *Sampling Methods for Censuses and Surveys*. London: Griffin.

Zarkovic, S.S. (1956). Note on the history of sampling methods in Russia. *Journal of the Royal Statistical Society*, Series A, 119, 336-338.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2005
Vol. 31, N° 2, pp. 153-164
Statistique Canada, N° 12-001-XPB au catalogue

Imputation hot deck pour le modèle de réponse

Wayne A. Fuller et Jae Kwang Kim [1]

Résumé

L'imputation hot deck est une procédure qui consiste à remplacer les réponses manquantes à certaines questions par des valeurs empruntées à d'autres répondants. L'un des modèles sur lesquels elle s'appuie est celui où l'on suppose que les probabilités de réponse sont égales dans les cellules d'imputation. Nous décrivons une version efficace de l'imputation hot deck pour le modèle de réponse dans les cellules et donnons un estimateur de la variance dont le traitement informatique est efficace. Nous détaillons une approximation de la procédure entièrement efficace dans laquelle un petit nombre de valeurs sont imputées pour chaque non-répondant. Nous illustrons les procédures d'estimation de la variance dans une étude de Monte Carlo.

Mots clés : Non-réponse, imputation fractionnaire; probabilité de réponse; estimation de la variance par rééchantillonnage.

1. Introduction

Dans les enquêtes par sondage, l'imputation est utilisée comme méthode de traitement de la non–réponse partielle. Dans le cas de l'imputation hot deck, les valeurs imputées sont des fonctions des répondants compris dans l'échantillon courant. Sande (1983) et Ford (1983) décrivent l'imputation hot deck. Kalton et Kasprzyk (1986), ainsi que Little et Rubin (2002) passent en revue diverses procédures d'imputation.

Dans l'une des versions de l'imputation hot deck, la valeur imputée est celle donnée par un répondant appartenant à la même cellule d'imputation, où les cellules d'imputation forment une subdivision exhaustive et disjointe de la population. Dans le cas de l'imputation hot deck aléatoire, des valeurs provenant de répondants appartenant à la même cellule d'imputation sont attribuées au hasard aux non-répondants. L'enregistrement qui fournit la valeur est appelé le *donneur* et celui dans lequel la valeur manque est appelé le *receveur*.

La variance est généralement plus grande pour l'estimateur imputé que pour l'échantillon complet, parce que la non-réponse réduit la taille de l'échantillon et que l'estimateur imputé peut contenir une composante due à l'imputation aléatoire. Rao et Shao (1992) ont proposé pour l'imputation hot-deck une méthode du jackknife ajusté où les unités de la première phase sont sélectionnées avec remise. Rao et Sitter (1995) discutent de la méthode d'estimation de la variance par le jackknife ajusté pour l'imputation par le ratio. Rao (1996) et Sitter (1997) utilisent la méthode du jackknife ajusté dans le cas de l'imputation par la régression. Shao, Chen et Chen (1998) appliquent la notion de Rao et Shao (1992) à la méthode des répliques répétées équilibrées (BRR). Shao et Steel (1999) proposent une estimation de la variance pour les données d'enquête avec imputation composite, où plus d'une méthode d'imputation est utilisée, et introduisent les fractions d'échantillonnage dans les expressions de la variance. Yung et Rao (2000) appliquent la méthode du jackknife ajusté à des estimateurs imputés construits en utilisant un échantillon stratifié a posteriori. Rubin (1987), ainsi que Rubin et Schenker (1986) proposent des méthodes d'imputation multiples. Tollefson et Fuller (1992), ainsi que Särndal (1992) proposent diverses méthodes d'imputation et les estimateurs correspondants de la variance. Kim et Fuller (2004) étudient l'utilisation de l'imputation fractionnée dans le cas du modèle où les observations dans une cellule d'imputation sont indépendantes et de même loi (iid).

Dans le présent article, nous examinons l'imputation hot deck pour une population subdivisée en cellules d'imputation. À la section 2, nous décrivons le modèle de réponse. À la section 3, nous introduisons l'imputation fractionnée entièrement efficace et présentons une méthode d'estimation de la variance pour l'estimateur par imputation, sous l'hypothèse que la probabilité de non-réponse est constante dans une cellule. À la section 4, nous proposons une modification de la méthode entièrement efficace avec utilisation d'un plus petit nombre de donneurs. À la section 5, nous donnons un exemple en vue d'illustrer la mise en œuvre de la méthode proposée. À la section 6, nous exposons les résultats d'une étude en simulation. Enfin, à la dernière section, nous résumons l'étude.

1. Wayne A. Fuller, Department of Statistics, Iowa State University, Ames, IA, 50011, États-Unis; Jae Kwang Kim, Department of Applied Statistics, Yonsei University, Séoul, 120-749, Corée.

2. Conditions de base

Considérons une population de N éléments identifiés par un ensemble d'indices $U = \{1, 2, \ldots, N\}$. À chaque unité i de la population est associée une variable étudiée y_i et un vecteur $\mathbf{x}_i$ de données auxiliaires. L'ensemble de vecteurs, $(y_i, \mathbf{x}_i)$, $i = 1, 2, \ldots, N$, est noté F.

Soit A les indices des éléments d'un échantillon sélectionné d'après un ensemble de règles probabilistes appelées *mécanisme d'échantillonnage*. Soit θ_N la quantité d'intérêt dans la population et $\hat{\theta}$ un estimateur de θ_N, pour l'échantillon complet, linéaire en y et écrivons que

$$\hat{\theta} = \sum_{i \in A} w_i \, y_i. \tag{1}$$

Si w_i est l'inverse de la probabilité de sélection, alors $\hat{\theta}$ est sans biais pour le total de population.

Soit A_R et A_M les ensembles d'indices pour les répondants et les non-répondants dans l'échantillon, respectivement. Définissons la fonction indicateur de réponse

$$R_i = \begin{cases} 1 & \text{si } i \in A_R \\ 0 & \text{si } i \in A_M \end{cases} \tag{2}$$

et posons que $\mathbf{R} = \{(i, R_i); i \in A\}$. La loi de $\mathbf{R}$ est appelée *mécanisme de réponse*.

Supposons que la population finie U soit constituée de G cellules d'imputation, où l'ensemble d'éléments dans la cellule g est U_g. Soit n_g le nombre d'éléments de l'échantillon compris dans la cellule d'imputation g et soit r_g, $r_g > 0$, le nombre de répondants dans la cellule d'imputation g. Supposons que nous ayons le modèle de réponse uniforme dans les cellules, où les r_g réponses dans une cellule sont équivalentes à un échantillon de Poisson tiré avec probabilités égales à partir des n_g éléments.

L'imputation fractionnaire est une méthode consistant à utiliser plus d'un donneur par receveur. Kalton et Kish (1984) ont proposé l'imputation fractionnaire comme méthode d'imputation efficace. Elle a été discutée par Fay (1996). Soit d_{ij} le nombre de fois que y_i est utilisé comme donneur pour la valeur manquante y_j et définissons $\mathbf{d} = \{d_{ij}; i \in A_R, j \in A_M\}$. La loi de $\mathbf{d}$ est appelée *mécanisme d'imputation*. Soit w_{ij}^* le facteur appliqué au poids original de l'élément j quand y_i est utilisé pour cet élément. Pour l'élément j, $j \in A_M$,

$$Y_{Ij} = \sum_{i \in A_R} w_{ij}^* \, y_i \tag{3}$$

est la moyenne pondérée des valeurs pour les répondants. Le facteur w_{ij}^* est appelé *fraction d'imputation*, c'est-à-dire la fraction de la réponse manquante y_j que fournit le donneur i. Notons que $w_{ii}^* = 1$ pour $i \in A_R$ et $w_{ij}^* = 0$ pour $i \neq j$, $i, j \in A_R$. La somme des facteurs d'imputation pour une réponse manquante est contrainte d'être égale à 1,

$$\sum_{i \in A_R} w_{ij}^* = 1, \quad \forall j \in A. \tag{4}$$

Un estimateur ayant les valeurs imputées définies par (3) et un facteur $w_{ij}^* < 1$ est appelé estimateur *par imputation fractionnaire*.

Nous pouvons écrire un estimateur par imputation linéaire en y sous la forme

$$\hat{\theta}_I = \sum_{i \in A_R} \left(\sum_{j \in A} w_j w_{ij}^* \right) y_i \tag{5}$$

$$=: \sum_{i \in A_R} \alpha_i y_i, \tag{6}$$

où la notation A =: B signifie que la définition de B est telle qu'il soit égal à A. La somme des $w_{ij}^* \, w_j$ sur l'ensemble des receveurs pour lesquels i est un donneur (y compris pour lui–même), noté α_i, est le poids total appliqué au donneur i. Si une unité répondante i n'est pas utilisée comme donneur, sauf pour elle–même, alors $\alpha_i = w_i$.

3. Imputation fractionnaire entièrement efficace

Supposons que tous les éléments d'une cellule d'imputation aient la même probabilité de répondre et supposons que les réponses soient indépendantes. Alors, nous pouvons obtenir la loi globale d'un estimateur imputé sous le modèle de réponse en utilisant la structure de probabilité de l'échantillonnage à plusieurs phases, où le modèle de réponse est traité comme étant la deuxième phase du mécanisme d'échantillonnage.

Si les probabilités de réponse dans une cellule sont uniformes, alors un estimateur raisonnable du total est la somme pondérée des estimateurs par le ratio

$$\hat{\theta}_{\mathrm{FE}} = \sum_{g=1}^{G} \left(\sum_{i \in A \cap U_g} w_i \right) \frac{\sum_{i \in A_R \cap U_g} w_i y_i}{\sum_{i \in A_R \cap U_g} w_i}. \tag{7}$$

Dans le contexte de l'échantillonnage à deux phases, Kott et Stukel (1997) ont donné à l'estimateur (7) le nom d'estimateur avec facteur d'extension repondéré. L'estimateur (7) est dit entièrement efficace parce qu'il ne contient aucune variabilité due à la sélection aléatoire des donneurs. Si les w_i sont les mêmes pour tous les éléments d'une cellule, le ratio

$$\left(\sum_{i \in A_R \cap U_g} w_i \right)^{-1} \sum_{i \in A_R \cap U_g} w_i \, y_i \tag{8}$$

est une moyenne simple et, donc, sans biais pour la moyenne de cellule, sachant qu'il existe au moins un répondant dans la cellule. Si les w_i d'une cellule ne sont pas égaux, alors (8) présente un biais de ratio. Il est possible que le nombre d'éléments dans une cellule, n_g, soit positif et

que le nombre de répondants, r_g, soit nul. Quand cela se produit en pratique, les cellules sont regroupées.

Nous pouvons obtenir les propriétés de grand échantillon de l'estimateur pour une série de populations et d'échantillons. Supposons que la population soit composée de G_v cellules disjointes et exhaustives, où v est l'indice de la série. Supposons que la variance d'un estimateur de la moyenne pour l'échantillon complet soit $O(n_v^{-1})$, où n_v est la taille de l'échantillon sélectionné à partir de la v^{e} population. Supposons que les réponses sont indépendantes. Alors, sous des conditions de régularité, nous pouvons nous servir des procédures utilisées par Kim, Navarro et Fuller (2005) dans la preuve de leur théorème 2.1 pour montrer que l'estimateur (7) satisfait

$$\hat{\theta}_{\text{FE}_v} = \hat{\theta}_v + \sum_{g_v=1}^{Gv} \sum_{i \in A_{gv}} w_{iv}(\pi_{gv}^{-1} R_{iv} - 1) e_{iv} + o_p(n_v^{-1/2} N_v), \quad (9)$$

où $e_{iv} = y_{iv} - \bar{Y}_{gv}$, A_{gv} est l'ensemble d'indices d'échantillon dans la g^{e} cellule pour le v^{e} échantillon, $\bar{Y}_{gv}$ est la moyenne de population de la variable y dans la cellule gv de population F_v, π_{gv} est la probabilité qu'un élément dans la cellule gv réponde, et F_v représente la v^{e} population. En outre

$$V(\tilde{\theta}_{\text{FE}v} \mid F_v) = V(\hat{\theta}_v \mid F_v) + E\left\{ \sum_{g_v=1}^{Gv} \pi_{gv}^{-1}(1-\pi_{gv}) \sum_{i \in Agv} w_{iv}^2 e_{iv}^2 \mid F_v \right\}, \quad (10)$$

où

$$\tilde{\theta}_{\text{FE}v} = \hat{\theta}_v + \sum_{g_v=1}^{Gv} \sum_{i \in A_{gv}} w_{iv}(\pi_{gv}^{-1} R_{iv} - 1) e_{iv}.$$

Nous pouvons appliquer l'estimateur (7) en utilisant une imputation fractionnaire dans laquelle chaque unité répondante figurant dans une cellule d'imputation est utilisée comme donneur pour chaque non-répondant compris dans la cellule. Alors, l'estimateur (7) peut s'écrire sous la forme de l'estimateur par imputation fractionnaire

$$\hat{\theta}_{\text{FEFI}} = \sum_{g=1}^{G} \sum_{j \in A \cap U_g} \sum_{i \in A_R \cap U_g} w_j w_{ij}^* y_i, \quad (11)$$

où w_j w_{ij}^* est le poids du donneur i pour le receveur j, w_{ij}^* est la fraction d'imputation du donneur i pour le receveur j définie dans (3), et

$$w_{ij}^* = \begin{cases} \left(\sum_{s \in A_R \cap U_g} w_s\right)^{-1} w_i R_i & \text{si } R_j = 0 \\ 1 & \text{si } R_j = 1 \text{ et } i = j. \end{cases} \quad (12)$$

L'estimateur (11) avec w_{ij}^* donné par (12), qui est algébriquement équivalent à (7), est appelé *estimateur par imputation entièrement efficace* (FEFI pour *fully efficient fractionally imputed*). L'estimateur par imputation fractionnaire a l'avantage de permettre d'estimer directement des fonctions de y, telles que la fraction inférieure à un nombre donné, d'après l'ensemble de données imputées fractionnaires.

Afin d'examiner l'estimation de la variance par rééchantillonnage, posons qu'un estimateur de la variance par rééchantillonnage pour l'échantillon complet est donné par

$$\hat{V}(\hat{\theta}) = \sum_{k=1}^{L} c_k (\hat{\theta}^{(k)} - \hat{\theta})^2, \quad (13)$$

où $\hat{\theta}^{(k)}$ est la k^{e} estimation de θ_N d'après les observations incluses dans la k^{e} réplique, L est le nombre de répliques, et c_k est un facteur associé à la réplique k déterminé par la méthode de rééchantillonnage. Pour une discussion de la répétition des échantillons d'enquête, voir Krewski et Rao (1981), ainsi que Rao, Wu et Yue (1992). Si l'estimateur original $\hat{\theta}$ est un estimateur linéaire de la forme (1), la k^{e} estimation répétée de $\hat{\theta}$ peut s'écrire

$$\hat{\theta}^{(k)} = \sum_{i \in A} w_i^{(k)} y_i, \quad (14)$$

où $w_i^{(k)}$ est le poids de rééchantillonnage de la i^{e} unité de la k^{e} réplique.

Nous proposons pour l'estimateur $\hat{\theta}_{\text{FEFI}}$ la réplique

$$\hat{\theta}_{\text{FEFI}}^{(k)} = \sum_{g=1}^{G} \left(\sum_{i \in A \cap U_g} w_i^{(k)} \right) \frac{\sum_{i \in A_R \cap U_g} w_i^{(k)} y_i}{\sum_{i \in A_R \cap U_g} w_i^{(k)}} =: \sum_{g=1}^{G} \sum_{j \in A \cap U_g} \sum_{i \in A_R \cap U_g} w_j^{(k)} w_{ij}^{*(k)} y_i. \quad (15)$$

Si nous utilisons la réplique (15), nous pouvons écrire l'estimateur de la variance par rééchantillonnage sous la forme

$$\hat{V}_{\text{FEFI}} = \sum_{k=1}^{L} c_k (\hat{\theta}_{\text{FEFI}}^{(k)} - \hat{\theta}_{\text{FEFI}})^2. \quad (16)$$

Les répliques données par (15) peuvent être calculées en deux étapes. Premièrement, nous créons la réplique habituelle en définissant les poids $w_i^{(k)}$ pour chaque élément. Deuxièmement, pour un non-répondant, nous utilisons comme fraction d'imputation par rééchantillonnage du donneur i au receveur j

$$w_{ij}^{*(k)} = \frac{w_i^{(k)}}{\sum_{s \in A_R \cap U_g} w_s^{(k)}}.$$

Notons que la somme des poids de rééchantillonnage fractionnaire des enregistrements donneurs pour chaque receveur est égale au poids de rééchantillonnage de chaque unité dans un échantillon complet.

La méthode proposée est étroitement associée à l'estimateur de la variance de Rao et Shao (1992). Voir aussi Yung et Rao (2000). Toutefois, l'utilisation de l'imputation fractionnaire simplifie beaucoup l'estimation de la variance. Dans la création des répliques, seuls les poids appliqués aux valeurs imputées changent. Il n'est pas nécessaire de recalculer les valeurs imputées et, une fois qu'ils sont calculés, les poids des répliques peuvent être utilisés pour n'importe quelle fonction lisse du vecteur y. En outre, les répliques fractionnaires rendent l'estimateur (16) approprié pour un vecteur de variables y.

Nous pouvons utiliser le théorème 3.1 de Kim, Navarro et Fuller (2005) pour montrer que, étant donné une méthode de production de répliques de l'échantillon complet convergente,

$$\hat{V}_{\text{FEFI}} = V(\tilde{\theta}_{\text{FE}v} \mid F_v) - N_v^{-2} \sum_{g_v=1}^{G_v} \sum_{i \in U_{gv}} \pi_{gv}^{-1}(1-\pi_{gv}) e_{iv}^2 + o_p(n_v^{-1}), \qquad (17)$$

où $\tilde{\theta}_{\text{FE}v}$ est défini dans (10), et où la loi a trait aux mécanismes d'échantillonnage et de réponse.

Si l'on peut ignorer la correction pour population finie, l'estimateur (16) est convergent pour $V\{\hat{\theta}_{\text{FE}}\}$. Si la taille d'échantillon est grande comparativement à N, alors un estimateur de

$$N_v^{-2} \sum_{g_v=1}^{G_v} \sum_{i \in U_{gv}} \pi_{gv}^{-1}(1-\pi_{gv})\, e_{iv}^2$$

devrait être ajouté à (16).

La méthode d'imputation et d'estimation de la variance décrite pour le modèle de réponse produit aussi des estimateurs convergents pour le modèle de moyenne de cellule. Sous ce modèle, les éléments contenus dans une cellule de la population finie sont une réalisation de variables aléatoires indépendantes et de même loi. La méthode d'imputation fondée sur le modèle de réponse n'est pas nécessairement entièrement efficace pour la moyenne de population sous le modèle de moyenne de cellule, mais on peut montrer que l'estimateur de la moyenne et l'estimateur de la variance de la moyenne estimée sont convergents.

4. Approximations de la méthode entièrement efficace

Aux sections précédentes, nous avons construit l'estimateur $\hat{\theta}_{\text{FEFI}}$ de façon à ce que la variance due à l'imputation soit nulle. L'application de la méthode d'imputation fractionnaire, telle qu'elle est décrite en (11), pourrait nécessiter l'utilisation d'un grand nombre de donneurs pour chaque receveur. Par conséquent, nous décrivons une procédure comportant un nombre fixe de donneurs par receveur qui est entièrement efficace pour le total général, mais qui n'est pas forcément entièrement efficace pour les sous-populations. La méthode consiste à affecter des donneurs pour produire une variance faible des valeurs imputées entre receveurs et à modifier la pondération des donneurs pour arriver à l'efficacité complète pour le total.

Supposons que M donneurs soient affectés à chaque receveur. Nous proposons d'affecter les donneurs aux receveurs de façon à approximer la distribution de tous les répondants dans la cellule. L'une des méthodes de sélection possibles consiste à tirer un échantillon stratifié pour chaque receveur. Une autre consiste à recourir à l'échantillonnage systématique avec probabilités proportionnelles aux poids pour sélectionner les donneurs pour chaque receveur. Les fractions initiales w_{ij0}^* sont affectées aux valeurs données. Dans le cas de l'échantillonnage systématique avec poids égaux, la fraction initiale w_{ij0}^* est M^{-1}.

Après avoir affecté les donneurs, nous corrigeons les fractions initiales, w_{ij0}^*, de sorte que la somme des poids donne l'estimateur entièrement efficace de la moyenne de y et que la fonction de distribution cumulative estimée d'après les poids soit une approximation de l'estimateur entièrement efficace de la fonction de distribution cumulative. La modification de la pondération par la régression a été proposée par Fuller (1984, 2003). Chen, Rao et Sitter (2000) discutent d'une méthode d'imputation efficace où l'on modifie les valeurs imputées plutôt que les poids. Soit $\mathbf{z}_{gj} = (z_{g\,j1}, z_{g\,j2}, \ldots, z_{g\,j\alpha})$ un vecteur défini par

$$\begin{aligned} z_{g\,j1} &= y_j \\ z_{g\,j2} &= 1 \quad \text{si } y_j \le L_2 \\ &= 0 \quad \text{autrement} \\ &\vdots \\ z_{g\,j\alpha} &= 1 \quad \text{si } L_{\alpha-1} < y_j \le L_\alpha \\ &= 0 \quad \text{autrement,} \end{aligned}$$

où $L_2, L_3, \ldots, L_\alpha$ divisent la fourchette de valeurs observées de y dans la cellule g en $\alpha - 1$ sections. Le nombre de sections que l'on peut utiliser dépend du nombre et du type d'observations dans la cellule, du nombre de receveurs et du nombre de donneurs par receveur. Si le nombre de donneurs par receveur est grand, il est possible d'ajuster l'ensemble de poids pour chaque receveur de façon à ce que la somme des w_{ij}^* sur i soit égale à l'unité pour chaque j et que la somme des $w_{ij}^* y_i$ sur i soit l'estimateur entièrement efficace pour chaque j. Dans la plupart des cas, les poids sont ajustés de sorte que la somme des w_{ij}^* sur i soit égale à l'unité pour chaque j et que les moyennes de cellule des valeurs imputées soient égales à l'estimateur entièrement efficace.

Soit $\bar{\mathbf{z}}_{\mathrm{FE},\,g}$ l'estimateur entièrement efficace pour la cellule *g*. Si nous utilisons des procédures de régression, les w_{ij}^{*} modifiés pour donner la moyenne de cellule entièrement efficace de $\mathbf{z}$, sont

$$w_{ij}^{*} = w_{ij0}^{*} + (\bar{\mathbf{z}}_{\mathrm{FE},g} - \bar{\mathbf{z}}_{g}^{*})\,\mathbf{S}_{\mathbf{z}\mathbf{z}g}^{-1} w_{ij0}^{*} (\bar{\mathbf{z}}_{g[i]j} - \bar{\mathbf{z}}_{g\cdot j})', \quad (18)$$

où

$$\mathbf{S}_{\mathbf{z}\mathbf{z}g} = \sum_{j\in A_{Lg}} b_j \sum_{i\in A_{Rg}} w_{ij0}^{*} (\mathbf{z}_{g[i]j} - \bar{\mathbf{z}}_{g\cdot j})'(\mathbf{z}_{g[i]j} - \bar{\mathbf{z}}_{g\cdot j}) d_{ij},$$

$$\bar{\mathbf{z}}_{g\cdot j} = \sum_{i\in A_{Rg}} w_{ij0}^{*}\, \mathbf{z}_{g[i]j}\, d_{ij},$$

$$\bar{\mathbf{z}}_{g}^{*} = \sum_{j\in A_{Lg}} b_j \sum_{i\in A_{Rg}} w_{ij0}^{*}\, \mathbf{z}_{g[i]j}\, d_{ij},$$

$$b_j = \left(\sum_{s\in A_{Lg}} w_s \right)^{-1} w_j,$$

A_{Lg} est l'ensemble d'indices des receveurs dans la cellule *g*, $\mathbf{z}_{g[i]j} = \mathbf{z}_{gi}$ est la valeur imputée d'après le donneur *i* au receveur *j*, et $\bar{\mathbf{z}}_{g\cdot j}$ est la moyenne pondérée des valeurs imputées pour le receveur *j* en utilisant le poids initial w_{ij0}^{*}.

Pour estimer la variance, nous créons des répliques de sorte que les poids appliqués aux donneurs reflètent l'effet de la suppression d'un élément sur l'estimateur entièrement efficace. Nous utilisons les mots « suppression » et « supprimer » pour identifier l'élément choisi pour la modification principale du poids pour l'estimation de la variance par rééchantillonnage.

Soit $w_i^{(k)}$ le poids attribué à l'élément *i* pour la k^{e} réplique pour l'estimation de la variance de l'estimateur pour l'échantillon complet. Alors, la réplique pour la moyenne entièrement efficace de *y* pour la cellule *g* est

$$\bar{\mathbf{z}}_{g}^{(k)} = \left[\sum_{i\in A_{Rg}} w_i^{(k)} \right]^{-1} \sum_{i\in A_{Rg}} w_i^{(k)} \mathbf{z}_i. \quad (19)$$

Les fractions de rééchantillonnage sont attribuées aux donneurs dans la cellule *g* de sorte que l'estimation de la moyenne de cellule par rééchantillonnage soit $\bar{\mathbf{z}}_{g}^{(k)}$. Nous assignons les poids fractionnaires initiaux $w_{ij0}^{*(k)}$, où $w_{ij0}^{*(k)}$ est faible, mais positif, si *i* est une unité supprimée pour la réplique *k*. Nous calculons les poids fractionnaires finaux $w_{ij}^{*(k)}$ selon la procédure (18) en remplaçant $\bar{\mathbf{z}}_{\mathrm{FE},g}$ par $\bar{\mathbf{z}}_{g}^{(k)}$ et w_{ij0}^{*} par $w_{ij0}^{*(k)}$. La procédure simule l'effet de la suppression d'un seul élément sur l'estimateur entièrement efficace.

5. Un exemple artificiel

Nous présentons ici un exemple fondé sur des données artificielles afin d'illustrer l'application de la méthode proposée. Supposons que nous observions deux variables d'intérêt, *x* et *y*, dans un échantillon de taille $n = 10$ obtenu par échantillonnage aléatoire simple. La variable *x* est une variable nominale comptant trois catégories, disons 1, 2 et 3, et la variable *y* est une variable continue. Il y a non–réponse partielle pour les deux variables et il existe un ensemble de cellules d'imputation pour chaque variable. Le tableau 5.1 donne les observations sur l'échantillon, où la non–réponse est représentée par *M*. Nous utilisons un poids unitaire pour simplifier la présentation. Nous divisons par dix pour obtenir les poids pour la moyenne.

Tableau 5.1
Un ensemble de données illustratif

Observation	Poids	Cellule pour *x*	Cellule pour *y*	*x*	*y*
1	1	1	1	1	7
2	1	1	1	2	M
3	1	1	2	3	M
4	1	1	1	M	14
5	1	1	2	1	3
6	1	2	1	2	15
7	1	2	2	3	8
8	1	2	1	3	9
9	1	2	2	2	2
10	1	2	1	M	M

Comme la variable *x* est une variable nominale à trois catégories, l'utilisation de trois fractions pour l'imputation fractionnaire donne des estimateurs entièrement efficaces pour la distribution de la variable *x*. Donc, dans le tableau 5.2, les poids pour les trois valeurs imputées de *x* pour la quatrième observation sont les fractions pour les trois catégories dans la cellule 1 pour *x*.

Si l'on utilise un sous–ensemble de donneurs pour chaque receveur, nous suggérons une méthode contrôlée de sélection des donneurs, telle que l'échantillonnage systématique. Dans notre exemple simple, nous pourrions facilement utiliser l'imputation fractionnaire avec les quatre réponses *y* dans la cellule 1, mais afin d'illustrer l'ajustement par la régression, nous n'en utilisons que trois. Voir le tableau 5.2.

Dans la situation où les réponses à deux questions manquent, plusieurs approches sont possibles, y compris la définition d'un troisième ensemble de cellules d'imputation pour ce genre de cas. Étant donné la petite taille de l'échantillon dans notre illustration, nous imputons sous l'hypothèse que *x* et *y* sont indépendantes dans les cellules. Donc, nous imputons quatre valeurs pour l'observation 10. Pour chacune des deux valeurs possibles de *x*, nous imputons deux valeurs possibles de *y*. Nous choisissons l'une des paires de valeurs de *y* imputées de façon qu'elles soient inférieures à la moyenne des réponses et l'autre, de façon à ce qu'elle soit plus grande que la moyenne. Voir les valeurs imputées pour l'observation 10 au tableau 5.2.

Tableau 5.2
Poids fractionnaires pour les moyennes

Observation	Poids	Donneur pour *y*	Cellule pour *x*	Cellule pour *y*	*x*	*y*
1	1,0000		1	1	1	7
2	0,2886	1	1	1	2	7
2	0,3960	6	1	1	2	15
2	0,3154	8	1	1	2	9
3	0,3333	5	1	2	3	3
3	0,3333	7	1	2	3	8
3	0,3334	9	1	2	3	2
4	0,5000		1	1	1	14
4	0,2500		1	1	2	14
4	0,2500		1	1	3	14
5	1,0000		1	2	1	3
6	1,0000		2	1	2	15
7	1,0000		2	2	3	8
8	1,0000		2	1	3	9
9	1,0000		2	2	2	2
10	0,2247	8	2	1	2	9
10	0,2753	4	2	1	2	14
10	0,2095	1	2	1	3	7
10	0,2905	6	2	1	3	15

Nous attribuons une fraction initiale égale à un tiers aux trois valeurs imputées pour les observations 3 et 4, et une fraction initiale égale à un quart aux quatre valeurs imputées pour l'observation 10. Puis, nous ajustons les poids fractionnaires en utilisant la méthode de régression de l'équation (18) pour donner la moyenne par imputation fractionnaire entièrement efficace (FEFI) de *y* comme estimateur, où l'estimateur entièrement efficace de la moyenne de *y* est

$$\bar{y}_{\text{FE}} = \sum_{g=1}^{2} \frac{n_g}{n} \bar{y}_{Rg} = 8{,}4833.$$

Nous contraignons les poids pour l'observation 10 de sorte que les fractions estimées pour les deux catégories de *x* soient les fractions de cellule. Alors, comme la moyenne pondérée de la variable nominale est contrôlée pour chaque individu, le vecteur $\mathbf{z}$ contient uniquement la variable *y*. Le tableau 5.2 donne les poids fractionnaires finaux calculés sous pondération par la régression.

Un analyste peut utiliser l'ensemble de données du tableau 5.2 et tout programme informatique pour échantillon complet pour calculer des estimations des fonctions de *y* et *x*, telles que la moyenne de *y* pour les catégories de *x*. L'ensemble de données fractionnaires est entièrement efficace pour toute fonction de la variable *x* et est également entièrement efficace pour la moyenne de la variable *y*.

Pour l'estimation de la variance par le jackknife, nous répétons le calcul des poids pour chaque réplique. Les estimations répétées des moyennes de cellule de *y* sont données au tableau 5.3 et les estimations répétées des fractions pour les catégories de *x* sont données au tableau 5.4. Nous utilisons les valeurs des tableaux 5.3 et 5.4 comme totaux de contrôle $\bar{\mathbf{z}}_{\text{FE},g}$ dans la pondération par la régression. Nous prenons $w_{ij0}^{*(k)} = 3^{-1}$ comme valeur initiale des fractions de rééchantillonnage pour l'observation 2 et $w_{ij0}^{*(k)} = 4^{-1}$ pour l'observation 10.

Le tableau 5.5 contient les poids jackknife pour l'ensemble de données obtenu par imputation fractionnaire du tableau 5.2. Les poids de rééchantillonnage sont utilisés de la même façon que les répliques pour un échantillon complet. Ils conviennent, avec les mises en garde de la section suivante, pour toute statistique pour laquelle le jackknife avec échantillon complet est approprié. Donc, la procédure est particulièrement séduisante pour un ensemble de données d'usage général, car l'analyste ne doit effectuer aucun calcul supplémentaire.

Nous obtenons l'estimateur entièrement efficace de la moyenne de *y* en considérant que les répondants représentent la deuxième phase d'un échantillon à deux phases. Un estimateur de variance pour échantillon à deux phases peut s'écrire

$$\hat{V} = \frac{1}{n}\sum_{g=1}^{2} \frac{n_g}{n}(\bar{y}_{Rg} - \bar{y}_{\text{FE}})^2 + \sum_{g=1}^{2}\left(\frac{n_g}{n}\right)^2 \frac{1}{r_g} s_{Rg}^2 = 3{,}043,$$

où s_{Rg}^2 est la variance d'échantillon intracellulaire pour la cellule *g*. Si nous utilisons les poids de rééchantillonnage du tableau 5.5, l'estimation de la variance par rééchantillonnage pour la moyenne de *y* est

$$\hat{V}_{\text{JK}}(\bar{y}_{\text{FI}}) = \sum_{k=1}^{10} 0{,}9\,(\bar{y}_{\text{FI}}^{(k)} - \bar{y}_{\text{FI}})^2 = 3{,}078.$$

La différence entre l'estimateur de la variance linéarisé et l'estimateur de la variance par le jackknife est

$$\sum_{g=1}^{2}\left(\frac{r_g}{r_g - 1}\frac{n-1}{n} - 1\right) s_{Rg}^2.$$

Donc, l'estimateur de la variance par le jackknife surestime légèrement la variance réelle dans notre exemple.

Tableau 5.3
Répliques jackknife de la moyenne de cellule de la variable *y*

Cellule	Réplique									
	1	2	3	4	5	6	7	8	9	10
1	12,67	11,25	11,25	10,33	11,25	10,00	11,25	12,00	11,25	11,25
2	4,33	4,33	4,33	4,33	5,00	4,33	2,50	4,33	5,50	4,33

Tableau 5.4
Répliques jackknife de la moyenne de cellule des variables nominales de la variable *x*

Cellule	Niveau de *x*	Réplique									
		1	2	3	4	5	6	7	8	9	10
1	1	0,33	0,67	0,67	0,50	0,33	0,50	0,50	0,50	0,50	0,50
	2	0,33	0,00	0,33	0,25	0,33	0,25	0,25	0,25	0,25	0,25
	3	0,33	0,33	0,00	0,25	0,33	0,25	0,25	0,25	0,25	0,25
2	1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	2	0,50	0,50	0,50	0,50	0,50	0,33	0,67	0,67	0,33	0,50
	3	0,50	0,50	0,50	0,50	0,50	0,67	0,33	0,33	0,67	0,50

Tableau 5.5
Poids jackknife pour l'imputation fractionnaire

Obs.	Réplique									
	1	2	3	4	5	6	7	8	9	10
1	0	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111
2	0,1664	0	0,3206	0,4205	0,3206	0,4563	0,3206	0,2392	0,3206	0,2724
2	0,6559	0	0,4400	0,3002	0,4400	0,2500	0,4400	0,5540	0,4400	0,5075
2	0,2888	0	0,3505	0,3904	0,3505	0,4048	0,3505	0,3179	0,3505	0,3312
3	0,3706	0,3706	0	0,3706	0,3226	0,3706	0,5018	0,3706	0,2867	0,3706
3	0,3697	0,3697	0	0,3697	0,5018	0,3697	0,0090	0,3697	0,6004	0,3697
3	0,3708	0,3708	0	0,3708	0,2867	0,3708	0,6003	0,3708	0,2240	0,3708
4	0,3703	0,7407	0,7407	0	0,3703	0,5556	0,5556	0,5556	0,5556	0,5556
4	0,3704	0	0,3704	0	0,3704	0,2777	0,2777	0,2777	0,2777	0,2777
4	0,3704	0,3704	0	0	0,3704	0,2778	0,2778	0,2778	0,2778	0,2778
5	1,1111	1,1111	1,1111	1,1111	0	1,1111	1,1111	1,1111	1,1111	1,1111
6	1,1111	1,1111	1,1111	1,1111	1,1111	0	1,1111	1,1111	1,1111	1,1111
7	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	0	1.,111	1,1111	1,1111
8	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	0	1,1111	1,1111
9	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	1,1111	0	1,1111
10	0,1624	0,2777	0,2777	0,3061	0,2777	0,2286	0,3474	0,3013	0,1520	0
10	0,3931	0,2778	0,2778	0,2494	0,2778	0,1417	0,3934	0,4395	0,2185	0
10	0,0932	0,2778	0,2778	0,3231	0,2778	0,4400	0,1483	0,0746	0,3171	0
10	0,4623	0,2778	0,2778	0,2324	0,2778	0,3008	0,2220	0,2957	0,4235	0

6. Études par simulation

6.1 Paramètres d'intérêt

Pour étudier les propriétés de la méthode d'imputation, nous avons réalisé une étude de Monte-Carlo. L'échantillon est stratifié, avec deux éléments par strate et deux cellules d'imputation, où les cellules recoupent les strates. La cellule 1 comprend 20 % de la population des strates 1 à 25 et 80 % de la population des strates 26 à 50. La probabilité de réponse est 0,7 pour la cellule 1 et 0,5 pour la cellule 2. Nous examinons deux variables. La variable D est toujours observée et définit une sous-population. La probabilité que $D = 1$ est de 0,25 pour la cellule 1 et de 0,40 pour la cellule 2. La variable y est sujette à la non–réponse avec probabilités de réponse dans les cellules constantes. La variable D est indépendante de y et de la probabilité de réponse. La variable y suit une loi normale, où les paramètres pour une population de 50 strates sont donnés au tableau 5.1. Dans le modèle de génération des données du tableau 6.1, il n'existe aucun effet de strate. Les paramètres d'intérêt sont : $\theta_1 =$ moyenne de y, $\theta_2 =$ moyenne de y pour $D = 1$, $\theta_3 =$ fraction de Y inférieure à deux, $\theta_4 =$ fraction de Y inférieure à un.

Tableau 6.1
Ensemble de paramètres A

		Cellule 1		Cellule 2	
Strates	Poids de l'élément	Moyenne	Variance	Moyenne	Variance
1 à 25	0,01	0,4	0,36	1,6	0,36
26 à 50	0,01	0,4	0,36	1,6	0,36

6.2 Méthodes d'estimation

Dans la simulation, nous avons utilisé $M = 5$ et $M = 3$ donneurs par receveur. Nous avons sélectionné des échantillons systématiques à titre de donneurs pour chaque receveur. Si le nombre de répondants dans la cellule est inférieur à M, chaque répondant est utilisé comme donneur pour chaque receveur et les w_{ij}^* sont proportionnels au poids w_i original des répondants. Si le nombre de répondants dans la cellule est supérieur à M, nous classons les donneurs par taille et les numérotons de 1 à r_g. Puis, nous plaçons les donneurs dans l'ordre , 3, 5, $\ldots, r_g, r_{g-1}, r_{g-3}, \ldots, 2$ pour les

valeurs impaires de r_g et dans l'ordre $1, 3, 5, \ldots, r_{g-1}, r_g, r_{g-2}, \ldots, 2$ pour les valeurs paires de r_g. Ensuite, nous calculons les sommes cumulées des poids et sélectionnons m_g échantillons systématiques de taille M, où $m_g = n_g - r_g$. Les sommes cumulées sont normalisées de sorte que la somme générale soit égale à l'unité, le nombre R_{Ng}, compris entre 0 et $0, 2m_g$, est sélectionné aléatoirement et les m_g échantillons sont les échantillons systématiques de taille M définis par le donneur associé à $R_{Ng} + 0,2(s-1) + (t-1)m_g^{-1}$, $s = 1, 2, 3, 4, 5$ pour les receveurs $t = 1, 2, \ldots, m_g$. Pour chaque donneur, la fraction d'imputation initiale est $w_{ij}^* = M^{-1}$.

Les fractions d'imputation initiale sont modifiées en utilisant la méthode de régression (18). Nous avons ordonné les donneurs dans une cellule du plus petit au plus grand et formé la somme cumulée des poids. Soit

$$S_{g,wt} = \sum_{i=1}^{t} w_{[i]}, i \in A_{Rg}, \tag{20}$$

où $w_{[i]}, i = 1, 2, \ldots, r_g$ est le poids de $y_{g,(i)}$ et les $y_{g,(1)} \le \ldots \le y_{g,(n)}$ sont les valeurs ordonnées de y dans la cellule g. Pour définir les bornes des groupes qu'il convient d'utiliser pour créer des fonctions indicateurs, posons que t_{*s} est le t pour lequel

$$\max \{S_{g,wt} : S_{g,wt} \le 0,2sS_{gw}\}$$

pour $s = 1, 2, 3, 4$, où S_{gw} est le total des poids pour les donneurs compris dans la cellule g. Définissons

$$\begin{aligned} z_{gi,s+1} &= 1 \quad \text{si } y_i \le y_{g,(t_{*s})} \text{ et } i \in A_{Rg} \\ &= 0 \quad \text{autrement} \end{aligned} \tag{21}$$

pour $s = 1, 2, 3, 4$ et soit $\mathbf{z}_{gj} = (y_{gj1}, z_{gj2}, \ldots, z_{gj5})$. L'estimateur par imputation, modifié par la régression, de la moyenne pour chacune des cinq variables comprises dans le vecteur z est l'estimateur entièrement efficace de la moyenne pertinente.

L'estimateur entièrement efficace (FE) avec unité supprimée pour la réplique k de la moyenne de cellule de $\mathbf{z}$ est défini en (19). Le poids fractionnaire initial du donneur k à l'élément j est fixé à $w_{kj0}^{*(k)} = 0,01w_{kj}^*$. Ce poids initial assure que le poids final soit faible, mais permet l'ajustement par la régression. Les poids finaux $w_{ij0}^{*(k)}$ sont calculés par la procédure de régression (18) en utilisant le poids initial $w_{ij}^{*(k)}$.

6.3 Résultats de l'étude de Monte-Carlo

Les résultats de Monte-Carlo pour les 5 000 échantillons générés par les paramètres du tableau 6.1 sont donnés aux tableaux 6.2 et 6.3. Nous présentons les résultats pour l'échantillon complet, pour l'imputation fractionnaire avec cinq donneurs, pour l'imputation fractionnaire avec trois donneurs et pour l'imputation multiple (MI) en utilisant le bootstrap bayésien approximatif (ABB) de Rubin et Schenker (1986) avec $M = 5$ et avec $M = 3$. Les procédures FI et MI sont toutes deux sans biais pour les quatre paramètres du tableau 6.2. La dernière colonne de ce tableau donne la variance de Monte-Carlo de l'estimateur divisée par la variance de Monte-Carlo de la procédure FI avec $M = 5$, exprimée en pourcentage. La procédure FI est de 5 % à 10 % plus efficace que la procédure MI avec $M = 5$ et de 9 % à 13 % plus efficace que la procédure MI avec $M = 3$.

Sous le modèle, la moyenne des valeurs observées n'est pas le meilleur estimateur de la moyenne de domaine. Dans cet exemple, l'estimateur FI est presque aussi efficace que l'estimateur pour l'échantillon complet. L'effet d'un nombre plus petit d'observations est compensé par l'utilisation d'un meilleur estimateur de la moyenne pour le domaine. Sous le modèle, l'indicateur de domaine est indépendant des valeurs de y, sachant la cellule. Par conséquent, il est efficace d'utiliser toutes les valeurs contenues dans la cellule comme donneurs, plutôt que simplement les répondants dans le domaine.

Les propriétés des estimateurs de la variance sont données au tableau 6.3. La colonne intitulée « moyenne relative » donne la moyenne estimée de Monte-Carlo des variances estimées divisées par la variance estimée de Monte-Carlo, où cette dernière est donnée au tableau 6.2. Les deux méthodes d'estimation de la variance semblent être quasiment sans biais pour la variance de la moyenne. La variance relative de l'estimateur de variance MI pour $M = 5$ est égale à près de deux fois celle de l'estimateur de variance FI pour $M = 5$. Pour $M = 3$, l'estimateur de variance MI vaut plus de trois fois l'estimateur de variance FI. La variance de l'estimateur de variance MI est grande, parce que la variance due aux observations manquantes est estimée avec quatre degrés de liberté pour $M = 5$ et avec deux degrés de liberté pour $M = 3$.

L'estimateur de variance MI de la moyenne de domaine est gravement biaisé. Cette propriété a été reconnue pour la première fois par Fay (1991, 1992) et étudiée par Meng (1994), ainsi que par Wang et Robins (1998). L'estimateur de variance FI pour la moyenne de domaine présente aussi un biais positif, quoique nettement plus faible que celui de MI. Nous pouvons réduire le biais dans l'estimateur de variance FI en augmentant M, mais le biais de MI dépend peu de M.

Tous les estimateurs de variance de $\hat{\theta}_4$ présentent un léger biais négatif. Nous pensons que l'estimateur FI est légèrement biaisé pour $\hat{\theta}_4$ parce que, bien que nous utilisions le vecteur z, les poids sont légèrement lissés par la procédure de régression. Il est connu que l'imputation multiple (MI) donne lieu à un biais de petit échantillon. Voir Kim (2002).

Tableau 6.2
Moyenne et variance des estimateurs ponctuels sous les conditions A (5 000 échantillons de taille 100)

Paramètre	Scénario d'imputation	Moyenne	Variance	Variance relative à FI (%)
Moyenne	Échantillon complet	1,00	0,00570	67
(θ_1)	FI(3)	1,00	0,00849	100
	ABB(3)	1,00	0,00926	109
	FI(5)	1,00	0,00849	100
	ABB(5)	1,00	0,00903	106
Moyenne de domaine	Échantillon complet	1,14	0,02020	99
(θ_2)	FI(3)	1,14	0,02050	100
	ABB(3)	1,14	0,02230	109
	FI(5)	1,14	0,02040	100
	ABB(5)	1,14	0,02170	106
$\Pr(Y < 2)$	Échantillon complet	0,87	0,00104	51
(θ_3)	FI(3)	0,87	0,00202	100
	ABB(3)	0,87	0,00228	113
	FI(5)	0,87	0,00202	100
	ABB(5)	0,87	0,00223	110
$\Pr(Y < 1)$	Échantillon complet	0,50	0,00208	66
(θ_4)	FI(3)	0,50	0,00313	100
	ABB(3)	0,50	0,00342	109
	FI(5)	0,50	0,00313	100
	ABB(5)	0,50	0,00329	105

Tableau 6.3
Moyenne relative, statistique *t* et variance relative pour les estimateurs de variance sous les conditions A (5 000 échantillons de taille 100)

Paramètre	Méthode	Moyenne relative (%)**	Statistique *t**	Variance relative (%)
Moyenne	FI(3)	100,1	0,05	5,66
(θ_1)	ABB(3)	99,6	−0,19	19,25
	FI(5)	100,1	0,03	5,65
	ABB(5)	98,2	−0,89	9,95
Moyenne de domaine	FI(3)	115,9	7,54	13,88
(θ_2)	ABB(3)	127,9	12,72	28,88
	FI(5)	106,6	3,14	11,62
	ABB(5)	128,4	13,43	20,03
$\Pr(Y < 2)$	FI(3)	103,9	1,86	13,90
(θ_3)	ABB(3)	100,8	0,36	48,42
	FI(5)	101,7	0,82	12,07
	ABB(5)	98,5	−0,67	25,10
$\Pr(Y < 1)$	FI(3)	98,5	−0,75	4,67
(θ_4)	ABB(3)	96,3	−1,80	18,51
	FI(5)	97,6	−1,20	4,45
	ABB(5)	96,7	−1,65	10,17

* Statistique pour l'hypothèse selon laquelle la variance estimée est sans biais.
** Moyenne de Monte-Carlo des estimations de variance divisée par la variance de Monte-Carlo des estimations, en pourcentage.

Dans un deuxième ensemble de paramètres, noté *C*, les moyennes étaient les suivantes :

Cellule 1 des strates 1 à 25; $\mu = 0{,}4$

Cellule 1 des strates 26 à 50; $\mu = 3{,}0$

Cellule 2 des strates 1 à 25; $\mu = 1{,}6$

Cellule 2 des strates 26 à 50; $\mu = 2{,}2$.

Tous les autres paramètres sont les mêmes que dans l'ensemble de paramètres A. Les propriétés des estimateurs sont données au tableau 6.4. L'imputation fractionnaire (FI) et l'imputation multiple (MI) produisent toutes deux des estimations sans biais des moyennes et de la moyenne de domaine. Comme pour l'ensemble de paramètres A, la procédure FI est de 8 % à 12 % plus efficace que la procédure MI pour $M = 5$ et de 14 % à 16 % plus efficace pour $M = 3$.

Les hypothèses requises pour l'estimation de variance MI ne sont pas satisfaites pour l'ensemble de paramètres C. Par conséquent, la variance MI estimée est fortement biaisée pour tous les paramètres. Voir le tableau 6.5. Pour $M = 5$, le biais dans la variance MI estimée est d'environ 17 % pour la variance de la moyenne globale et de près de 50 % pour la moyenne de domaine. Le biais de la variance MI de la moyenne est plus faible pour une variable binomiale que pour une variable continue, parce que l'effet de stratification est plus faible dans le premier cas.

Les propriétés des variances estimées pour la procédure FI sont semblables à celles obtenues pour l'ensemble de paramètres A. La variance de la moyenne de domaine présente un biais positif d'environ 23 % pour $M = 3$ et d'environ 6 % pour $M = 5$.

La variance de l'estimation de la variance MI est de 2,4 à 3,5 fois plus élevée que la variance de l'estimation de la variance FI pour $M = 5$ et de 3 à 7 fois plus élevée pour $M = 3$, ce qui démontre la supériorité nette de l'estimateur de variance FI pour cette configuration.

Tableau 6.4
Moyenne et variance des estimateurs ponctuels sous les conditions C (5 000 échantillons de taille 100)

Paramètre	Scénario d'imputation	Moyenne	Variance	Variance relative à FI (%)
Moyenne	Échantillon complet	2,10	0,00500	48
(θ_1)	FI(3)	2,10	0,01050	100
	ABB(3)	2,10	0,01220	116
	FI(5)	2,10	0,01050	100
	ABB(5)	2,10	0,01150	110
Moyenne de domaine	Échantillon complet		0,02530	102
(θ_2)	FI(3)	2,01	0,02510	101
	ABB(3)	2,01	0,02850	115
	FI(5)	2,01	0,02480	100
	ABB(5)	2,01	0,02710	109
$\Pr(Y < 2)$	Échantillon complet		0,00127	45
(θ_3)	FI(3)	0,45	0,00281	100
	ABB(3)	0,45	0,00322	115
	FI(5)	0,45	0,00280	100
	ABB(5)	0,45	0,00314	112
$\Pr(Y < 1)$	Échantillon complet		0,00107	54
(θ_4)	FI(3)	0,15	0,00199	100
	ABB(3)	0,15	0,00226	114
	FI(5)	0,15	0,00199	100
	ABB(5)	0,15	0,00214	108

Tableau 6.5
Moyenne relative, statistique *t* et variance relative pour les estimateurs de variance sous les conditions C (5 000 échantillons de taille 100)

Paramètre	Méthode	Moyenne relative (%)	Statistique *t**	Variance relative (%)
Moyenne	FI(3)	100,9	0,41	6,42
(θ_1)	ABB(3)	116,7	7,31	40,14
	FI(5)	100,8	0,39	6,42
	ABB(5)	117,1	7,99	22,29
Moyenne de domaine	FI(3)	122,7	10,78	16,23
(θ_2)	ABB(3)	144,4	19,79	46,05
	FI(5)	106,1	2,95	11,95
	ABB(5)	148,7	22,51	32,49
$\Pr(Y < 2)$	FI(3)	104,4	2,18	6,63
(θ_3)	ABB(3)	114,7	6,54	42,32
	FI(5)	101,8	0,89	6,42
	ABB(5)	112,1	5,74	20,67
$\Pr(Y < 1)$	FI(3)	102,3	1,13	11,08
(θ_4)	ABB(3)	101,3	0,58	39,14
	FI(5)	99,9	–0,04	10,05
	ABB(5)	102,2	1,04	23,60

* Statistique pour l'hypothèse selon laquelle la variance estimée est sans biais.

7. Résumé

Dans l'imputation fractionnaire, plusieurs donneurs sont utilisés pour chaque valeur manquante et une fraction du poids du non-répondant est attribuée à chaque donneur. Si l'on utilise tous les donneurs, la procédure est entièrement efficace, sous le modèle, pour toutes les fonctions d'un vecteur **y**. Nous montrons que l'utilisation de l'imputation fractionnaire avec un petit nombre d'imputations par non-répondant peut donner un estimateur entièrement efficace de la moyenne. Les estimations d'autres paramètres, comme les estimations de la distribution cumulative sont presque entièrement efficaces.

L'imputation fractionnaire permet de construire des répliques d'usage général pour l'estimation de la variance. Il est possible d'utiliser un seul ensemble de répliques pour estimer la variance dans le cas de variables imputées, de variables observées sur l'ensemble des répondants et, sous les hypothèses du modèle, pour des fonctions de deux types de variables. Les répliques donnent des estimations des variances des moyennes de domaine dont le biais est nettement plus faible que celui des estimations par imputation multiple. Le biais tend vers zéro quand la valeur de M augmente et, dans la simulation, est modéré pour $M = 5$. L'estimateur de la variance par rééchantillonnage est facile à appliquer au moyen d'un logiciel de rééchantillonnage, tel que Wesvar.

L'imputation fractionnaire avec un nombre fixe de donneurs par receveur est un peu plus efficace pour la moyenne que l'imputation multiple avec le même nombre de donneurs. L'imputation fractionnaire donne des estimations de variance dont le biais est plus faible et dont la variance est nettement plus faible que les estimateurs par imputation multiple avec le même nombre d'imputations.

8. Remerciements

La présente étude a été financée partiellement aux termes d'un sous-contrat entre Westat et la Iowa State University en vertu du contrat n° ED-99-CO-0109 établi entre Westat et le Department of Education, ainsi que du contrat de coopération 13-3AEU-0-80064 conclu entre la Iowa State University, le U.S. National Agricultural Statistics Service et le U.S. Bureau of the Census. Nous remercions Jean Opsomer et Damiao Da Silva de leurs commentaires constructifs.

Bibliographie

Chen, J., Rao, J.N.K. et Sitter, R.R. (2000). Efficient random imputation for missing data in complex surveys. *Statistica Sinica*, 10, 1153-1169.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of Bureau of the Census Annual Research Conference*, American Statistical Association, 429-440.

Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section*, American Statistical Association, 227-232.

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

Ford, B.M. (1983). An overview of hot-deck procedures. Dans *Incomplete Data in Sample Surveys*, Vol. 2. New York: Academic Press, 185-207.

Fuller, W.A. (1984). Application de la méthode des moindres carrés et de techniques connexes aux plans de sondage complexes. *Techniques d'enquête*, 10, 107-130.

Fuller, W.A. (1996). *Introduction to Statistical Time Series*. 2ième édition. New York: John Wiley & Sons, Inc.

Fuller, W.A. (2003). Estimation for multiple phase samples. Dans *Analysis of Survey Data*, (Éds. R.L. Chambers et C.J. Shinner). Wiley, Chichister, England, 307-322.

Kalton, G., et Kasprzyk, D. (1986). Le traitement des données d'enquêtes manquantes. *Techniques d'enquête*, 12, 1-17.

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. (2002). A note on approximate Bayesian bootstrap imputation. *Biometrika*, 89, 470-477.

Kim, J.K., et Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.

Kim, J.K., Navarro, A. et Fuller, W.A. (2005). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, à paraître.

Kott, P.S., et Stukel, D.M. (1997). La méthode du jackknife convient-elle à un échantillon à deux phases? *Techniques d'enquête*, 23, 89-98.

Krewski, D., et Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9, 1010-1019.

Little, R.J.A., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. 2ième édition. New York: John Wiley & Sons, Inc.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9, 538-573.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.

Rao, J.N.K., et Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

Rao, J.N.K., et Sitter, R.R. (1995). Variance estimation under two-phase sampling with applications to imputation for missing data *Biometrika*, 82, 453-460.

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

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

Rubin, D.B. (1987). *Multiple Imputation For Nonresponse In Surveys*. New York: John Wiley & Sons, Inc.

Sande, I.G. (1983). Hot-deck imputation procedures. *Incomplete Data in Sample Surveys*, Vol. 3. New York: Academic Press, 339-349.

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., Chen, Y. et Chen, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

Shao, J., et Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

Tollefson, M., et Fuller, W.A. (1992). Variance estimation for sampling with random imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 140-145.

Wang, N., et Robins, J.M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85, 935-948.

Yung, W., et Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of American Statistical Association*, 95, 903-915.

Techniques d'enquête, décembre 2005
Vol. 31, N° 2, pp. 165-174
Statistique Canada, N° 12-001-XPB au catalogue

Estimation de la variance avec imputation hot deck : Une étude par simulation de trois méthodes

J. Michael Brick, Michael E. Jones, Graham Kalton et Richard Valliant [1]

Résumé

Les méthodes d'estimation de la variance des estimations par sondage applicables à des données complètes sont biaisées lorsque certaines données sont imputées. Nous recourons à la simulation pour comparer l'efficacité de la méthode assistée par modèle, de la méthode du jackknife ajusté et de la méthode d'imputation multiple pour estimer la variance d'un total quand les réponses à certaines questions ont été imputées par la méthode hot deck. La simulation vise à étudier les propriétés des estimations de la variance des estimations imputées de totaux pour la population dans son ensemble et pour certains domaines provenant d'un plan d'échantillonnage stratifié non proportionnel à un degré quand les hypothèses sous-jacentes, comme l'absence de biais dans l'estimation ponctuelle et l'hypothèse des réponses manquantes au hasard dans les cellules hot deck, ne sont pas vérifiées. Les estimateurs de la variance des estimations pour l'ensemble de la population produisent des intervalles de confiance dont le taux de couverture s'approche du taux nominal, même en cas d'écarts modestes par rapport aux hypothèses, mais il n'en est pas ainsi des estimations par domaine. La couverture est surtout sensible au biais dans les estimations ponctuelles. Comme le démontre la simulation, même si une méthode d'imputation donne des estimations presque sans biais pour la population dans son ensemble, les estimations par domaine peuvent être fort biaisées.

Mots clés : Jackknife ajusté; estimation par domaine; estimation de la variance assistée par modèle; imputation multiple; non-réponse.

1. Introduction

L'imputation est un moyen fréquemment utilisé dans les recherches par sondage pour remplacer les réponses manquantes à certaines questions, de façon à produire des ensembles de données complets pour la diffusion au grand public ou l'analyse générale. Il est généralement reconnu que traiter des valeurs imputées comme s'il s'agissait de valeurs observées produit un biais par défaut dans les estimations de la variance des estimations par sondage. Par conséquent, le taux de couverture des intervalles de confiance est inférieur au taux nominal. Le biais dans les estimations de la variance a tendance à croître avec le taux de non-réponse partielle et peut être considérables si ce taux est élevé.

Nous étudions ici trois méthodes mises au point pour estimer la variance en présence de données imputées, à savoir une méthode assistée par modèle (Särndal 1992), une méthode du jackknife ajusté (Rao et Shao 1992) et une méthode d'imputation multiple (Rubin 1987). Chacune de ces méthodes a été évaluée théoriquement et par des méthodes de simulation, principalement sous des conditions concordant avec les hypothèses des méthodes. Dans la présente étude, nous utilisons la simulation pour comparer les trois méthodes dans des conditions expérimentales identiques sous lesquelles certaines hypothèses que requièrent les méthodes ne tiennent pas. L'objectif est d'examiner la performance relative des méthodes dans des situations susceptibles de se produire en pratique. D'autres études en simulation des méthodes d'estimation de la variance avec données imputées ont généralement été plus limitées. Même l'étude par simulation de plus grande portée réalisée par Lee, Rancourt et Särndal (2001) était basée sur de petites populations et n'incluait pas l'imputation multiple.

Nous utilisons un échantillon stratifié non proportionnel à un degré, sélectionné à partir d'un ensemble de données de population réelles, pour évaluer ces méthodes d'estimation de la variance dans des conditions réalistes. Nous attribuons les valeurs imputées par la méthode hot deck, qui est l'une des méthodes d'imputation les plus populaires en recherche par sondage. Puisque l'imputation hot deck est une forme d'imputation par la régression (Kalton et Kaspryzk 1986), restreindre l'étude par simulation à la méthode hot deck n'est pas une caractéristique critique en ce qui concerne l'étude des effets sur l'estimation de la variance. Nous étudions l'estimation pour les totaux de population, ainsi que pour les totaux de domaine. Dans le cas des estimations par domaine, nous supposons que l'indicateur de domaine est connu pour tous les membres de l'échantillon.

Dans les simulations, nous utilisons trois combinaisons distinctes de mécanismes de génération de données manquantes et de formation de cellules hot deck pour évaluer les propriétés des méthodes d'estimation de la variance sous des conditions qui violent à des degrés divers les hypothèses

1. J. Michael Brick, Michael E. Jones et Graham Kalton, Westat, 1650 Research Boulevard, Rockville, MD 20850; Richard Valliant, University of Michigan, 1218 Lefrak Hall, College Park, MD 20742.

qui sous-tendent les méthodes. Les trois méthodes d'estimation de la variance que nous étudions reposent chacune sur l'hypothèse que les données manquent au hasard dans chaque cellule hot deck. En outre, la méthode assistée par modèle (MA) et la méthode d'imputation multiple (MI) reposent sur l'hypothèse qu'un modèle simple avec moyenne et variance communes est vérifié dans chaque cellule. L'étude de la robustesse des méthodes d'estimation de la variance est un aspect important de la simulation, parce qu'en pratique, les hypothèses qui sous- tendent les méthodes ne sont pour ainsi dire jamais entièrement satisfaites.

À la section suivante, nous décrivons brièvement trois méthodes d'estimation de la variance avec données imputées par la méthode hot deck. À la troisième section, nous décrivons la population étudiée, le plan d'échantillonnage utilisé dans les simulations et les méthodes appliquées pour générer les données manquantes et mettre en œuvre les imputations hot deck. À la quatrième section, nous donnons les résultats des simulations. Enfin, à la dernière section, nous présentons certaines conclusions concernant les méthodes et leur applicabilité.

2. Description des méthodes d'estimation de la variance

Nous représentons l'échantillon complet par *A*, le sous-ensemble qui répond à une question par A_R, et le sous-ensemble qui ne répond pas à la question par A_M. Pour les imputations, nous répartissons les unités en cellules hot deck portant l'indice $g = 1, \ldots, G$, où le sous-ensemble de n_{Rg} répondants dans la cellule g est A_{Rg}, et le sous- ensemble de non-répondants est A_{Mg}. Pour chaque unité pour laquelle une valeur manque, la méthode hot deck consiste à sélectionner aléatoirement dans la même cellule hot deck un répondant qui deviendra le donneur de la valeur imputée.

Sous imputation hot deck, les donneurs sont souvent sélectionnés dans une cellule par échantillonnage aléatoire simple avec remise (easar), par échantillonnage aléatoire simple sans remise (eassr) ou par échantillonnage avec probabilité proportionnelle au poids de sondage avec remise (epppsar). Puisque les résultats de la simulation obtenus en utilisant les méthodes easar et epppsar sont fort semblables, seuls les résultats pour la seconde–nommée hot deck pondérée–sont présentés ici. L'estimateur imputé d'un total de population est $\hat{\theta}_I = \sum_{i \in A_R} w_i y_i + \sum_{i \in A_M} w_i y_i^*$, où w_i est le poids de sondage, y_i est la valeur déclarée et y_i^* est la valeur imputée pour l'unité *i* dans l'ensemble de non-répondants.

2.1 Estimation de la variance assistée par modèle

L'approche assistée par modèle (MA) avec imputation hot deck repose sur l'hypothèse que les données manquent aléatoirement dans les cellules et qu'un modèle pour la génération des *y* est vérifié. Un modèle naturel en cas d'imputation hot deck est que les y_i sont générés de façon indépendante et identique dans les cellules hot deck, c'est-à-dire, $y_{gi} \overset{\text{iid}}{\sim} (\mu_g, \sigma_g^2)$ pour la cellule *g*. Sous l'approche assistée par modèle, les inférences dépendent de la validité des hypothèses du modèle.

Särndal (1992) a décomposé la variance totale de l'estimateur imputé en trois composantes notées V_{SAM}, V_{IMP} et V_{MIX}. Les estimateurs utilisés pour ces composantes dans les simulations sont ceux donnés dans Brick, Kalton et Kim (2004). L'estimateur MA de la variance est égal à la somme des estimations composantes : $\hat{V}_{\text{MA}} = \hat{V}_{\text{SAM}} + \hat{V}_{\text{IMP}} + 2\hat{V}_{\text{MIX}}$. Les estimateurs $\hat{V}_{\text{IMP}}$ et $\hat{V}_{\text{MIX}}$ requièrent un estimateur de la variance élémentaire dans chaque cellule hot deck. Puisque les simulations ont indiqué que la différence entre les estimateurs pondéré et non pondéré était faible, nous ne discutons que de l'estimateur pondéré de σ_g^2, c'est-à-dire $\hat{\sigma}_g^2 = n_{Rg}(n_{Rg} - 1)^{-1} \sum_{A_{Rg}} w_i (y_i - \bar{y}_{Rg})^2 \times (\sum_{A_{Rg}} w_i)^{-1}$, avec $\bar{y}_{Rg} = \sum_{A_{Rg}} w_i y_i (\sum_{A_{Rg}} w_i)^{-1}$.

2.2 Estimation de la variance par le jackknife ajusté

L'estimateur de la variance par le jackknife ajusté (AJ) de Rao et Shao (1992) pour un échantillon stratifié avec imputation et facteurs de correction pour population finie (*fcpf*) ignorable est

$$\hat{V}(\hat{\theta}_I) = \sum_{h=1}^{L} \sum_{k=1}^{n_h} \frac{n_h - 1}{n_h} (\hat{\theta}_{Ih}^{(k)} - \hat{\theta}_I)^2,$$

où n_h est le nombre d'unités échantillonnées dans la strate *h*,

$$\hat{\theta}_{Ih}^{(k)} = \sum_{g=1}^{G} \left\{ \sum_{(hi) \in A_{Rg}} w_{hi}^{(k)} y_{hi} + \sum_{(hj) \in A_{Mg}} w_{hj}^{(k)} (y_{hj}^* + \hat{y}_{Rg}^{(k)} - \bar{y}_{Rg}) \right\}$$

est l'estimateur ajusté quand l'unité *k* est omise,

$$\hat{y}_{Rg}^{(k)} = \sum_{(hi) \in A_{Rg}} w_{hi}^{(k)} y_{hi} \Big/ \sum_{(hi) \in A_{Rg}} w_{hi}^{(k)},$$

$$\bar{y}_{Rg} = \sum_{(hi) \in A_{Rg}} w_{hi} y_{hi} \Big/ \sum_{(hi) \in A_{Rg}} w_{hi},$$

$w_{hi}^{(k)}$ est le poids de l'unité *hi* ajusté pour tenir compte de l'omission de l'unité *k*. La notation $(hi) \in B$ dénote que l'unité *i* dans la strate *h* fait partie de l'ensemble *B*. Cette procédure requiert le calcul de $\sum n_h$ estimations répétées, $\hat{\theta}_{Ih}^{(k)}$. Une stratégie utilisée fréquemment pour réduire les calculs consiste à combiner les unités en strates de variance (par exemple, voir Rust et Rao 1996). Soit h^* une strate de variance combinée et *k* un groupe d'unités d'échantillonnage dans la strate combinée. Toutes les unités

échantillonnées sont affectées à l'un des groupes. Alors, l'estimateur de la variance par le jackknife ajusté groupé est

$$\hat{V}_{AJ} = \sum_{h^*} \sum_{k=1}^{n_{h^*}} \frac{n_{h^*(k)}}{n_{h^*}} (\hat{\theta}_{Ih^*}^{(k)} - \hat{\theta}_I)^2,$$

où n_{h^*} est le nombre d'unités d'échantillonnage dans la strate de variance combinée h^*, $n_{h^*(k)}$ est le nombre d'unités retenues dans la strate h^* quand les unités du groupe k sont supprimées et, correspondant à $\hat{\theta}_{Ih}^{(k)}$, $\hat{\theta}_{Ih^*}^{(k)}$ est l'estimation imputée ajustée pour l'ensemble de la population quand les unités du groupe k dans la strate h^* sont supprimées. Les unités retenues provenant de la strate de plan de sondage h qui figurent dans la strate de variance combinée h^* reçoivent un poids de rééchantillonnage $w_{h^*i}^{(k)} = n_{h^*}(n_{h^*(k)})^{-1} w_{h^*i}$.

La méthode AJ repose sur l'hypothèse d'un modèle à probabilité de réponse uniforme dans chaque cellule hot deck, mais, contrairement à la méthode MA, ne nécessite pas d'hypothèse concernant la loi de distribution. Sous le modèle de probabilité de réponse uniforme sans hypothèse concernant la loi de distribution, l'utilisation d'une méthode hot deck pondérée est nécessaire pour produire des estimations imputées sans biais.

Quand ils ont élaboré la théorie de la méthode AJ, Rao et Shao (1992) ont supposé que les facteurs *fcpf* étaient ignorables. Cependant, dans les simulations, ils ne le sont pas dans certaines strates, leur valeur variant d'environ 0,05 à 0,24. Shao et Steel (1999), ainsi que Lee, Rancourt et Särndal (1995) donnent des méthodes pour tenir compte des *fcpf* non négligeables. Dans la simulation, nous avons utilisé l'ajustement pour les *fcpf* proposé par Lee, Rancourt et Särndal (1995) parce qu'il est facile à appliquer. Sans l'ajustement pour les *fcpf*, l'estimateur de la variance AJ surestime considérablement les variances dans les simulations.

2.3 Imputation multiple

L'imputation multiple (MI) est décrite en détail dans Rubin (1987), ainsi que dans Little et Rubin (2002). Le résumé présenté ici a trait à son application en présence d'imputation hot deck. Comme pour l'approche assistée par modèle, nous supposons que, dans les cellules hot deck, les réponses manquent de façon aléatoire et que les y sont des variables aléatoires indépendantes de moyenne et variance communes. Pour chaque unité pour laquelle une valeur manque, M valeurs sont imputées, ce qui crée M ensembles de données complets.

Pour éviter de sous-estimer les variances par la méthode MI, il faut modifier la méthode hot deck. Rubin et Schenker (1986) ont proposé le bootstrap bayesien approximatif (ABB) pour l'échantillonnage aléatoire simple avec imputation hot deck en cas d'utilisation de la méthode MI. Pour les simulations, nous avons modifié l'ABB afin de tenir compte de l'échantillonnage pppsar des enregistrements donneurs. Nous avons créé un groupe d'enregistrements donneurs pour l'ABB dans chaque cellule en sélectionnant les répondants avec probabilité proportionnelle à w_i. (Aucun article publié ne discute de l'application des méthodes ABB avec des poids inégaux. A posteriori, nous pensons qu'une méthode ABB non pondérée aurait été préférable. L'utilisation d'un ABB non pondéré avec une imputation hot deck pppsar produit des estimations ponctuelles sans biais des totaux de population sous le modèle de probabilité de réponse).

3. Conception de l'étude par simulation

3.1 Description de la population étudiée et du plan d'échantillonnage

La base de sondage pour les simulations est un sous-ensemble du fichier des districts scolaires publics extraits du Common Core of Data (CCD) de 1999–2000 assemblé par le U.S. National Center for Education Statistics. La base de sondage finale comprend 11 941 districts.

Pour les simulations, nous avons sélectionné un échantillon de 1 020 districts scolaires conformément à un plan d'échantillonnage aléatoire simple stratifié. Nous avons créé 12 strates par croisement de 4 catégories de nombre d'élèves (taille du district) et 3 catégories de pourcentage d'élèves se trouvant au seuil de pauvreté ou sous celui-ci (situation de pauvreté). Les strates et les nombres de districts dans la base de sondage sont présentés au tableau 1. Celui-ci donne aussi les tailles d'échantillon de strate et les taux d'échantillonnage utilisés dans les simulations.

Le tableau contient également les moyennes et les écarts-types de strate pour les deux variables étudiées, c'est-à-dire le nombre d'élèves dans le district et le nombre de district dont le niveau d'enseignement le plus faible est la prématernelle. Nous avons choisi d'étudier ces variables, parce qu'elles sont typiques de nombreuses estimations calculées d'après le genre de plan de sondage susmentionné.

En plus des estimations pour la population dans son ensemble, nous avons estimé la valeur des deux variables étudiées pour deux domaines, définis comme étant les districts situés dans la région du Nord-Est et ceux situés dans des régions non métropolitaines. Les moyennes pour ces domaines sont très différentes des moyennes pour la population dans son ensemble pour les deux variables étudiées.

3.2 Mécanismes de génération des données manquantes et méthodes d'imputation

Par construction, l'information sur les deux variables étudiées est disponible pour tous les districts compris dans

la base de sondage. Pour générer des valeurs manquantes, nous avons attribué des indicateurs de réponse aux unités échantillonnées comprises dans les « cellules de réponse ». Dans certains cas, ces dernières sont les strates d'échantillonnage, et sont nommées cellules STR, tandis que dans d'autres, il s'agit de cellules nommées cellules HD. Ces dernières ont été définies par croisement de quatre régions géographiques et d'une catégorisation à quatre niveaux du nombre d'enseignants équivalent temps plein dans le district. Les cellules HD sont plus ou moins corrélées aux strates d'échantillonnage, mais chacune contient des unités provenant de plus d'une strate.

Dans une cellule de réponse donnée, nous avons réparti aléatoirement les unités échantillonnées entre les catégories manquante ou non manquante selon un taux précisé. Pour chaque type de cellule de réponse, nous avons choisi trois scénarios pour attribuer les taux de réponses manquantes. Dans deux scénarios, ce taux variait selon la cellule de réponse, tandis que dans le troisième, il était constant dans toutes les cellules.

Nous avons réalisé les simulations en tirant d'abord un échantillon aléatoire simple stratifié d'après les tailles d'échantillon de strate du tableau 1. Après avoir sélectionné l'échantillon, nous avons attribué aléatoirement la situation de réponse (répondant/non-répondant) à chaque unité échantillonnée conformément au scénario de réponse donné. Pour les méthodes MA et AJ, nous avons utilisé les procédures d'imputation hot deck pondérée décrites plus haut pour imputer les valeurs manquantes. Pour la méthode MI, nous avons d'abord créé un groupe de donneurs en utilisant l'ABB pondéré, puis nous avons utilisé la méthode hot deck pondérée pour chacun des $M = 5$ ensembles de données imputés. Nous avons calculé les nombres totaux estimés d'élèves et de districts avec prématernelle pour l'échantillon simulé avec valeur imputée, ainsi que les estimations de la variance de ces estimations selon les trois méthodes d'estimation de la variance. (Pour les simulations exécutées pour lesquelles il n'a pas été possible de calculer la variance estimée ou pour lesquelles la taille d'échantillon dans une cellule était inférieure à 2, l'échantillon a été supprimé. Le nombre maximal d'échantillons supprimés sur l'ensemble des 10 000 exécutions de chacune des simulations était de 2 pour la méthode MA et de 28 pour la méthode AJ le nombre d'échantillons AJ supprimés n'a été de 28 que pour une seule exécution; le nombre le plus grand suivant était de 3). La méthode AJ portait sur 3 strates de variance combinées et 40 groupes d'unités par strate pour un total de 120 répliques. Les trois strates combinées, formées d'après des strates ayant environ les mêmes *fcpf*, étaient constituées des strates 1 à 6, 7 à 9 et 10 à 12. Pour vérifier le groupement, nous nous sommes assurés que la procédure d'estimation de la variance par le jackknife groupé donnait essentiellement les mêmes estimations moyennes de la variance et taux de couverture des intervalles de confiance que le jackknife non groupé dans le cas de la réponse complète. Le processus complet a été répété 10 000 fois pour chaque scénario de réponse.

Une caractéristique de la conception des simulations est que les moyennes des deux domaines considérés diffèrent souvent beaucoup des moyennes de population complète selon la strate et la cellule HD. Une remarque importante en ce qui concerne les estimations par domaine est que les imputations ont été faites en sélectionnant des donneurs à partir de l'ensemble des répondants dans une cellule hot deck, sans reconnaître précisément le domaine comme cela pourrait être fait en pratique dans certains cas. Après avoir fait les imputations pour l'échantillon complet, nous avons estimé le total pour un domaine par $\hat{\theta}_I = \sum_{i\in A_R} \delta_i w_i y_i + \sum_{j\in A_M} \delta_j w_j y_j^*$, où $\delta_i = 1$ si l'unité i est dans le domaine et 0 autrement.

Tableau 1
Définitions des strates, chiffres de population, tailles d'échantillon, taux d'échantillonnage, moyennes et écarts-types du nombre d'élèves et proportions de districts offrant une prématernelle

Strate	Taille du district	Situation de pauvreté	N_h	n_h	Taux d'échant.	Nombre d'élèves		Proportion avec prématernelle
						Moyenne	é.-t.	
1	1	1	615	32	0,0520	270,0	155,0	0,44
2	1	2	1 147	59	0,0514	263,3	175,0	0,49
3	1	3	1 292	66	0,0511	243,5	142,5	0,49
4	2	1	1 720	111	0,0645	1 607,2	837,0	0,44
5	2	2	2 305	149	0,0646	1 429,7	784,1	0,52
6	2	3	1 893	122	0,0644	1 427,8	788,8	0,63
7	3	1	692	75	0,1084	4 695,3	1 360,6	0,35
8	3	2	579	63	0,1088	4 728,5	1 365,0	0,51
9	3	3	527	57	0,1082	4591,8	1 380,3	0,63
10	4	1	342	83	0,2427	16 003,4	12 670,2	0,51
11	4	2	449	110	0,2450	17 577,3	14 246,7	0,58
12	4	3	380	93	0,2447	19 331,8	16 142,7	0,68
Total			11 941	1 020		3 237,9	6 770,5	0,52

Trois des quatre combinaisons possibles de mécanisme de réponse (cellules STR ou HD) et de formation de cellules hot deck (cellules STR ou HD) ont été étudiées dans les simulations. Nous nommons ces combinaisons STR/STR, HD/HD et STR/HD, où le premier ensemble de lettres désigne le mécanisme de réponse et le deuxième, le type de cellule hot deck. Les trois ensembles de valeurs du taux de réponse étaient de 0,2 à 0,6 espacées uniformément entre les cellules de réponse, une valeur constante de 0,7 dans toutes les cellules, et de 0,6 à 0,9 répartie uniformément entre les cellules. Les trois combinaisons cellule de réponse/cellule hot deck et les trois ensembles de taux de réponse ont généré neuf scénarios de simulation distincts pour chaque estimation.

3.3 Hypothèses concernant les modèles de réponse et la structure de la population

Deux modèles interviennent dans les simulations. Le modèle de population repose sur l'hypothèse que les valeurs de y dans chaque cellule hot deck sont indépendantes et ont la même espérance. Le modèle de réponse repose sur l'hypothèse qu'il existe une probabilité de réponse uniforme dans chaque cellule hot deck. Si les deux modèles tiennent, alors l'utilisation d'une imputation hot deck non pondérée ou pondérée produira une estimation sans biais du total global de population. Par contre, si l'on suppose que seul le modèle de réponse est vérifié, alors l'utilisation d'une imputation hot deck pondérée est nécessaire pour produire une estimation sans biais de ce total. Puisque l'imputation hot deck pondérée est utilisée dans les simulations, il suffit que le modèle de probabilité de réponse soit satisfait pour obtenir une estimation ponctuelle sans biais du total global de population. Le modèle de probabilité de réponse est vérifié pour toutes les combinaisons STR/STR et HD/HD, ainsi que pour la combinaison STR/HD avec taux de réponse constant; cependant, il ne tient pas pour les deux autres combinaisons STR/HD. La théorie de la méthode AJ d'estimation de la variance des totaux de population a été élaborée en supposant seulement que le modèle de probabilité de réponse est vérifié. Les théories MA et MI reposent sur l'hypothèse que les deux modèles sont vérifiés.

La possibilité d'utiliser uniquement le modèle de probabilité de réponse et l'imputation hot deck pondérée pour produire des estimations sans biais des totaux de population ne s'étend généralement pas à l'estimation des totaux de domaine. Si le domaine recoupe les cellules hot deck, il est nécessaire d'utiliser un modèle de population qui suppose que l'espérance des valeurs de domaine est la même que celle des valeurs hors domaine dans chaque cellule hot deck. Cependant, si les cellules hot deck sont définies de façon que chaque domaine comprenne la population complète dans un sous-ensemble des cellules hot deck, alors la situation pour l'estimation ponctuelle et l'estimation de la variance est la même que celle mentionnée plus haut pour les totaux globaux de population.

De façon générale, nous avons construit les scénarios de simulation de sorte que les cellules hot deck n'englobent pas les domaines, afin de refléter le fait qu'en pratique, il est essentiellement impossible d'intégrer tous les domaines dans un schéma d'imputation. Plus précisément, dans les simulations, les districts de la région du Nord-Est (NE) et ceux des régions statistiques non métropolitaines (NMSA) ne sont pas reliés aux définitions de strate du tableau 1 (qui sont utilisées comme cellules hot deck dans certains cas). En outre, les districts compris dans le domaine NMSA peuvent se retrouver dans toutes les cellules HD. Cependant, le domaine NE est un sous-ensemble de quatre cellules HD. Donc, la définition des cellules HD concorde plus avec l'estimation des totaux de domaine NE que celles des totaux de domaine NMSA.

3.4 Statistiques sommaires

Le biais relatif d'une estimation ponctuelle est estimé par $\text{relbiais}(\hat{\theta}_I) = \text{biais}(\hat{\theta}_I)/\theta_N$, où $\text{biais}(\hat{\theta}_I) = \sum_s (\hat{\theta}_{Is} - \theta_N)/10\,000$, $\hat{\theta}_{Is}$ est l'estimation d'après l'échantillon s, et θ_N est le paramètre de population finie. La variance empirique de $\hat{\theta}_I$ est $\text{Var}(\hat{\theta}_I) = \sum_s (\hat{\theta}_{Is} - \overline{\theta}_I)^2/10\,000$, où $\overline{\theta}_I = \sum_s \hat{\theta}_{Is}/10\,000$. L'estimation moyenne de la variance pour une méthode particulière est $v = \sum_s v_s/10\,000$, où v_s est la variance estimée pour l'exécution de simulation s.

Les pourcentages d'intervalles qui comprennent θ_N sont fondés sur les intervalles de confiance à 95 % nominaux $(\hat{\theta}_I \pm t\hat{V}^{1/2})$ calculés pour chacune des 10 000 simulations pour chaque scénario de simulation. Un élément dont il faut tenir compte ici est la précision des estimations de la variance d'après un plan d'échantillonnage stratifié non proportionnel et son incidence sur la décision d'utiliser une approximation normale ou bien des intervalles t pour calculer les intervalles de confiance. Nous avons constaté que, dans la plupart des cas, l'utilisation de la loi t n'avait aucun effet important pour les méthodes MA et AJ, et nous avons par conséquent utilisé un multiplicateur de 1,96 pour les intervalles de confiance basés sur ces méthodes. Rubin et Schenker (1986) proposent d'utiliser une loi t avec λ degrés de liberté pour les intervalles de confiance pour la méthode MI où

$$\lambda = (M-1)\left(1 + \frac{M}{M+1}\frac{U}{B}\right)^2.$$

Puisque l'utilisation de 1,96 avec la méthode MI a donné des intervalles dont la couverture est nettement trop faible, nous utilisons la loi t avec λ degrés de liberté pour les intervalles de confiance MI.

4. Résultats des simulations

À la présente section, nous présentons les principaux résultats des simulations, en commençant par la performance des trois méthodes d'estimation de la variance des estimations pour l'ensemble de la population, puis nous donnons les résultats pour les estimations par domaine. Les résultats essentiels sont résumés graphiquement ici, mais les tableaux contenant les données détaillées peuvent être consultés dans Brick, Jones, Kalton et Valliant (2004).

4.1 Estimations pour l'ensemble de la population

La figure 1 illustre les résultats des simulations pour l'estimation du nombre total d'élèves et du nombre de districts offrant la prématernelle d'après les 10 000 échantillons pour chacun des 9 scénarios de simulation. Y sont présentés le biais relatif de l'estimateur imputé, l'estimation moyenne de la variance en pourcentage de la variance empirique et le taux de couverture de l'intervalle de confiance.

Théoriquement, les estimations ponctuelles sont sans biais sous imputation hot deck pondérée si toutes les unités dans une cellule hot deck ont la même probabilité de réponse. Comme nous l'avons mentionné plus haut, cette condition est vérifiée pour les combinaisons STR/STR et HD/HD, ainsi que pour la combinaison STR/HD avec une probabilité de réponse globale uniforme. Le graphique des biais relatifs de la figure 1 concorde avec ce résultat théorique dans les limites de l'erreur de simulation. Bien que le biais relatif des estimations ponctuelles pour les deux autres scénarios STR/HD soit faible (systématiquement inférieur à 3 %), il peut néanmoins être important si les erreurs-types des estimations sont également faibles. Cochran (1977, page 12) montre que, si le ratio du biais à l'erreur-type est relativement grand, alors le taux de couverture peut être nettement plus faible que le taux nominal. Pour les estimations pour l'ensemble de la population avec cette taille d'échantillon, les ratios ne sont jamais supérieurs à 0,4, mais prennent une valeur beaucoup plus grande pour les estimations par domaine, comme nous l'exposons plus loin.

Le graphique des ratios de l'estimation moyenne de la variance à la variance empirique (*v*/Var dans les figures) pour les trois méthodes montre que ces estimations présentent un biais assez faible dans la plupart des cas, compris dans la fourchette de plus ou moins 8 % par rapport à la variance réelle simulée. Bien que les ratios pour toutes les méthodes varient entre les neuf scénarios, les ratios MI sont un peu plus variables que ceux calculés pour les deux autres méthodes.

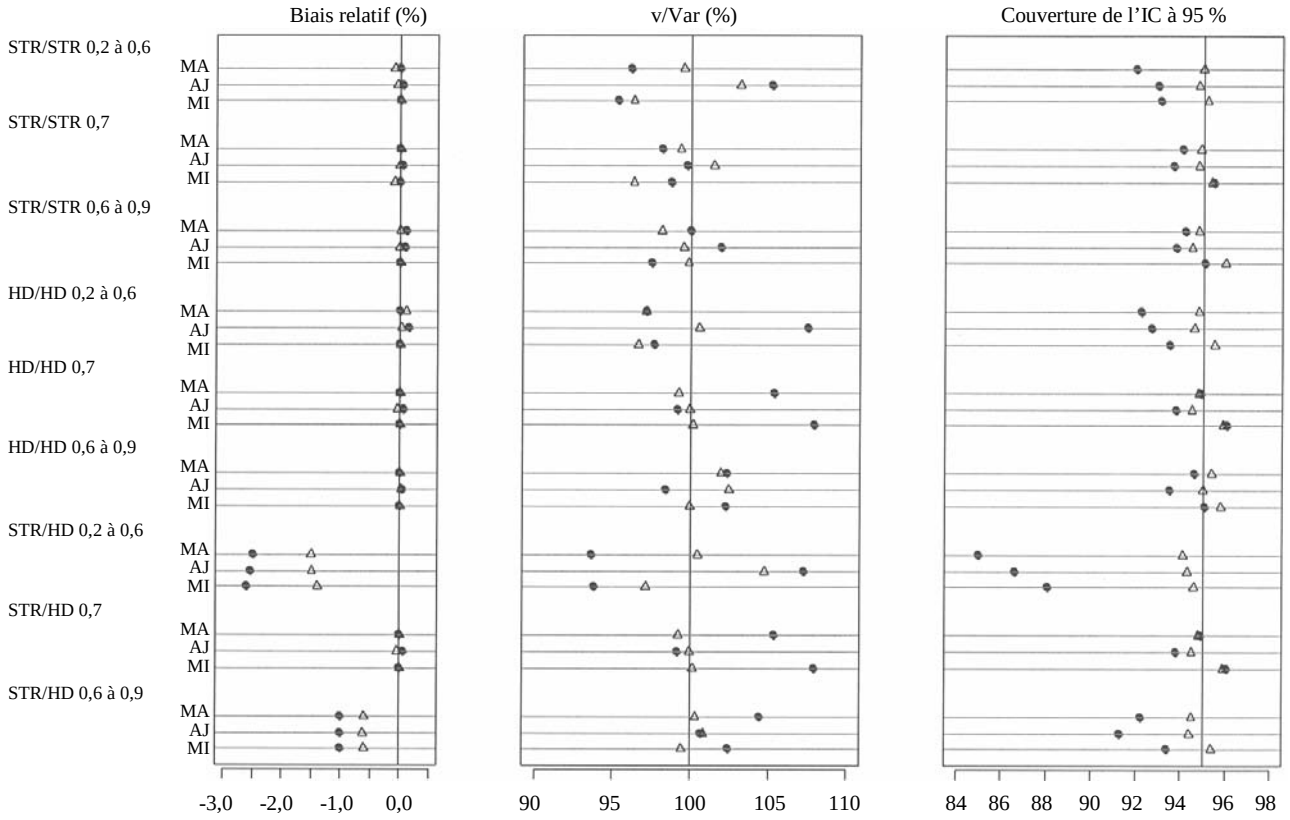

Figure 1. Biais relatifs, rapports des variances et couverture de l'intervalle de confiance à 95 % pour le nombre d'élèves (•) et le nombre de districts offrant la prématernelle (Δ).

L'une des principales raisons du calcul des variances est la production d'intervalles de confiance. Le panneau de droite de la figure 1 montre que les taux de couverture des intervalles de confiance des estimations s'approchent généralement du taux nominal de 95 %, particulièrement pour la statistique de la prématernelle. Pour les deux statistiques et pour toutes les méthodes et tous les scénarios, les taux de couverture sont compris entre 91 % et 96 %, à l'exception du nombre d'élèves pour le scénario STR/HD 0,2 à 0,6. Dans ce cas, pour lequel le taux de non-réponse est extrêmement élevé, les taux de couverture de 88 % ou moins pour les trois méthodes sont dus au biais relativement grand dans l'estimation ponctuelle. Dans l'ensemble, les trois méthodes d'estimation de la variance produisent des intervalles de confiance dont la couverture est nettement meilleure que celle des intervalles basés sur les estimations naïves de la variance (Brick et coll. 2004).

Pour les méthodes MA et AJ, les taux de couverture des intervalles de confiance sont essentiellement équivalents. Pour la méthode MI, ils sont généralement un peu plus élevés que pour les deux méthodes susmentionnées. Les taux de couverture MI sont aussi légèrement plus proche du taux nominal pour la variable de nombre d'élèves. La plupart des écarts sont faibles.

Les taux de couverture supérieur et inférieur de l'intervalle de confiance sont les mêmes pour les trois méthodes d'estimation de la variance. Dans le cas du nombre d'élèves, qui est une variable fortement asymétrique, les taux de couverture dans les queues de la distribution sont inégaux à cause de la corrélation entre le total estimé et les estimations de l'erreur-type. La couverture asymétrique dans les queues est également associée à des taux de couverture globaux plus faibles.

Les méthodes MA et AJ produisent des intervalles de confiance dont la longueur moyenne est presque la même pour les divers scénarios et variables. Comme la méthode MI utilise des valeurs provenant de la loi *t*, ses intervalles sont de 10 % à 20 % plus longs que les intervalles MA et AJ quand les taux de réponse sont faibles. Pour les taux de réponse plus élevés, la longueur des intervalles MI varie d'à peu près la même valeur à 5 % de plus que celle des intervalles produits par les deux autres méthodes. On pourrait, évidemment, raccourcir les intervalles de confiance MI en augmentant la valeur de M (Rubin 1987, chapitre 4), même si $M = 5$ est la valeur typique pour les applications.

4.2 Estimations par domaine

L'estimation des caractéristiques de domaine qui ne sont pas explicitement intégrées dans le scénario d'imputation peut être problématique lorsque le taux de données manquantes n'est pas négligeable. Kalton et Kaspryzk (1986), ainsi que Rubin (1996) et de nombreux autres auteurs ont discuté ce point et ont recommandé vivement d'inclure autant de variables que possible que dans le processus d'imputation. Cependant, étant donné le grand nombre d'analyses par domaine planifiées et ponctuelles réalisées d'après des données d'enquête, il n'est pas raisonnable de supposer que l'on peut tenir compte de tous les domaines dans un scénario d'imputation. Par conséquent, lors de la conception des simulations, nous avons omis sciemment d'inclure explicitement les domaines dans la définition des cellules hot deck. Dans le cas de l'imputation multiple, beaucoup d'attention a été accordée au problème d'estimation de la variance des estimations par domaine (par exemple, Fay 1992; Meng 1994; Rubin 1996).

Dans les simulations, nous estimons les totaux pour deux domaines, à savoir, les nombres de districts scolaires dans le domaine NE et dans le domaine NMSA. Les figures 2 et 3 présentent les résultats des simulations pour le domaine NE et pour le domaine NMSA, respectivement, dans le même format que celui utilisé auparavant. Notons que les échelles des figures 2 et 3 diffèrent l'une de l'autre et sont fort différentes de celles utilisées pour les estimations pour l'ensemble de la population.

Pour le domaine NE, les estimations ponctuelles présentent un biais positif important pour les combinaisons STR/STR. Les cellules hot deck basées sur STR ne sont pas reliées à la région et, par conséquent, les districts NE pour lesquels des données manquent ont des donneurs provenant d'autres régions, dont les caractéristiques sont différentes. En revanche, l'inclusion de la région dans la construction des cellules d'imputation HD élimine le biais des estimations ponctuelles pour les combinaisons HD/HD et pour la combinaison STR/HD avec probabilité de réponse globale uniforme, et réduit le biais pour les autres combinaisons STR/HD.

Les trois méthodes d'estimation de la variance requièrent des estimations ponctuelles sans biais et la théorie qui les sous-tend ne fournit aucune orientation quant à la façon dont elles fonctionneront sous les conditions que nous étudions. Les estimations de la variance sont approximativement sans biais pour les trois méthodes quand les estimations ponctuelles par domaine sont sans biais ou ne présentent qu'un biais faible. Cependant, la figure 2 montre que, pour la combinaison STR/STR, où les estimations ponctuelles sont sérieusement biaisées, les estimations de la variance surestiment habituellement les variances empiriques.

La figure 2 montre que les taux de couverture pour les scénarios HD/HD et STR/HD–pour lesquels le biais relatif des estimations ponctuelles est nul ou faible–sont compris entre 92 % et 96 % pour tous ces scénarios et méthodes d'estimation de la variance, sauf un. Fait exception la combinaison STR/HD avec taux de réponse compris entre

0,2 et 0,6, dont le taux de couverture pour le nombre d'élèves peut être aussi faible que 86 %.

Pour les scénarios STR/STR, la figure 2 indique que toutes les méthodes ont tendance à produire une couverture supérieure au taux nominal pour le nombre d'élèves et inférieure au taux nominal pour le nombre de districts offrant la prématernelle. L'écart entre les taux de couverture pour les deux variables est dû à la taille du biais relatif des estimations ponctuelles et des estimations de la variance.

Si nous passons aux estimations pour le domaine NMSA à la figure 3, il convient de souligner que la situation de région métropolitaine n'est explicitement incluse ni dans la définition de STR ni dans celle de HD, bien qu'elle soit clairement corrélée à la taille et, donc, à STR. Pour tous les scénarios, les estimations ponctuelles du nombre d'élèves dans le domaine NMSA présentent un biais positif important. La couverture des intervalles de confiance MA est uniformément égale ou supérieure au taux nominal, principalement à cause du biais positif extrême dans les estimations de la variance. La couverture des intervalles AJ s'approche du taux nominal pour les scénarios HD/HD et STR/HD, mais est inférieure à ce taux pour les trois scénarios STR/STR. Les profils de couverture pour les intervalles de confiance MI sont semblables à ceux observés pour la méthode AJ, sauf que la couverture des intervalles MI est appréciablement inférieure au taux nominal pour le scénario HD/HD avec taux de réponse de 0,2 à 0,6.

Les estimations ponctuelles du nombre de districts avec prématernelle dans le domaine NMSA ont un biais relatif négatif modéré pour chacun des neuf scénarios. Pour les trois méthodes d'estimation de la variance, la couverture des intervalles de confiance s'approche du taux nominal, sans la surcouverture observée pour les estimations pour le domaine NE.

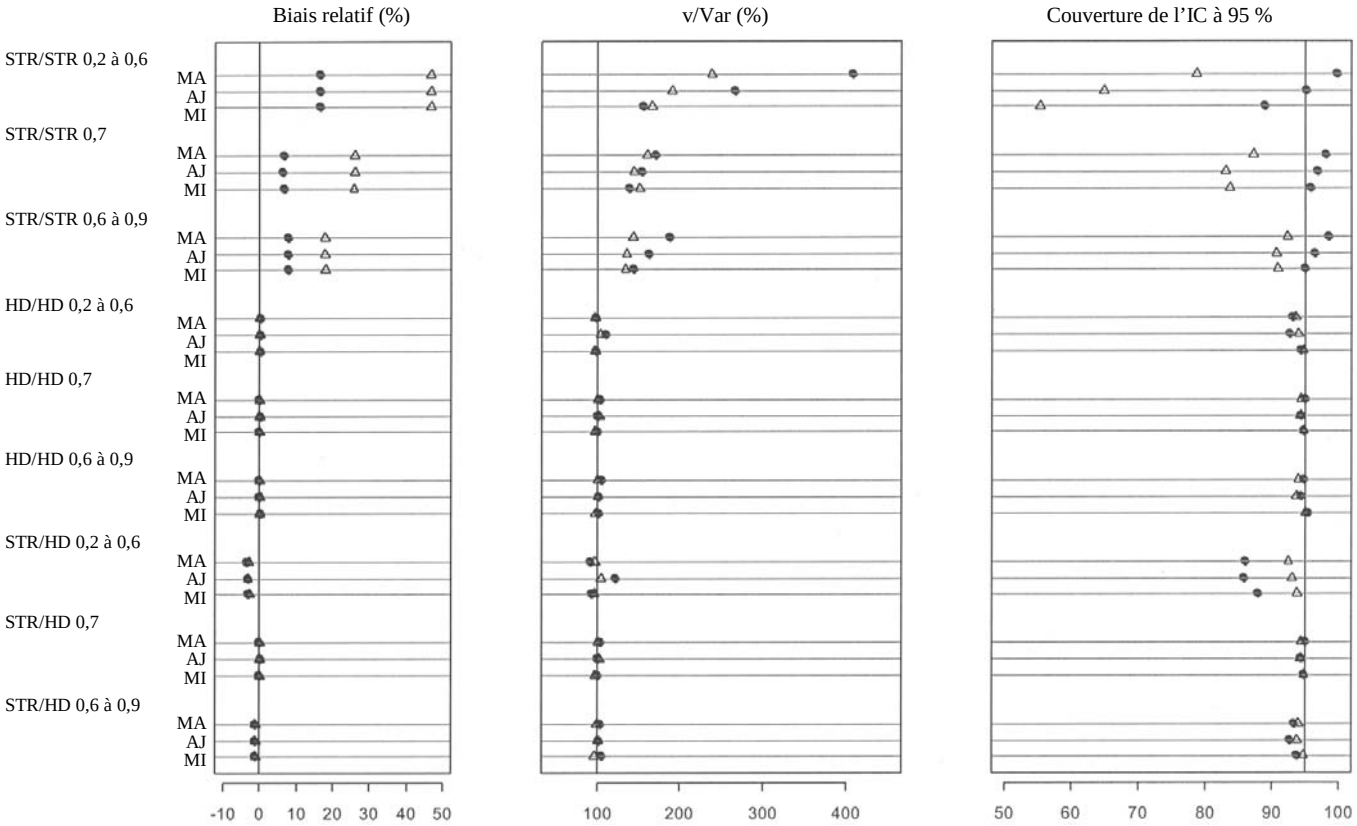

Figure 2. Biais relatifs, rapports des variances et couverture des intervalle de confiance à 95 % pour le nombre d'élèves (•) et le nombre de districts offrant la prématernelle (Δ) dans le Nord-Est.

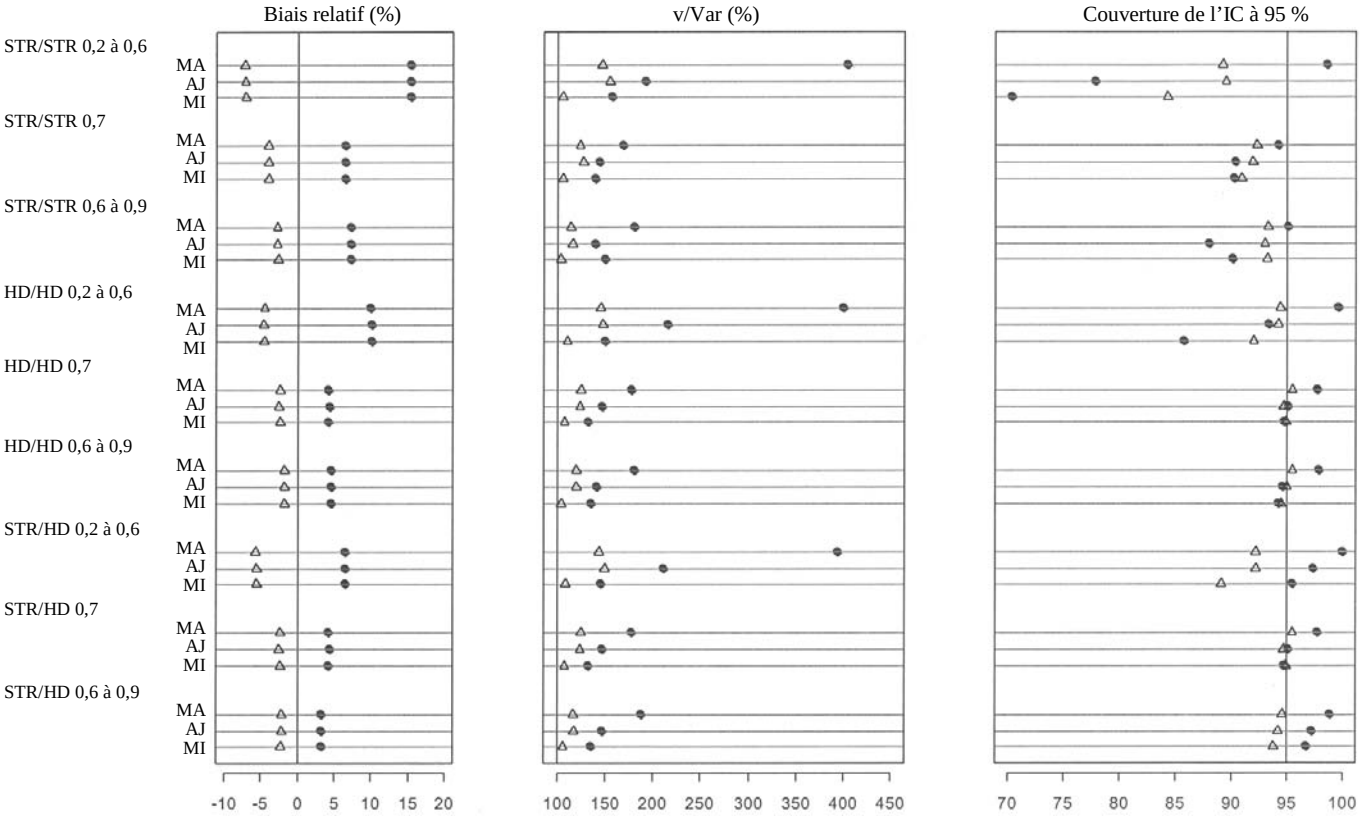

Figure 3. Biais relatifs, rapports des variances et couverture des intervalles de confiance à 95 % pour le nombre d'élèves (•) et le nombre de districts offrant la prématernelle (Δ) dans les régions non métropolitaines.

5. Conclusion

Nos simulations avaient pour but d'examiner les propriétés de trois estimateurs de la variance des estimations de totaux imputés d'après un plan d'échantillonnage stratifié sous divers mécanismes de réponse avec imputation hot deck pondérée. Les conditions des simulations reflètent celles auxquelles ont peut s'attendre en pratique en ce sens que les hypothèses qui sous-tendent les méthodes sont violées de diverses façons. Les trois méthodes produisent des estimations nettement meilleures que l'estimateur naïf de la variance. Elles donnent toutes trois de très bons résultats quand les estimations ponctuelles sont sans biais. Si le biais dans les estimations ponctuelles est important, aucune des méthodes ne produit des intervalles de confiance dont la couverture correspond au taux nominal. L'obtention de taux de couverture médiocres pour des estimations ponctuelles biaisées n'est pas inattendue, puisqu'il en est également ainsi quand il n'y a pas de données manquantes. Lorsque le biais des estimations ponctuelles est relativement faible, les taux de couverture réels obtenus pour les trois méthodes d'estimation de la variance sont parfois supérieurs et parfois inférieurs au taux nominal. Dans ce cas, la tendance des trois méthodes à surestimer la variance produit souvent des taux de couverture proches du taux nominal. Les faibles taux de réponse sont associés à un taux de couverture trop faible, dû en grande en partie aux biais plus importants dans les estimations ponctuelles.

En général, les écarts entre les taux de couverture obtenus pour les trois méthodes sont faibles et ne permettent pas d'affirmer que l'une des méthodes est supérieure aux autres en général. Pour des taux de réponse très faibles, la longueur moyenne des intervalles de confiance pour la méthode MI est appréciablement plus grande que celle observée pour les méthodes MA et AJ, mais l'utilisation d'un plus grand nombre d'ensembles d'imputations avec la méthode MI pourrait corriger ce problème. Il convient toutefois de souligner que ces simulations ne portent que sur le cas de l'échantillonnage à un degré. Il pourrait exister des écarts entre les longueurs des intervalles de confiance produits par les diverses méthodes en cas d'échantillons en grappes. Cette possibilité devrait faire l'objet d'études futures.

Les résultats de la présente étude donnent aux praticiens de l'imputation hot deck des preuves empiriques que toutes les méthodes d'estimation de la variance donnent de bons résultats en cas d'échantillonnage à un seul degré à condition que l'estimation ponctuelle soit sans biais, même si d'autres hypothèses sont violées. Les estimations pour des domaines qui ne sont pas pris en compte dans le scénario

d'imputation sont susceptibles de présenter un biais important. Si les estimations ponctuelles sont fortement biaisées, les méthodes peuvent produire des intervalles de confiance dont la couverture est de loin inférieure au taux nominal. Les analystes des ensembles de données imputées devraient déterminer si la méthode d'imputation qui a été utilisée produira vraisemblablement des estimations approximativement sans biais, particulièrement pour les estimations par domaine. Sinon, il pourrait être nécessaire qu'ils réimputent les réponses manquantes pour obtenir des estimations ponctuelles moins biaisées. Recommander aux imputeurs de tirer parti d'autant de variables explicatives que possible dans le processus d'imputation n'est pas un conseil nouveau, mais les conclusions qui se dégagent de nos simulations prouvent son importance.

Remerciements

Les auteurs remercient l'Institute for Education Sciences du National Center for Education Statistics d'avoir appuyé cette étude, en particulier Marilyn Seastrom. Nous tenons aussi à remercier les examinateurs de leurs commentaires constructifs.

Bibliographie

Brick, J.M., Kalton, G. et Kim, J.K. (2004). Estimation de variance pour l'imputation hot deck à l'aide d'un modèle. *Techniques d'enquête*, 30, 63-72.

Brick, J.M., Jones, M., Kalton, G. et Valliant, R. (2004). A simulation study of three methods of variance estimation with hot deck imputation for stratified samples. Préeparé sous contract No. RN95127001 to the National Center for Education Statistics. Rockville, MD: Westat, Inc.

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

Fay, R.E. (1992). When are imputations from multiple imputation valid. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 227-232.

Kalton, G., et Kasprzyk, D. (1986). Le traitement des données d'enquêtes manquantes. *Techniques d'enquête*, 12, 1-17.

Lee, H., Rancourt, E. et Särndal, C.-E. (1995). Jackknife variance estimation for data with imputed values. *Proceedings of the Statistical Society of Canada Survey Methods Section*, 111-115.

Lee, H., Rancourt, E. et Särndal, C.-E. (2001). Variance estimation from survey data under single imputation. Dans *Survey Nonresponse* (Éds. R.M. Groves, D.A. Dillman, J.L. Eltinge et R.J.A Little), Chapître 21, New York: John Wiley & Sons Inc.

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

Meng, X.-L. (1994). Multiple imputation inferences with uncongenial sources of input. (avec discussion). *Statistical Science*, 9, 538-573.

Rao, J.N.K., et Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

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

Rubin, D.B. (1996). Multiple imputation after 18+ years (avec discussion). *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 nonignorable nonresponse. *Journal of the American Statistical Association*, 81, 361-374.

Rust, K., et Rao, J.N.K. (1996). Variance estimation for complex estimators in sample surveys. *Statistics in Medicine*, 5, 381-397.

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 estimation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Techniques d'enquête, décembre 2005
Vol. 31, N° 2, pp. 175-183
Statistique Canada, N° 12-001-XPB au catalogue

La pondération pour la non-réponse augmente-t-elle la variance des moyennes de sondage?

Roderick J. Little et Sonya Vartivarian [1]

Résumé

La pondération pour la non-réponse est une méthode courante de traitement de la non-réponse totale dans les sondages. Elle vise à réduire le biais dû à la non-réponse, mais produit souvent un accroissement de la variance. Par conséquent, son efficacité est souvent considérée comme un compromis entre le biais et la variance. Cette vision est cependant simpliste, car la pondération pour la non-réponse peut, en fait, réduire le biais ainsi que la variance. Pour réduire le biais de non-réponse, une covariable de repondération doit avoir deux caractéristiques : elle doit être corrélée à la probabilité de réponse, d'une part, et à la variable d'intérêt, d'autre part. Si cette deuxième caractéristique existe, la repondération peut réduire plutôt qu'augmenter la variance d'échantillonnage. Nous présentons une analyse détaillée du biais et de la variance dans le cas d'une pondération pour l'estimation d'une moyenne de sondage au moyen de cellules d'ajustement. L'analyse donne à penser que la caractéristique la plus importante des variables à inclure dans la repondération est qu'elles soient prédictives des variables d'intérêt; la prédiction de la propension à répondre est un objectif secondaire, quoiqu'utile. Nous proposons des estimations empiriques de la racine carrée de l'erreur quadratique moyenne pour déterminer dans quelles circonstances la repondération est efficace et nous les évaluons au moyen d'une étude en simulation. Un estimateur composite simple fondé sur la racine de l'erreur quadratique moyenne empirique donne de meilleurs résultats que l'estimateur pondéré dans les simulations.

Mots clés : Données manquantes; correction pour la non-réponse; poids de sondage; non-réponse à une enquête.

1. Introduction

Dans la plupart des enquêtes, certaines personnes échantillonnées ne fournissent pas d'information parce qu'on n'a pas pu prendre contact avec elles ou qu'elles ont refusé de répondre (non-réponse *totale*). La méthode la plus courante de correction de la non-réponse totale est la pondération, où les répondants et les non-répondants sont classés dans des cellules d'ajustement d'après des données sur des covariables dont les valeurs sont connues pour toutes les unités échantillonnées, et un poids de non-réponse est calculé pour les cas compris dans une cellule proportionnellement à l'inverse du taux de réponse dans la cellule. Souvent, on multiplie le poids de sondage par ces poids de non-réponse et on normalise le poids global de sorte que la somme des poids des cellules soit égale au nombre de répondants dans l'échantillon. Oh et Scheuren (1983) donnent une bonne vue d'ensemble de la pondération pour la non-réponse. Une approche apparentée est la post-stratification (Holt et Smith 1979), qui s'applique lorsque la distribution de la population entre les cellules d'ajustement peut être déterminée d'après des sources externes, comme le recensement. Le poids est alors proportionnel au ratio du chiffre de population au nombre de répondants dans une cellule.

La pondération pour la non-réponse, ou repondération, est considérée principalement comme un moyen de réduire le biais dû à la non-réponse totale. Ce rôle de la repondération est analogue à celui des poids de sondage et est relié à la propriété d'absence de biais par rapport au plan de sondage de l'estimateur du total d'Horvitz-Thompson (Horvitz et Thompson 1952), où les unités sont pondérées par l'inverse de leur probabilité de sélection. La pondération pour la non-réponse peut être considérée comme une extension naturelle de cette idée, où les unités comprises dans l'échantillon sont pondérées par l'inverse de leur probabilité d'inclusion, estimée comme étant le produit de la probabilité de sélection et de la probabilité de réponse, sachant que l'unité a été sélectionnée; l'inverse de la seconde probabilité est le poids de non-réponse. Bien que certains modéliseurs soutiennent que la repondération en vue de corriger le biais n'est pas nécessaire dans les modèles où la pondération n'est pas associée aux variables d'intérêt, en pratique, peu d'entre eux sont prêts à émettre une hypothèse aussi forte.

Les poids de sondage réduisent le biais au prix d'un accroissement de la variance, si la variance du résultat est constante. Étant donné l'analogie entre les poids de non-réponse et les poids de sondage, il paraît plausible que la pondération pour la non-réponse réduise aussi le biais au prix d'un accroissement de la variance des estimations par sondage. La notion de compris entre le biais et la variance est abordée dans certaines discussions de la repondération pour la non-réponse (Kalton et Kasprzyk 1986; Kish 1992; Little, Lewitzky, Heeringa, Lepkowski et Kessler 1997).

1. Roderick J. Little, University of Michigan, États-Unis. Courriel : rlittle@umich.edu; Sonya Vartivarian, Mathematica Policy Research, Inc. 600 Maryland Ave SW, Suite 550, Washington, D.C. 20024-2512. Courriel : SVartivarian@Mathematica-MPR.com.

Kish (1992) présente une formule simple pour l'accroissement proportionnel de la variance dû à la pondération, disons L, sous l'hypothèse que la variance des observations est approximativement constante :

$$L = cv^2, \qquad (1)$$

où cv est le coefficient de variation des poids des répondants.

L'équation (1) est une bonne approximation si la variable de cellule d'ajustement est faiblement associée aux variables d'intérêt. Cependant, puisqu'elle donne une approximation de la variance plutôt que de l'erreur quadratique moyenne, elle ne mesure pas la réduction possible du biais dû à la non-réponse qui est l'objectif principal de la repondération et elle ne s'applique pas aux résultats qui sont associés à la variable de cellule d'ajustement, pour lesquels la pondération pour la non-réponse peut en fait réduire la variance. Le fait que la pondération pour la non-réponse puisse réduire la variance est implicite dans la formule de Oh et Scheuren (1983) et est mentionné dans Little (1986) lorsque des cellules d'ajustement sont créées par stratification visant à prédire la moyenne. Ce fait se dégage aussi de la méthode connexe de post-stratification pour correction de la non-réponse (Holt et Smith 1979).

La variabilité des poids proprement dite ne se traduit pas nécessairement par des estimations ayant une plus forte variance : une estimation pour laquelle la valeur de L est élevée peut avoir une plus petite variance qu'une estimation dont la valeur de L est faible, comme l'illustrent les simulations présentées à la section 3. En outre, les situations où la repondération réduit le plus efficacement le biais dû à la non-réponse sont précisément celles où elle a tendance à réduire, et non à accroître, la variance et où l'équation (1) ne s'applique pas. Ces situations diffèrent du cas des poids de sondage et sont reliées à la « superefficacité » que l'on peut obtenir lorsque les poids sont estimés à partir de l'échantillon plutôt que d'être des constantes fixées; voir, par exemple, Robins, Rotnitsky et Zhao (1994).

Nous proposons un perfectionnement simple de l'équation (1), à savoir l'équation (14) données plus loin, qui reflète à la fois les composantes de biais et de variance, que la variable de cellule d'ajustement soit associée ou non aux résultats, et est, par conséquent, un indicateur plus précis de la valeur de la pondération des estimations et des variables de cellule d'ajustement. Dans les enquêtes polyvalentes comportant de nombreux résultats, l'approche type consiste à appliquer la même pondération pour la non-réponse à toutes les variables, en supposant implicitement que la valeur de la réduction du biais dû à la non-réponse pour certaines variables fait plus que compenser l'accroissement éventuel de la variance pour d'autres. Notre estimation empirique de l'erreur quadratique moyenne permet un simple perfectionnement de cette stratégie, à savoir la restriction de la repondération au sous-ensemble de variables pour lesquelles elle réduit l'erreur quadratique moyenne estimée. Nous évaluons cette stratégie composite dans l'étude en simulation présentée à la section 3 et montrons qu'elle présente certains avantages par rapport à la repondération de toutes les variables de résultat. Comme nous le mentions à la section 4, d'autres approches présentent d'encore meilleures propriétés statistiques, mais elles produisent des poids différents pour chaque variable, ce qui rend leur mise en œuvre et leur explication aux utilisateurs des données d'enquête plus fastidieuses.

2. Repondération pour la non-réponse pour une moyenne

Supposons qu'on sélectionne un échantillon de n unités. Nous envisageons l'inférence pour la moyenne de population d'une variable étudiée Y sujette à la non-réponse. Par souci de simplicité et pour nous concentrer sur la question de la correction pour la non-réponse, nous supposons que les unités sont sélectionnées par échantillonnage aléatoire simple. En général, les remarques faites ici au sujet de la repondération pour la non-réponse s'appliquent aussi à des plans de sondage complexes, quoique les détails techniques deviennent plus compliqués.

Nous supposons que les répondants et les non-répondants peuvent être classés dans C cellules d'ajustement d'après une covariable X. Soit M un indicateur de données manquantes dont la valeur est 0 pour les répondants et 1 pour les non-répondants. Soit n_{mc} le nombre d'individus échantillonnés avec $M = m, X = c; m = 0, 1; c = 1, \ldots, C$, $n_{+c} = n_{0c} + n_{1c}$ représente le nombre d'individus échantillonnés dans la cellule c, $n_0 = \sum_{c=1}^{C} n_{0c}$ et $n_1 = \sum_{c=1}^{C} n_{1c}$, les nombres totaux de répondants et de non-répondants, et $p_c = n_{+c} / n$, $p_{0c} = n_{0c} / n_0$, les proportions de cas échantillonnés et répondants dans la cellule c. Nous comparons deux estimations de la moyenne de population μ de Y, à savoir la moyenne non pondérée

$$\bar{y}_0 = \sum_{c=1}^{C} p_{0c} \bar{y}_{0c}, \qquad (2)$$

où $\bar{y}_{0c}$ est la moyenne pour les répondants dans la cellule c, et la moyenne pondérée

$$\bar{y}_w = \sum_{c=1}^{C} p_c \bar{y}_{0c} = \sum_{c=1}^{C} w_c p_{0c} \bar{y}_{0c}, \qquad (3)$$

où les répondants dans la cellule c sont pondérés par l'inverse du taux de réponse $w_c = p_c / p_{0c}$. L'estimateur (3) peut être considéré comme un cas particulier de l'estimateur par régression, où les valeurs manquantes sont imputées par la régression de Y sur les indicateurs pour les

cellules d'ajustement. Nous comparons le biais et l'erreur quadratique moyenne de (2) et (3) sous le modèle suivant, qui traduit les caractéristiques importantes du problème. Nous supposons que, sachant la taille d'échantillon *n*, les cas échantillonnés suivent une loi multinomiale sur le tableau de contingence $(C \times 2)$ basé sur la classification de *M* et *X*, avec les probabilités de cellule

$$\Pr(M = 0,\ X = c) = \phi\pi_{0c};\ \Pr(M = 1,\ X = c) = (1-\phi)\pi_{1c},$$

où $\phi = \Pr(M = 0)$ est la probabilité marginale de réponse. La loi conditionnelle de *X* sachant $M = 0$ et n_0 est multinomiale avec les probabilités de cellule $\Pr(X = c \mid M = 0) = \pi_{0c}$, et la loi marginale de *X* sachant *n* est multinomiale avec l'indice *n* et les probabilités de cellule

$$\Pr(X = c) = \phi\pi_{0c} + (1-\phi)\pi_{1c} = \pi_c,$$

disons. Nous supposons que la loi conditionnelle de *Y* sachant $M = m,\ X = c$ est de moyenne μ_{mc} et de variance constante σ^2. Les moyennes de *Y* pour les répondants et les non-répondants sont

$$\mu_0 = \sum_{c=1}^{C} \pi_{0c}\mu_{0c},\ \mu_1 = \sum_{c=1}^{C} \pi_{1c}\mu_{1c},$$

respectivement, et la moyenne globale de *Y* est $\mu = \phi\mu_0 + (1-\phi)\mu_1$.

Sous ce modèle, la moyenne et la variance conditionnelles de $\bar{y}_w$ sachant $\{p_c\}$ sont, respectivement, $\sum_{c=1}^{C} p_c\mu_{0c}$ et $\sigma^2 \sum_{c=1}^{C} p_c^2 / n_{0c}$. Donc, le biais de $\bar{y}_w$ est

$$b(\bar{y}_w) = \sum_{c=1}^{C} \pi_c(\mu_{0c} - \mu_c),$$

où π_c et μ_c sont la proportion et la moyenne de population de *Y* dans la cellule *c*. Cela peut s'écrire

$$b(\bar{y}_w) = \tilde{\mu}_0 - \mu, \tag{4}$$

où $\tilde{\mu}_0 = \sum_{c=1}^{C} \pi_c \mu_{0c}$ est la moyenne des répondants « corrigée » pour les covariables et $\mu = \sum_{c=1}^{C} \pi_c \mu_c$ est la vraie moyenne de population de *Y*. La variance de $\bar{y}_w$ est égale à la somme de la valeur prévue de la variance conditionnelle et de la variance de son espérance conditionnelle, et est approximativement

$$V(\bar{y}_w) = (1+\lambda)\sigma^2 / n_0 + \sum_{c=1}^{C} \pi_c(\mu_{0c} - \tilde{\mu}_0)^2 / n, \tag{5}$$

où $\lambda = \sum_{c=1}^{C} \pi_{0c}((\pi_c / \pi_{0c} - 1)^2)$ est l'analogue de population de la variance des poids de non-réponse $\{w_c\}$, qui est identique à *L* dans l'équation (1), puisque les poids sont rééchelonnés de sorte que leur moyenne soit égale à un. La formule de la variance de la moyenne pondérée dans Oh et Scheuren (1983), dérivée sous la perspective de quasi-randomisation, se réduit à (5) si l'on suppose que la variance dans les cellules est constante et que l'on ignore les corrections pour population finie et les termes d'ordre $1/n^2$. L'erreur quadratique moyenne (eqm) de $\bar{y}_w$ est alors

$$\text{eqm}(\bar{y}_w) = b^2(\bar{y}_w) + V(\bar{y}_w). \tag{6}$$

L'erreur quadratique moyenne de la moyenne non pondérée (2) est donnée par

$$\text{eqm}(\bar{y}_0) = b^2(\bar{y}_0) + V(\bar{y}_0), \tag{7}$$

où

$$b(\bar{y}_0) = b(\bar{y}_w) + \mu_0 - \tilde{\mu}_0, \tag{8}$$

est le biais et

$$V(\bar{y}_0) = \sigma^2 / n_0 + \sum_{c=1}^{C} \pi_{0c}(\mu_{0c} - \mu_0)^2 / n_0, \tag{9}$$

est la variance. Donc, la différence (disons Δ) entre les erreurs quadratiques moyennes est

$$\begin{aligned} &\Delta = \text{eqm}(\bar{y}_0) - \text{eqm}(\bar{y}_w) = B + V_1 - V_2, \text{ où} \\ &B = (\mu_0 - \tilde{\mu}_0)^2 + 2(\mu_0 - \tilde{\mu}_0)(\tilde{\mu}_0 - \mu), \\ &V_1 = \sum_{c=1}^{C} \pi_{0c}(\mu_{0c} - \mu_0)^2 / n_0 - \sum_{c=1}^{C} \pi_c(\mu_{0c} - \tilde{\mu}_0)^2 / n, \\ &V_2 = \lambda\sigma^2 / n_0 \end{aligned} \tag{10}$$

l'équation (10) et son interprétation détaillée fournissent les résultats principaux de l'article; il convient de souligner que les termes positifs dans (10) favorisent l'estimateur pondéré $\bar{y}_w$.

a) Le premier terme *B* représente l'effet sur l'erreur quadratique moyenne de la réduction du biais due à l'ajustement sur les covariables. Il est d'ordre un et domine de plus en plus la EQM à mesure qu'augmente la taille de l'échantillon. Si $\mu \le \tilde{\mu}_0 < \mu_0$ ou $\mu_0 < \tilde{\mu}_0 \le \mu$, alors la repondération a réduit le biais de la moyenne des répondants et les deux composantes de *B* sont positives. En particulier, si les données manquantes le sont au hasard (Rubin 1976; Little et Rubin 2002), en ce sens que les répondants constituent un échantillon aléatoire des cas échantillonnés dans chaque cellule *c*, alors $\tilde{\mu}_0 = \mu$ et la repondération élimine le biais de la moyenne non pondérée. La correction du biais est

$$\mu_0 - \tilde{\mu}_0 \simeq \sum_{c=1}^{C} \pi_{0c}(1 - w_c)(\mu_{0c} - \mu_0),$$

si l'on ignore les différences entre les poids et leurs espérances. Il s'agit de zéro à $O(1)$ si la non-réponse n'est pas reliée aux cellules d'ajustement (auquel cas $w_c \approx 1$ pour tout *c*) ou que le résultat n'est pas relié aux cellules d'ajustement (auquel cas $\mu_{0c} \approx \mu_0$ pour tout *c*). Donc, une réduction importante du biais nécessite des variables de cellules d'ajustement reliées à la fois à la non-réponse et au résultat d'intérêt, fait qui a été souligné par plusieurs auteurs. On pense souvent que le conditionnement sur les caractéristiques observées des non-répondants réduira le biais, mais il

convient de souligner que cela n'est pas garanti; il est possible que la moyenne corrigée s'écarte davantage, en moyenne, de la moyenne réelle que la moyenne non corrigée, auquel cas la repondération empire le biais.

b) L'effet de la repondération sur la variance est représenté par $V_1 - V_2$.

c) Pour les résultats Y qui ne sont pas reliés aux cellules d'ajustement, $\mu_{0c} = \mu_0$ pour tout c, $V_1 = 0$, et la repondération accroît la variance, puisque V_2 est positive. L'équation (10) réduit alors la version en population de la formule (1) de Kish. Les variables d'ajustement de cellule qui sont de bons prédicteurs de la non-réponse font plus de tort que de bien dans cette situation, puisqu'elles accroissent la variance des poids sans réduire le biais; mais il n'existe aucun compromis entre le biais et la variance pour ces résultats, puisqu'il n'y aucune réduction du biais.

d) Si la variable de cellule d'ajustement X n'est pas reliée à la non-réponse, alors λ est $O(1/n)$ et, par conséquent, V_2 a un ordre de variabilité plus faible que V_1. Le terme V_1 a tendance à être positif, puisque $\sum_{c=1}^{C} \pi_{0c}(\mu_{0c} - \mu_0)^2 \simeq \sum_{c=1}^{C} \pi_{0c}(\mu_{0c} - \tilde{\mu}_0)^2$, et le diviseur n dans le deuxième terme est plus grand que le diviseur n_0 dans le premier terme. Donc, dans ce cas, la repondération a tendance à n'avoir aucun effet sur le biais, mais elle réduit la variance dans la mesure où X est un bon prédicteur du résultat. Ceci contredit la notion selon laquelle la repondération accroît la variance. La « superefficacité » mentionnée plus haut qui résulte de l'estimation des poids de non-réponse à partir de l'échantillon est illustrée par le fait que, si les données sont manquantes complètement au hasard, alors le « vrai » poids de non-réponse est une constante pour toutes les unités répondantes. Par conséquent, la pondération au moyen des « vrais » poids produit (2), qui est moins efficace que la pondération par les poids « estimés », qui produit (3).

e) Si la variable de cellule d'ajustement est un bon prédicteur du résultat et également un prédicteur de la non-réponse, alors la valeur de V_2 est de nouveau faible, parce que la variance résiduelle σ^2 est réduite et celle de V_1 est généralement positive en vertu d'un argument semblable à celui exposé au point d). Le terme $\sum_{c=1}^{C} \pi_{0c}(\mu_{0c} - \mu_0)^2$ peut s'écarter plus de $\sum_{c=1}^{C} \pi_c(\mu_{0c} - \tilde{\mu}_0)^2$, parce que les poids sont moins semblables, mais cette différence pourrait être positive ou négative, et les différents diviseurs semblent plus susceptibles de déterminer le signe et la taille de V_1. Donc, la repondération a tendance à réduire à la fois le biais et la variance dans ce cas-ci.

f) L'équation (9) peut être appliquée au cas de la post-stratification sur les chiffres de population, en posant que n représente la taille de la population plutôt que la taille d'échantillon. Si l'on suppose que la population est grande, le deuxième terme de V_1 disparaît essentiellement, ce qui augmente la possibilité de réduction de la variance quand les variables formant les post-strates sont prédictives du résultat. Cette observation est une réplication de résultats antérieurs sur la post-stratification (Holt et Smith 1979; Little 1993).

Un sommaire qualitatif simple des résultats a) à f) de la section 2 est présenté au tableau 1, qui indique la direction du biais et la variance quand les associations entre les cellules d'ajustement et le résultat ainsi que l'indicateur de données manquants sont fortes ou faibles. De toute évidence, la repondération n'est efficace que pour les résultats qui sont associés à la variable de cellule d'ajustement, puisque autrement, elle accroît la variance sans réduction compensatoire du biais. Pour les résultats qui sont associés à la variable de cellule d'ajustement, la repondération accroît la précision et réduit également le biais si les variables de cellule d'ajustement sont reliées à la non- réponse.

Tableau 1
Effet de la repondération sur le biais et sur la variance d'une moyenne, selon la force de l'association des variables de cellule d'ajustement avec la non-réponse et le résultat

	Association avec le résultat	
Association avec la non-réponse	Faible	Forte
Faible	Cellule 1 Biais : --- Var : ---	Cellule 3 Biais : --- Var : ↓
Forte	Cellule 2 Biais : --- Var : ↑	Cellule 4 Biais : ↓ Var : ↓

Il est utile de disposer d'estimations de la EQM de $\bar{y}_0$ et $\bar{y}_w$ qui peuvent être calculées d'après les données observées. Soit $s_{0c}^2 = \sum_{i \in c}(y_i - \bar{y}_{0c})^2/(n_{0c} - 1)$ la variance d'échantillon des répondants dans la cellule c, $s^2 = \sum_{c=1}^{C}(n_{0c} - 1)s_{0c}^2/(n_{0c} - C)$ la variance globale à l'intérieur des cellules et $s_0^2 = \sum_{i=1}^{n_0}(y_i - \bar{y}_0)^2/(n_0 - 1)$ la variance d'échantillon totale des valeurs des réponses. Nous utilisons les expressions approximativement sans biais qui suivent, sous l'hypothèse que les données sont manquantes au hasard (MAR pour *missing at random*) :

$$\hat{\text{eqm}}(\bar{y}_0) = \hat{B}^2(\bar{y}_0) + \hat{V}(\bar{y}_0), \qquad (11)$$

où $\hat{V}(\bar{y}_0) = s_0^2/n_0$ et

$$\hat{B}^2(\bar{y}_0) = \max\{0, (\bar{y}_w - \bar{y}_0)^2 - V_d\}$$

$$V_d = (n_1/n)^2 \begin{pmatrix} \sum_{c=1}^{C} p_{1c}(\bar{y}_{0c} - \bar{y}_0^{(1)})^2 / n_1 \\ + \sum_{c=1}^{C} p_{0c}(\bar{y}_{0c} - \bar{y}_0)^2 / n_0 \\ + s^2 \sum_{c=1}^{C} (p_{1c} - p_{0c})^2 / n_{0c} \end{pmatrix}, \qquad (12)$$

où $\bar{y}_0^{(1)} = \sum_{c=1}^{C} p_{1c}\bar{y}_{0c}$, et V_d estime la variance de $(\bar{y}_w - \bar{y}_0)$ et est inclus dans (12) comme correction du biais pour $(\bar{y}_w - \bar{y}_0)^2$ en tant qu'estimation de $B^2(\bar{y}_0)$, en suivant l'exemple de Little et coll. (1997). En outre

$$\hat{\text{eqm}}(\bar{y}_w) = \hat{V}(\bar{y}_w) = (1+L)s^2/n_0 + \sum_{c=1}^{C} p_c(\bar{y}_{0c} - \bar{y}_w)^2/n. \quad (13)$$

Si nous soustrayons (11) de (13), l'écart entre les erreurs quadratiques moyennes de $\bar{y}_w$ et $\bar{y}_0$ est alors estimé par

$$D = Ls^2/n_0 - (s_0^2 - s^2)/n_0 + \sum_{c=1}^{C} p_c(\bar{y}_{0c} - \bar{y}_w)^2/n - \hat{B}^2(\bar{y}_0). \quad (14)$$

Il s'agit du perfectionnement de (1) que nous proposons, qui est représenté par le premier terme du deuxième membre de (14).

3. Étude en simulation

Nous incluons des simulations pour illustrer le biais et la variance de la moyenne pondérée et non pondérée pour des ensembles de paramètres représentant chaque cellule du tableau 1. Nous comparons aussi les approximations analytiques de l'erreur quadratique moyenne (MSE) dans les équations (6) et (7) et leurs estimations fondées sur un échantillon (11) et (13) à la MSE empirique sur des échantillons répétés.

3.1 Paramètres de superpopulation

Les spécifications de la simulation pour la loi conjointe de X et M sont décrites au tableau 2. L'échantillon suit approximativement une loi uniforme sur la variable de cellule d'ajustement X, qui compte $C = 10$ cellules. Nous choisissons deux taux de réponse marginaux, soit 70 %, qui correspond à une valeur de sondage typique et 52 %, qui est une valeur plus extrême pour accentuer les différences entre les méthodes. Nous simulons trois lois de M sachant X afin de modéliser une association forte, moyenne ou faible.

Les lois simulées de la variable d'intérêt Y sachant $M = m$, $X = c$ sont présentées au tableau 3. Elles ont toutes la forme

$$[Y \mid M = m, X = c] \sim N(\beta_0 + \beta_1 X, \sigma^2).$$

Trois ensembles de valeurs de (β_1, σ^2) sont simulés pour modéliser les associations forte, moyenne et faible entre Y et X. L'ordonnée à l'origine β_0 est choisie de sorte que la moyenne globale de Y soit $\mu = 26{,}3625$ pour chaque scénario.

Nous avons simulé un biais d'échantillon répété de taille $n = 400$ et $n = 2\,000$ pour chaque combinaison de paramètres dans les tableaux 2 et 3. Les échantillons pour lesquels $n_{0c} = 0$ pour tout c ont été exclus, puisque l'estimation pondérée ne peut être calculée; en pratique, certaines cellules seraient probablement groupées dans de tels cas. Les nombres de simulations exclues sont présentés au tableau 4.

Tableau 2
Pourcentage de cas échantillonnés dans la cellule d'ajustement X et la cellule d'indication de données manquantes M

a. Taux de réponse global = 52 %

	Association entre M et X	X	1	2	3	4	5	6	7	8	9	10
1.	Forte	$M = 0$	0,55	1,00	4,01	4,52	5,04	5,55	6,06	6,58	9,14	9,96
		$M = 1$	8,69	9,00	6,01	5,53	5,04	4,54	4,04	3,54	1,02	0,20
2.	Moyenne	$M = 0$	2,77	3,50	4,01	4,52	5,04	5,55	6,06	6,58	7,11	7,62
		$M = 1$	6,47	6,50	6,01	5,53	5,04	4,54	4,04	3,54	3,05	2,54
3.	Faible	$M = 0$	4,62	5,15	5,21	5,28	5,34	5,40	5,45	5,52	5,58	5,64
		$M = 1$	4,62	4,85	4,81	4,77	4,73	4,69	4,65	4,60	4,57	4,52

b. Taux de réponse global = 70 %

	Association entre M et X	X	1	2	3	4	5	6	7	8	9	10
1.	Forte	$M = 0$	0,55	3,00	6,51	7,04	7,55	8,07	8,59	9,11	9,64	9,96
		$M = 1$	8,69	7,00	3,51	3,02	2,52	2,02	1,52	1,01	0,51	0,20
2.	Moyenne	$M = 0$	4,44	5,30	5,81	6,33	6,85	7,37	7,88	8,40	8,93	9,45
		$M = 1$	4,80	4,70	4,21	3,72	3,22	2,72	2,22	1,72	1,22	0,71
3.	Faible	$M = 0$	6,19	6,85	6,91	6,98	7,05	7,11	7,17	7,24	7,31	7,37
		$M = 1$	3,05	3,15	3,11	3,07	3,02	2,98	2,93	2,88	2,84	2,79

Tableau 3
Paramètres pour $[Y \mid M = m, X = c] \sim N(\beta_0 + \beta_1 c, \sigma^2)$

	Association entre Y et X	β_1	σ^2	ρ^2
1.	Forte	4,75	46	$\approx 0{,}80$
2.	Moyenne	3,70	122	$\approx 0{,}48$
3.	Faible	0,00	234	0,00

Nous avons simulé un biais d'échantillon répété de taille $n = 400$ et $n = 2\,000$ pour chaque combinaison de paramètres dans les tableaux 2 et 3. Les échantillons pour lesquels $n_{0c} = 0$ pour tout c ont été exclus, puisque l'estimation pondérée ne peut être calculée; en pratique, certaines cellules seraient probablement groupées dans de tels cas. Les nombres de simulations exclues sont présentés au tableau 4.

Tableau 4
Nombre de répliques exclues parce qu'une cellule ne contenait aucun répondant

		Taux de réponse	
Association entre M et X	Association entre Y et X	52%	70%
Forte	Forte	134	113
	Moyenne	120	117
	Faible	131	104
Moyenne	Faible	1	0

3.2 Comparaisons du biais, de la variance et de la racine de l'erreur quadratique moyenne, et leurs estimations

Les résultats sommaires du calcul empirique du biais et de la racine de l'erreur quadratique moyenne (REQM) sont présentés au tableau 5. Nous pouvons comparer la valeur empirique de la REQM de la moyenne pondérée aux estimations suivantes, présentées au tableau 5, qui correspondent à la moyenne sur 1 000 répliques, à savoir la REQM estimée d'après l'équation (1) empirique de Kish, c'est-à-dire

$$\hat{\text{eqm}}_{\text{Kish}}(\bar{y}_w) = (1 + L)s_Y^2 / n_0,$$

$$\text{où } s_Y^2 = \sum_{i=1}^{n_0} (y_i - \bar{y}_0)^2 / (n_0 - 1), \qquad (15)$$

la REQM analytique provenant des équations (6) et (7), et la REQM estimée d'après les équations (11) et (13).

Comme l'ont suggéré Oh et Scheuren (1983), nous incluons dans les deux dernières colonnes du tableau 5 le biais et la REQM empirique moyens d'une moyenne composite basée sur le choix, entre $\bar{y}_w$ et $\bar{y}_0$, de celle dont l'estimation de l'erreur quadratique moyenne fondée sur l'échantillon est la plus faible. Le biais empirique relativement au paramètre de population est présenté pour tous les estimateurs. Nous incluons aussi le biais et la RMSE de la moyenne avant suppression des cas de non-réponse.

Le tableau 5a donne les résultats des simulations pour un taux de réponse de 52 %. Les lignes sont étiquetées d'après les quatre cellules du tableau 1, où les associations moyenne et forte sont combinées. Pour chaque ligne, la plus faible des REQM pour les moyennes des répondants non pondérée et pondérée figure en caractères gras, indiquant la supériorité de la méthode correspondante.

Les quatre premières lignes du tableau 5a correspondent à la cellule 4 du tableau 1, avec une association moyenne/forte entre Y et X, et une association moyenne/forte entre M et X. Dans ces cas, la REQM de $\bar{y}_w$ est beaucoup plus faible que celle de $\bar{y}_0$, témoignant du fait qu'une part importante du biais de $\bar{y}_0$ est éliminée par la repondération.

Les deux lignes suivantes du tableau 5a correspondent à la cellule 3 du tableau 1, avec une association moyenne/forte entre Y et X, et une association faible entre M et X. Dans ces cas, le biais de $\bar{y}_0$ n'est plus important, mais la précision de $\bar{y}_w$ s'est améliorée, particulièrement quand l'association entre Y et X est forte. Il s'agit de cas où la repondération réduit la variance au lieu de l'accroître. Les estimations analytiques de la RMSE et les estimations sur échantillon s'approchent des estimations empiriques de la RMSE, tandis que l'équation de Kish la surestime, comme le prédit la théorie exposée à la section 2.

Les deux lignes suivantes du tableau 5a correspondent à la cellule 2 du tableau 1, où l'association entre Y et X est faible et l'association entre M et X est moyenne ou forte. Dans ces cas, la MSE de $\bar{y}_w$ est plus grande que celle de $\bar{y}_0$. Ces cas illustrent des situations où la repondération accroît la variance, sans réduction compensatoire du biais. La dernière ligne correspond à la cellule 1 du tableau 1, avec des associations faibles entre M et X et entre Y et X. La moyenne non pondérée a une REQM plus faible dans ces conditions, mais l'accroissement de la REQM dû à la repondération est négligeable. Pour les trois dernières lignes du tableau 5a, la REQM pour l'équation de Kish est semblable à celle produite au moyen de la formule analytique de la section 2 et aux estimations empiriques fondées sur cette formule, et toutes ces estimations s'approchent de la REQM empirique.

Les deux dernières colonnes du tableau 5a donnent le biais et la REQM empiriques avec la méthode composite consistant à choisir $\bar{y}_w$ ou $\bar{y}_0$ d'après la REQM estimée. Pour les simulations des six premières lignes, l'estimateur composite est le même que $\bar{y}_w$ et, par conséquent, détecte et élimine le biais de la moyenne non pondérée. Pour les simulations dans la cellule 1 (dernière ligne), l'estimateur composite donne les mêmes résultats que $\bar{y}_w$ ou $\bar{y}_0$, comme il fallait s'y attendre puisque $\bar{y}_w$ et $\bar{y}_0$ donnent des résultats très semblables dans ce cas. Pour les simulations dans la cellule 2 qui ne sont pas favorables à la pondération, la racine de l'erreur quadratique moyenne de l'estimateur

composite est plus faible que celle de $\bar{y}_w$, mais considérablement plus élevée que celle de $\bar{y}_0$, ce qui donne à penser que, pour les conditions de cette simulation, la EQM empirique donne la capacité limitée de choisir le meilleur estimateur dans des échantillons individuels.

Néanmoins, l'estimateur composite est, dans l'ensemble, le meilleur des trois pris en considération dans cette simulation.

Le tableau 5b donne les résultats pour le taux de réponse de 70 %. Le profil des résultats est fort semblable à celui du tableau 5a. Comme prévu, les différences entre les méthodes sont plus petites, quoiqu'elles demeurent considérables pour nombre de lignes du tableau.

Tableau 5a

Résultats sommaires des estimateurs basés sur 1 000 échantillons répétés pour $C = 10$ cellules d'ajustement, limitées aux échantillons répétés pour lesquels $n_{0c} > 0$ pour tout c. Le taux de réponse est de 52 %. Les valeurs sont multipliées par 1 000

Association avec les cellules d'ajustement basée sur X				Moyenne non pondérée				Moyenne pondérée					Moyenne avant suppression		Moyenne composite	
Cellule	(M, X)	(Y, X)	n	Biais emp.	REQM emp.	REQM analytique[1]	REQM est.[2]	Biais emp.	REQM emp.	REQM Kish[3]	REQM analytique[4]	REQM est.[5]	Biais emp.	REQM emp.	Biais emp.	REQM emp.
4	Forte	Forte	400	6 955	7 024	7 055	6 974	0	**1 057**	1 410	956	988	−38	795	0	1 057
			2 000	7 008	7 020	7 006	7 015	−2	**424**	608	427	434	12	342	−2	424
4	Forte	Moyenne	400	5 376	5 471	5 536	5 404	−33	**1 264**	1 510	1 216	1 297	−21	776	−33	1 264
			2 000	5 424	5 441	5 466	5 466	−41	**561**	650	545	559	−30	338	−41	561
4	Moyenne	Forte	400	3 664	3 794	3 809	3 754	−4	**816**	1 071	835	842	6	741	−4	816
			2 000	3 703	3 731	3 700	3 712	7	**369**	473	373	374	4	337	7	369
4	Moyenne	Moyenne	400	2 838	3 006	3 042	2 991	−18	**938**	1 095	954	970	−9	747	−18	938
			2 000	2 864	2 900	2 898	2 893	−2	**426**	483	426	428	6	335	-2	426
3	Faible	Forte	400	476	1 148	1 113	1 178	40	**823**	1 050	823	828	30	764	40	823
			2 000	376	587	614	595	−11	**361**	465	368	368	−3	333	−11	361
3	Faible	Moyenne	400	350	1 106	1 095	1 134	13	**927**	1 063	925	939	−16	762	13	927
			2 000	287	565	563	559	−20	**429**	470	413	414	−22	353	−20	429
2	Forte	Faible(0)	400	56	**1 070**	1 056	1 275	96	1 658	1 613	1 518	1 631	28	793	83	1 410
			2 000	−11	**464**	473	567	−26	698	698	679	699	−19	337	−25	620
2	Moyenne	Faible(0)	400	9	**1 042**	1 053	1 077	−27	1 122	1 112	1 097	1 125	21	772	−12	1 074
			2 000	−4	**474**	471	480	−11	491	491	491	493	11	340	-9	481
1	Faible	Faible(0)	400	−30	**1 038**	1 050	1 055	−30	1 053	1 064	1 050	1 076	−30	752	−30	1 040
			2 000	−2	**472**	469	469	−1	474	470	469	471	−8	343	−1	472

[1] Calculée en utilisant l'équation (7)
[2] Calculée en utilisant l'équation (11)
[3] Calculée en utilisant l'équation (15)
[4] Calculée en utilisant l'équation (6)
[5] Calculée en utilisant l'équation (13)

Tableau 5b

Résultats sommaires des estimateurs basés sur 1 000 échantillons répétés pour $C = 10$ cellules d'ajustement, limitées aux échantillons répétés pour lesquels $n_{0c} > 0$ pour tout c. Le taux de réponse est de 70 %. Les valeurs sont multipliées par 1 000

Association avec les cellules D'ajustement basée sur X				Moyenne non pondérée				Moyenne pondérée					Moyenne avant suppression		Moyenne composite	
Cellule	(M, X)	(Y, X)	n	Biais emp.	REQM emp.	REQM analytique[6]	REQM est.[7]	Biais emp.	REQM emp.	REQM Kish[8]	REQM analytique[9]	REQM est.[10]	Biais emp.	REQM emp.	Biais emp.	REQM emp.
4	Forte	Forte	400	4 692	4 810	4 893	4 860	−133	**1 129**	1 192	889	894	−129	998	−133	1 129
			2 000	4 827	4 841	4 839	4 854	−20	**400**	529	398	405	−5	334	−20	400
4	Forte	Moyenne	400	3 581	3 716	3 855	3 733	−133	**1 266**	1 250	1 075	1 097	−128	917	−127	1 284
			2 000	3 763	3 784	3 778	3 777	−9	**501**	554	481	490	11	343	−9	501
4	Moyenne	Forte	400	2 666	2 812	2 878	2 837	−58	**803**	910	794	796	−49	772	−58	803
			2 000	2 732	2 760	2 767	2 761	−6	**353**	406	355	355	−9	333	−6	353
4	Moyenne	Moyenne	400	2 104	2 282	2 315	2 291	−28	**833**	924	854	861	−43	751	−28	833
			2 000	2 146	2 180	2 170	2 165	13	**370**	411	382	382	10	334	13	370
3	Faible	Forte	400	217	906	954	980	−81	**797**	911	790	793	−77	771	−81	797
			2 000	312	513	506	502	2	**365**	405	353	353	4	349	2	365
3	Faible	Moyenne	400	251	922	942	960	15	**804**	916	845	852	26	727	15	804
			2 000	224	454	472	471	−14	**370**	408	378	379	−15	327	−14	370
2	Forte	Faible(0)	400	0	**952**	915	1 131	35	1 445	1 349	1 298	1 358	1	807	26	1 292
			2 000	−11	**416**	409	485	−41	608	598	580	599	−4	347	−31	535
2	Moyenne	Faible(0)	400	22	**911**	910	920	24	942	936	930	946	2	757	21	925
			2 000	23	**418**	407	411	20	425	416	416	417	15	344	19	420
1	Faible	Faible(0)	400	1	**914**	914	912	2	917	916	914	926	−5	751	1	914
			2 000	4	**402**	408	408	4	403	409	408	410	6	331	4	402

[6] Calculée en utilisant l'équation (7)
[7] Calculée en utilisant l'équation (11)
[8] Calculée en utilisant l'équation (15)
[9] Calculée en utilisant l'équation (6)
[10] Calculée en utilisant l'équation (13)

4. Discussion

Les résultats des sections 2 et 3 ont des incidences importantes en ce qui concerne l'utilisation de la pondération comme outil de correction pour la non-réponse totale. Les enquêtes comptent souvent de nombreuses variables de résultat auxquelles est habituellement appliquée la même pondération. L'analyse de la section 2 et les simulations de la section 3 donnent à penser que l'on pourrait obtenir de meilleurs résultats en estimant l'erreur quadratique moyenne des moyennes pondérée et non pondérée, et en limitant la pondération aux cas pour lesquels cette relation est importante. Une approche plus perfectionnée consiste à appliquer des modèles à effets aléatoires pour réduire les coefficients de pondération, de telle façon que la réduction soit plus importante pour les résultats qui ne sont pas fortement corrélés aux covariables (par exemple, Elliott et Little 2000). Une autre option souple est l'imputation fondée sur les modèles de prédiction, puisque ces derniers permettent d'utiliser des prédicteurs échelonnés par intervalles ainsi que des prédicteurs catégoriques, et de laisser tomber les interactions afin d'intégrer un plus grand nombre d'effets principaux. L'imputation multiple (Rubin 1987) peut être utilisée pour propager l'incertitude.

Quand on dispose de beaucoup d'information sur les covariables, une façon intéressante d'aborder la généralisation à des corrections par catégorie de pondération consiste à créer un score de propension pour chaque répondant d'après une régression logistique de l'indicateur de non-réponse sur les covariables, puis de créer des cellules d'ajustement de ce score. Les méthodes axées sur le score de propension ont été élaborées au départ dans le contexte des cas appariés et des témoins dans les études par observation (Rosenbaum et Rubin 1983), mais sont appliquées assez fréquemment aujourd'hui dans le contexte de la non-réponse totale (Little 1986; Czajka, Hirabayashi, Little et Rubin 1987; Ezzati et Khare 1992). Ici, l'analyse laisse entendre que, pour que cette approche soit productive, le score de propension doit être prédicteur des résultats. Vartivarian et Little (2002) considèrent des cellules d'ajustement basées sur la classification conjointe en fonction de la propension à répondre, ainsi que des prédicteurs sommaires des résultats afin de tirer parti des associations résiduelles entre les covariables et le résultat après correction pour tenir compte du score de propension. L'exigence que les variables de cellule d'ajustement prédisent les résultats étaye cette approche.

L'analyse présentée ici pourrait être étendue de plusieurs façons. Les termes de deuxième ordre figurant dans l'expression de la variance sont ignorés ici; s'ils étaient inclus, ils pénaliseraient la repondération fondée sur un grand nombre de petites cellules d'ajustement. Les corrections pour population finie pourraient être incluses, mais il semble peu probable qu'elles aient une incidence sur les principales conclusions. Il serait intéressant de voir dans quelle mesure les résultats peuvent être généralisés à des plans d'échantillonnage complexes comportant une mise en grappe et une stratification. En outre, il serait utile d'analyser minutieusement les effets de la pondération pour la non-réponse sur le biais et la variance d'autres statistiques que les moyennes, comme les moyennes de sous-classe ou les coefficients de régression. Nous nous attendons à ce que nombre de ces analyses indiquent aussi qu'il est important que les variables de cellule d'ajustement prédisent le résultat, mais d'autres points intéressants pourraient également s'en dégager.

Remerciements

L'étude a été financée par la subvention SES-0106914 de la National Science Foundation. Nous remercions un rédacteur adjoint et trois examinateurs de leurs commentaires constructifs au sujet d'ébauches antérieures.

Bibliographie

Czajka, J.L., Hirabayashi, S.M., Little, R.J.A. et Rubin, D.B. (1987). Evaluation of a new procedure for estimating income aggregates from advance data. Dans *Statistics of Income and Related Administrative Record Research*: *1986-1987*, U.S. Department of the Treasury, 109-136.

Elliott, M.R., et Little, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16, 191-209.

Ezzati, T., et Khare, M. (1992). Nonresponse adjustments in a National Health Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 339-344.

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

Holt, D., et Smith, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical, A*, 142, 33-46.

Kalton, G., et Kasprzyk, D. (1986). Le traitement des données d'enquêtes manquantes. *Techniques d'enquête*, 12, 1-16.

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

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

Little, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

Little, R.J.A., Lewitzky, S. Heeringa, S., Lepkowski, J. et Kessler, R.C. (1997). An assessment of weighting methodology for the national comorbidity study. *American Journal of Epidemiology*, 146, 439-449.

Little, R.J.A., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2 ième édition. New York: John Wiley & Sons, Inc.

Oh, H.L., et Scheuren, F.S. (1983). Weighting adjustments for unit nonresponse. Dans *Incomplete Data in Sample Surveys*, 2, Theory and Bibliographies, (Éds. W.G. Madow, I. Olkin et D.B. Rubin), Academic Press, New York, 143-184.

Robins, J.M., Rotnitsky, A. et Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846-866.

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

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

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

Vartivarian, S., et Little, R.J.A. (2002). On the formation of weighting adjustment cells for unit nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2005
Vol. 31, N° 2, pp. 185-200
Statistique Canada, N° 12-001-XPB au catalogue

Fonctions de variance-covariance pour les moyennes de domaine des questions avec valeurs ordonnées

Alistair James O'Malley et Alan Mark Zaslavsky [1]

Résumé

De nombreuses analyses statistiques, particulièrement l'analyse multiniveaux, requièrent l'estimation d'une matrice des variances-covariances d'échantillonnage. Dans le cas de problèmes univariés, des fonctions reliant la variance à la moyenne ont été utilisées pour obtenir des estimations de la variance, en regroupant l'information sur l'ensemble des unités ou des variables. Nous présentons des fonctions de variance et de corrélation pour des moyennes multivariées de questions d'enquête avec valeurs ordonnées, pour des données complètes, ainsi que pour des données avec non-réponse structurée. Nous élaborons aussi des méthodes permettant d'évaluer l'ajustement du modèle et de calculer des estimateurs composites qui combinent des prédictions directes et fondées sur un modèle. Nous utilisons des données d'enquête provenant de la Consumer Assessments of Health Plans Study (CAHPS®) pour illustrer l'application de la méthodologie.

Mots clés : Fonction de variance; fonction de corrélation; modèle hiérarchique; réponse ordonnée, non-réponse; enchaînement de questions.

1. Introduction

Les données d'enquête sont souvent utilisées pour obtenir des mesures permettant de faire des comparaisons entre domaines d'estimation. Dans notre exemple pratique, des enquêtes sont réalisées pour recueillir des déclarations sur les expériences vécues en ce qui concerne les régimes d'assurance-maladie (entités qui administrent les soins de santé) auprès des membres inscrits; de même, une enquête pourrait être conçue pour évaluer les écoles en faisant passer des tests à un échantillon d'élèves.

Une part essentielle de l'analyse des données d'enquête est le calcul des variances d'échantillonnage ou de la matrice des covariances d'échantillonnage d'un estimateur multivarié. L'approche type en échantillonnage consiste à calculer les variances directement pour chaque estimateur dans chaque domaine. Les estimations directes de la variance peuvent être instables si le nombre de répondants à une question est faible parce que la taille de l'échantillon pour un domaine est petite, vu que la question s'applique seulement à une fraction des répondants (comme les utilisateurs d'équipement spécialisé dans les enquêtes sur la santé), ou parce que nous souhaitons calculer les moyennes pour un petit sous-groupe (comme les personnes atteintes de maladie chronique).

En modélisant les estimations de la variance sous forme de fonctions des moyennes d'unité (domaine), nous pouvons regrouper l'information sur l'ensemble des unités pour obtenir des estimations plus stables. Bien que la modélisation puisse introduire un biais, pour les petites unités, ce problème est compensé par la réduction de la variation d'échantillonnage. On peut aussi envisager de généraliser les estimations de la variance sur l'ensemble des questions en plus des domaines, ou à la place de ceux-ci. Cette approche convient lorsqu'il existe des groupes de questions pour lesquelles il est probable que la même relation moyenne-variance soit vérifiée. Cependant, si le nombre de domaines est beaucoup plus grand que celui des questions, l'amélioration éventuelle la plus importante s'obtient en généralisant sur l'ensemble des domaines plutôt que sur l'ensemble des questions.

Une *fonction de variance généralisée* (FVG) est un modèle mathématique qui décrit la relation entre la variance ou la variance relative d'un estimateur et son espérance. Si plusieurs estimations sont produites d'après le même échantillon, Wolter (1985, chapitre 5) propose le modèle

$$V / M^2 = \theta_0 + \theta_1 / M,$$

où M et V représentent la valeur prévue et la variance de l'estimateur, respectivement. Une forme de ce genre pourrait convenir pour des variables, comme le revenu ou la richesse, pour lesquelles un coefficient de variation quasi constant serait plausible, parce que la moyenne et l'écart-type sont proportionnels à la longueur de la période de référence. La modélisation du coefficient de variation est donc pertinente surtout dans les situations où les variables ont un contenu semblable, mais des échelles différentes d'étendue non restreinte (par exemple, données sur le revenu recueillies mensuellement et annuellement). Dans notre problème, les questions ont des valeurs ordonnées, si bien qu'un modèle du coefficient de variation n'est pas un

1. Alistair James O'Malley et Alan Mark Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115-5899, États-Unis. Courriel : omalley@hcp.med.harvard.edu et zaslavsk@hcp.med.harvard.edu.

choix naturel. D'autres FVG proposées ont également une forme simple (Woodruff 1992; Otto et Bell 1995).

Trouver une FVG appropriée peut simplifier les calculs et rendre les estimations de la variance plus stables. En outre, résumer les estimations de la variance d'échantillonnage sous la forme d'une fonction facilite la présentation de grandes quantités de statistiques (Wolter 1985, pages 201–202). Enfin, modéliser les variances sous forme de fonctions des moyennes facilite la réestimation itérative des variances d'échantillonnage en cas de modélisation hiérarchique. En pratique, la décision d'utiliser des fonctions de variance dans un contexte de modélisation hiérarchique dépend de la qualité de l'ajustement de la FVG; l'utilisation de cette dernière ne vaut la peine que si l'ajustement est suffisamment bon.

Les études antérieures sur les FVG sont assez rares. Wolter (1985, chapitre 5) donne un aperçu, mais ne fournit que quelques références, comme le font aussi Valliant, Dorfman et Royall (2000, pages 344 à 348). Valliant (1992a, 1992b) utilise des FVG pour lisser des indices variables en fonction du temps dans les analyses de séries chronologiques. Woodruff (1992) utilise des FVG pour estimer la variance de la variation de l'emploi dans la Current Employment Survey, et Wolter (1985, pages 208 à 217) illustre l'utilisation des FVG au moyen de données provenant de la Current Population Survey. Des FVG sont également utilisées dans la National Health Interview Survey (Valliant et coll. 2000, page 344).

Huff, Eltinge et Gershunskaya (2002), ainsi que Cho, Eltinge, Gershunskaya et Huff (2002) considèrent l'utilisation de FVG pour la Current Employment Survey et la Consumer Expenditure Survey réalisées aux États-Unis. Eltinge (2002) utilise des FVG pour estimer une matrice des covariances d'échantillonnage complète, lorsque les échantillons sont trop petits pour produire des estimations stables pour toutes les régions, et estime les composantes de l'erreur quadratique moyenne (MSE pour *mean squared error*) du modèle de FVG. Otto et Bell (1995) ajustent des FVG au revenu médian, au revenu par habitant et au taux de pauvreté selon le groupe d'âge dans la Current Population Survey, en supposant que l'interdépendance des taux au cours du temps est autorégressive et que les matrices des covariances d'échantillonnage suivent une loi de Wishart.

Notre étude prolonge les travaux antérieurs sur les FVG dans quatre directions. En premier lieu, nous utilisons la FVG pour faire une généralisation sur l'ensemble des domaines plutôt que sur l'ensemble des questions. Donc, nous ne supposons pas que les diverses questions ont la même FVG, bien qu'il pourrait être raisonnable d'ajuster des modèles de la même forme pour les questions ayant des catégories de réponses semblables. En deuxième lieu, nous élaborons des FVG pour la matrice des covariances complète, qui doivent être estimées pour une inférence conjointe sur plusieurs résultats. En troisième lieu, nous nous concentrons sur la relation entre les moyennes et les variances des questions à format de réponse ordonné souvent utilisées dans les questionnaires d'enquête, plutôt que sur des réponses continues homoscédastiques. Enfin, nous tenons compte explicitement des profils de non-réponse due à des enchaînements de questions structurés. Alors qu'on peut ignorer la non-réponse partielle structurée (sauf son effet sur la taille d'échantillon) dans le cas de l'estimation univariée, il faut en tenir compte explicitement pour modéliser les relations bivariées, parce qu'elle a une incidence sur la covariance d'échantillonnage des moyennes des questions. De surcroît, comme le nombre de réponses varie selon la question, nous ne pouvons modéliser les covariances d'échantillonnage au moyen de la loi de Wishart, qui ne possède qu'un seul paramètre pour la taille d'échantillon.

Nous commençons par décrire l'estimation directe des variances et des covariances, y compris le cas où des données manquent à cause d'enchaînements de questions. À la section 3, nous présentons des modèles pour les fonctions de variance et de covariance généralisées (FVCG) et nous exposons nos stratégies d'ajustement et d'évaluation de modèles, et de combinaison d'estimations directes et de prédictions par modèle. À la section 4, nous appliquons nos méthodes à une grande enquête sur les soins de santé. À la section 5, pour conclure, nous décrivons des applications et des extensions de nos méthodes.

2. Estimations directes des variances d'échantillonnage des moyennes de domaine

Nous indiçons les observations par domaine (indice *h*), par question (indices *i* et *j*) et par répondant (indices *k* et *l*); $y_{h,ik}$ et $r_{h,ik}$ représentent le résultat et l'indicateur de réponse du sujet *k* dans le domaine *h* pour la question *i*. Nous supprimons l'indice inférieur de question quand nous faisons référence à l'ensemble des questions pour un répondant ou un domaine, et nous n'avons pas besoin d'utiliser d'indice inférieur de répondant quand nous discutons des moyennes, des variances et des corrélations de questions.

L'estimation directe de la matrice des covariances d'échantillonnage des moyennes de domaine (donc, « estimation de variance ») débute par l'expression des moyennes sous forme de fonctions des totaux des résultats et des indicateurs de réponse. Nous remplaçons $y_{h,ik}$ par 0 pour les observations manquantes, de sorte que les totaux soient définis en présence d'enchaînements de questions. Suivant la notation de Särndal, Swenson et Wretman (1992, pages 24 à 28; 36 à 42), soit U_h et S_h la population et

l'échantillon, respectivement, pour le h^{e} domaine, $Y_{h,i} = \sum_{U_h} y_{h,ik}$, $R_{h,i} = \sum_{U_h} r_{h,ik}$, $\hat{Y}_{h,i} = \sum_{S_h} \breve{y}_{h,ik}$, et $\hat{R}_{h,i} = \sum_{S_h} \breve{r}_{h,ik}$, où $\breve{y}_{h,ik} = y_{h,ik}/\pi_{h,k}$, $\breve{r}_{h,ik} = r_{h,ik}/\pi_{h,k}$, et $\pi_{h,k} = \text{pr}(k \in S_h)$.

Le vecteur des résultats moyens pour la population d'éléments compris dans le domaine *h* est

$$M_h = f(Y_h, R_h) = \left(\frac{Y_{h,1}}{R_{h,1}}, \ldots, \frac{Y_{h,I}}{R_{h,I}} \right),$$

où $Y_h = (Y_{h,1}, \ldots, Y_{h,I})$ et $R_h = (R_{h,1}, \ldots, R_{h,I})$. Un estimateur est donné par

$$f(\hat{Y}_h, \hat{R}_h) = \left(\frac{\hat{Y}_{h,1}}{\hat{R}_{h,1}}, \ldots, \frac{\hat{Y}_{h,I}}{\hat{R}_{h,I}} \right).$$

Un développement en série de Taylor de premier ordre de $f(\hat{Y}_h, \hat{R}_h)$ autour de $f(Y_h, R_h)$ produit l'approximation

$$\text{var}(f(\hat{Y}_h, \hat{R}_h)) \approx V_h = f'(Y_h, R_h)\, \text{var}(\hat{Y}_h, \hat{R}_h)\, f'(Y_h, R_h)^T,$$

où $f'(Y_h, R_h)$ est le Jacobien de $f(Y_h, R_h)$. Souvent, il est informatiquement plus facile de commencer par calculer $u_{h,k} = f'(Y_h, R_h) z_{h,k}$, où $z_{h,k} = (y_{h,k}, r_{h,k})$, puis d'évaluer la variance sous la forme

$$\begin{aligned} V_h &= \text{var}\left(\sum_{S_h} \breve{u}_{h,k} \right) \\ &= \text{var}\left(\sum_{U_h} \breve{u}_{h,k}\, I_{h,k} \right) \\ &= \sum_{k,l \in U_h} \Delta_{h,kl}\, \breve{u}_{h,k}\, \breve{u}_{h,l}^T, \end{aligned}$$

où $I_{h,k} = 1$ si $k \in S_h$ (indiquant que le k^{e} membre du domaine *h* est échantillonné) et 0 autrement, $\Delta_{h,kl} = \pi_{h,kl} - \pi_{h,k}\pi_{h,l}$, et $\pi_{h,kl} = \text{pr}(k, l \in S_h)$. Un estimateur de V_h est

$$\hat{V}_h = \sum_{k,l \in S_h} \breve{\Delta}_{h,kl}\, \breve{u}_{h,k}\, \breve{u}_{h,l}^T, \qquad (1)$$

où $\breve{\Delta}_{h,kl} = \Delta_{h,kl}/\pi_{h,kl}$.

Pour décrire l'évaluation de $\hat{V}_h$, nous ne devons considérer qu'un seul élément diagonal (c'est-à-dire, variance) et un seul élément hors diagonale (c'est-à-dire, covariance). La sous-matrice du Jacobien formée par les i^{e} et j^{e} questions est donnée par

$$f'(Y_h, R_h) = \begin{pmatrix} \dfrac{1}{R_{h,i}} & 0 & -\dfrac{Y_{h,i}}{R_{h,i}^2} & 0 \\ 0 & \dfrac{1}{R_{h,j}} & 0 & -\dfrac{Y_{h,j}}{R_{h,j}^2} \end{pmatrix}.$$

Pour $z_{h,k} = (y_{h,ik}, y_{h,jk}, r_{h,ik}, r_{h,jk})$, il s'ensuit que

$$u_{h,k} = f'(Y_h, R_h)\, z_{h,k} = \begin{pmatrix} \dfrac{1}{R_{h,i}}(y_{h,ik} - M_{h,i}\, r_{h,ik}) \\ \dfrac{1}{R_{h,j}}(y_{h,jk} - M_{h,j}\, r_{h,jk}) \end{pmatrix},$$

où $M_{h,i} = Y_{h,i}/R_{h,i}$ est le résultat moyen de la i^{e} question dans le domaine *h*. Donc,

$$\hat{V}_{h,ii} = \frac{1}{R_{h,i}^2} \sum_{k,l \in S_h} \breve{\Delta}_{h,kl} (\breve{y}_{h,ik} - M_{h,i}\, \breve{r}_{h,ik})(\breve{y}_{h,il} - M_{h,i}\, \breve{r}_{h,il}) \qquad (2)$$

et

$$\begin{aligned} \hat{V}_{h,ij} = \frac{1}{R_{h,i}\, R_{h,j}} \sum_{k,l \in S_h} \breve{\Delta}_{h,kl} &(\breve{y}_{h,ik} - M_{h,i}\, \breve{r}_{h,ik}) \\ &\times (\breve{y}_{h,jl} - M_{h,j}\, \breve{r}_{h,jl}). \end{aligned} \qquad (3)$$

Pour évaluer (2) et (3), nous faisons une autre approximation en substituant $\hat{R}_{h,i} = \sum_{S_h} \breve{r}_{h,ik}$ et $\hat{M}_{h,i} = \sum_{S_h} \breve{y}_{h,ik} / (\sum_{S_h} \breve{r}_{h,ik})$ à $R_{h,i}$ et $M_{h,i}$.

Si les taux d'échantillonnage sont faibles ou que nous souhaitons faire des prédictions pour une grande superpopulation (par exemple, tous les participants possibles à un régime d'assurance-maladie plutôt que ceux couramment inscrits seulement), $\breve{\Delta}_{h,kl} = 1 - \pi_{h,k} \approx 1$ si $k = l$, $\breve{\Delta}_{h,kl} \approx 0$ si $k \neq 1$, et le plan d'échantillonnage s'approche de l'échantillonnage avec remise. Sous le plan d'échantillonnage avec remise, les estimateurs approximativement sans biais sont

$$\hat{V}_{h,ii} = \frac{1}{\hat{R}_{h,i}^2} \sum_{k \in S_h} (\breve{y}_{h,ik} - \hat{M}_{h,i}\, \breve{r}_{h,ik})^2 \qquad (4)$$

et

$$\hat{V}_{h,ij} = \frac{1}{\hat{R}_{h,i}\, \hat{R}_{h,j}} \sum_{k \in S_h} (\breve{y}_{h,ik} - \hat{M}_{h,i}\, \breve{r}_{h,ik})(\breve{y}_{h,jk} - \hat{M}_{h,j}\, \breve{r}_{h,jk}). \qquad (5)$$

Ces estimateurs peuvent être généralisés pour prendre en compte la mise en grappes.

En cas d'échantillonnage avec probabilités égales dans les domaines, (4) et (5) se réduisent à

$$\hat{V}_{h,ii} = \frac{1}{\hat{R}_{S_{h,i}}^2} \sum_{k \in S_h} (y_{h,ik} - \hat{M}_{h,i}\, r_{h,ik})^2 \qquad (6)$$

et

$$\hat{V}_{h,ij} = \frac{1}{\hat{R}_{S_{h,i}} \hat{R}_{S_{h,j}}} \sum_{k \in S_h} (y_{h,ik} - \hat{M}_{h,i}\, r_{h,ik})(y_{h,jk} - \hat{M}_{h,j}\, r_{h,jk}), \qquad (7)$$

où $\hat{R}_{S_{h,i}}$ est le nombre de répondants à la question *i* dans le domaine *h*.

3. Modèles pour les fonctions de variance

À la présente section, nous proposons des spécifications de modèle pour les variances et pour les corrélations d'échantillon pour les réponses complètes ou pour celles avec enchaînements de questions structurés. Puis, nous discutons des stratégies d'ajustement et d'évaluation des modèles. Nous supposons que ces domaines sont des strates non chevauchantes, de sorte que les erreurs d'échantillonnage pour divers domaines soient indépendantes.

Nous transformons les évaluations ordonnées pour les amener à l'intervalle $[0,1]$ par la transformation $p_{h,i} = (B_{h,i} - M_{h,i})/(B_{h,i} - A_{h,i})$, où $A_{h,i}$ et $B_{h,i}$ sont les catégories de réponse minimale et maximale pour la question i dans le domaine h, respectivement. Nous nous concentrons sur la modélisation des variances pour les grandes valeurs de $M_{h,i}$ (petites valeurs de $p_{h,i}$) parce que, dans l'exemple que nous avons choisi, les résultats moyens sont habituellement proches de l'extrémité supérieure de l'échelle.

3.1 Fonctions de variance

Afin de tenir compte du nombre variable de répondants sur les domaines et les questions, et des différences d'échelle, nous normalisons les estimateurs de variance donnés par (6) pour la taille d'échantillon et faisons un rééchelonnement :

$$\tilde{V}_{h,ii} = \frac{\hat{R}_{S_{h,i}}\,\hat{V}_{h,ii}}{(B_{h,i} - A_{h,i})^2}.$$

Sous échantillonnage avec probabilités inégales dans les domaines, nous pourrions utiliser un facteur de normalisation qui tient compte des pondérations. Une normalisation possible consiste à multiplier $\hat{V}_{h,ii}$ par $\hat{R}^{*}_{S_{h,i}} = (\sum \breve{r}_{h,ik})^2 / (\sum \breve{r}^2_{h,ik})$, où $\breve{r}_{h,ik}$ est l'indicateur de réponse à la question i pour le k^{e} sujet dans le h^{e} domaine, à la place de $\hat{R}_{S_{h,i}}$. Cette approximation, proposée par Kish (1965), possède une justification fondée sur un modèle (Gabler, Haeder et Lahiri 1999). Elle donne de bons résultats si les probabilités d'échantillonnage varient moyennement dans l'échantillon, mais peut être inefficace si la variation est excessive (Korn et Graubard 1999, page 173; Spencer 2000).

Comme, dans notre exemple, les questions ont des valeurs ordonnées, la variance doit tendre vers 0 quand $p_{h,i} \to 0$ ou $p_{h,i} \to 1$. Un prédicteur ayant manifestement cette propriété est la fonction de variance de la loi de Bernoulli, $p_{h,i}(1 - p_{h,i})$. Celle-ci est vérifiée exactement pour les questions dichotomiques et pourrait être une approximation utile pour les questions comportant au moins trois catégories de réponse.

Comme autres solutions que le modèle de variance de Bernoulli, nous considérons des modèles contenant diverses fonctions polynomiales et autres des moyennes comme prédicteur. De tous les modèles envisagés, la famille de modèles quadratiques a donné des résultats d'ajustement aussi bons que n'importe quelle autre. Nous nous concentrons sur les modèles quadratiques qui suivent.

$$\text{Modèle V1:}\quad \tilde{V}_{h,ii} = \beta_{1i}\, p_{h,i}, \tag{8}$$

$$\text{Modèle V2:}\quad \tilde{V}_{h,ii} = \beta_{2i}\, p_{h,i}(1 - p_{h,i}), \tag{9}$$

$$\text{Modèle V3:}\quad \tilde{V}_{h,ii} = \beta_{1i}\, p_{h,i} + \beta_{2i}\, p_{h,i}(1 - p_{h,i}). \tag{10}$$

Donc, nous considérons un modèle de variance linéaire V1, un modèle de type binomial V2 et un modèle de variance quadratique général V3. Tous ces modèles assurent correctement que $\tilde{V}_{h,ii} = 0$ quand $p_{h,i} = 0$, mais seul V2 assure que $\tilde{V}_{h,ii} = 0$ quand $p_{h,i} = 1$. La logique qui sous-tend V1 est que les relations sont souvent approximativement linéaires sur de petits intervalles. Aussi bien V1 que V2 sont des sous-modèles du modèle quadratique à deux paramètres V3. Nous avons également considéré des modèles pour $\log(\tilde{V}_{h,ii})$, mais ceux-ci n'ont pas donné un aussi bon ajustement.

Le modèle V3 est équivalent au modèle proposé par Wolter (1985, chapitre 5); l'équivalence se voit en exprimant le deuxième membre de V3 en fonction de $p_{h,i}$ et de $p^2_{h,i}$, puis en divisant les deux membres par $p^2_{h,i}$ pour obtenir la variance relative. Cependant, les estimations des paramètres obtenues par ajustement des deux formes du modèle peuvent différer selon les hypothèses de modélisation utilisées.

3.2 Fonctions de corrélation avec données complètes

Comme les corrélations sont indépendantes de l'échelle des données, nous les modélisons et nous dérivons les covariances d'échantillonnage, au lieu de modéliser directement les covariances. Nous modélisons les corrélations d'échantillon

$$\hat{\rho}_{h,ij} = \frac{\hat{V}_{h,ij}}{(\hat{V}_{h,ii}\,\hat{V}_{h,jj})^{1/2}},$$

par la voie des valeurs transformées non contraintes $Z_{h,ij} = \log\{(1 + \hat{\rho}_{h,ij})/(1 - \hat{\rho}_{h,ij})\}$. Contrairement aux modèles de variance, les modèles de corrélation peuvent inclure une ordonnée à l'origine non contrainte, puisque la corrélation n'est sujette à aucune contrainte naturelle quand $p_{h,i}$ ou $p_{h,j}$ s'approche de 0 ou de 1.

Puisque $\hat{\rho}_{h,ij}$ est une fonction des premier et deuxième moments des questions i et j, il semble raisonnable de se concentrer d'abord sur les modèles linéaires et quadratiques pour $Z_{h,ij}$. Comme pour les fonctions de variance, nous constatons qu'une gamme plus étendue de modèles

(par exemple, les modèles avec logarithme des moyennes comme prédicteur) n'améliorent pas considérablement l'ajustement. En dernière analyse, nous retenons la série de modèles emboîtés qui suit.

$$\text{Modèle C1:}\quad Z_{h,ij} = \alpha_{0ij}, \tag{11}$$

$$\text{Modèle C2:}\quad Z_{h,ij} = \alpha_{0ij} + \alpha_{3ij}\ p_{h,i}\ p_{h,j}, \tag{12}$$

$$\text{Modèle C3:}\quad Z_{h,ij} = \alpha_{0ij} + \alpha_{1ij}(p_{h,i} + p_{h,j}) + \alpha_{3ij}\ p_{h,i}\ p_{h,j}, \tag{13}$$

$$\text{Modèle C4:}\quad Z_{h,ij} = \alpha_{0ij} + \alpha_{1ij} p_{h,i} + \alpha_{2ij} p_{h,j} + \alpha_{3ij}\ p_{h,i}\ p_{h,j}, \tag{14}$$

$$\text{Modèle C5:}\quad Z_{h,ij} = \alpha_{0ij} + \alpha_{1ij} p_{h,i} + \alpha_{2ij} p_{h,j} + \alpha_{3ij}\ p_{h,i}\ p_{h,j} + \alpha_{4ij}\ p_{h,i}^2 + \alpha_{5ij}\ p_{h,j}^2, \tag{15}$$

Le modèle C3 est le modèle C4 avec la contrainte $\alpha_{1ij} = \alpha_{2ij}$.

3.3 Prédiction des covariances avec données manquantes structurées

Lorsque les données comportent des enchaînements de questions, les corrélations d'échantillon des évaluations pour l'ensemble des répondants qui ont répondu aux deux questions peuvent être modélisées au moyen des modèles (11) à (15), comme dans le cas de la réponse complète. Il est facile d'estimer les covariances d'échantillon correspondantes en utilisant les fonctions de variance ajustées pour rééchelonner les corrélations prévues. Cependant, comme la covariance d'échantillonnage reflète la variabilité dans l'ensemble du processus d'échantillonnage, et non simplement la variabilité dans la sous-population de répondants qui ont répondu aux deux questions, la relation entre la covariance d'échantillon et la covariance d'échantillonnage est plus complexe que si les données étaient complètes. À la présente section, nous dérivons la relation entre la covariance d'échantillon pour l'ensemble de répondants qui ont répondu aux deux questions et la covariance d'échantillonnage. Cela nous permet d'appliquer des modèles de corrélation tels que (11) à (15) à des données avec enchaînements de questions.

Pour toute paire de questions, il existe quatre schémas de données, à savoir une réponse aux deux questions, une réponse et une question sautée (deux schémas), et deux questions sautées. Nous étendons notre notation en introduisant un indice supérieur représentant la situation de réponse à une deuxième question. Soit $\hat{Y}^1_{h,ij} = \sum_{S_h} \breve{y}_{h,ik}\breve{r}_{h,jk}$, $\hat{Y}^0_{h,ij} = \sum_{S_h} \breve{y}_{h,ik}(1-\breve{r}_{h,jk})$, $\hat{R}^1_{h,ij} = \sum_{S_h} \breve{r}_{h,ik}\breve{r}_{h,jk}$, $\hat{R}^0_{h,ij} = \sum_{S_h} \breve{r}_{h,ik}(1-\breve{r}_{h,jk})$, $\hat{M}^1_{h,ij} = \hat{Y}^1_{h,ij} / \hat{R}^1_{h,ij}$, $\hat{M}^0_{h,ij} = \hat{Y}^0_{h,ij} / \hat{R}^0_{h,ij}$. Alors

$$\hat{M}_{h,i} = \frac{\hat{R}^1_{h,ij}\,\hat{M}^1_{h,ij} + \hat{R}^0_{h,ij}\,\hat{M}^0_{h,ij}}{\hat{R}_{h,i}}.$$

Dans le cas de l'échantillonnage avec probabilités égales, le remplacement de $\hat{M}_{h,i}$ par l'expression susmentionnée dans (7) donne

$$\tilde{V}_{h,ij} = \frac{\hat{R}^1_{h,ij}}{\hat{R}_{h,i}\,\hat{R}_{h,j}}\left\{\hat{C}^1_{h,ij} + \frac{\hat{R}^0_{h,ij}\,\hat{D}_{h,ij}\,\hat{R}^0_{h,ji}\,\hat{D}_{h,ji}}{\hat{R}_{h,i}\,\hat{R}_{h,j}}\right\}, \tag{16}$$

où $\hat{D}_{h,ij} = \hat{M}^1_{h,ij} - \hat{M}^0_{h,ij}$. Ici, $\hat{C}^1_{h,ij} = \sum_S (\breve{y}_{h,ik} - \hat{M}^1_{h,ij}\ \breve{r}_{h,ik})(\breve{y}_{h,ik} - \hat{M}^1_{h,ji}\ \breve{r}_{h,jk}) / \hat{R}^1_{h,ij}$ est la covariance d'échantillon normalisée des évaluations pour l'ensemble des répondants qui ont répondu aux deux questions (que l'on peut prédire en utilisant les fonctions de corrélation et de variance et, dans le cas de l'échantillonnage avec probabilités inégales, en appliquant un facteur de normalisation). Lorsque les probabilités d'échantillonnage ne sont pas égales, l'équation (16) n'est vérifiée exactement que si $\sum_S \breve{r}_{h,jk}(\breve{y}_{h,ik} - \hat{M}^1_{h,ik}\ \breve{r}_{h,ik}) = 0$. Par conséquent, nous pouvons nous attendre à ce que (16) donne une bonne approximation si les probabilités d'échantillonnage pour une question ne sont pas fortement corrélées aux résidus d'une autre question. En général, il convient de vérifier s'il est approprié d'utiliser (16) pour les plans d'échantillonnage avec probabilités inégales.

Les différences estimées entre les moyennes $\hat{D}_{h,ij}$ déterminent la contribution du schéma de réponse à la covariance d'échantillonnage. Nous pouvons modéliser $\hat{D}_{h,ij}$ ou $\hat{D}_{h,ij}\,\hat{D}_{h,ji}$ dans le processus d'obtention d'estimations lissées de $\tilde{V}_{h,ij}$. Dans notre application, les $\hat{D}_{h,ij}$ sont généralement petites. Comme le deuxième terme de (16) est un produit de deux facteurs de petite taille ($\hat{D}_{h,ij}$ et $\hat{D}_{h,ji}$), la contribution de $\hat{D}_{h,ij}$ à (16) est faible et il suffit d'utiliser un modèle simple pour $\hat{D}_{h,ij}$, comme une constante pour chaque paire de questions. Cependant, une constante propre devrait être estimée pour chaque paire de questions.

3.4 Ajustement et évaluation des modèles

Nous estimons les paramètres de la fonction de variance ou de la fonction de corrélation par régression par les moindres carrés repondérée itérativement. La pondération est importante quand le nombre de réponses varie considérablement d'un domaine à l'autre, comme dans le cas de notre exemple.

À la présente section, nous utilisons un indice pour les domaines (*h*) et pour les répondants (*k*), mais non pour les questions, car la même méthodologie s'applique à chaque modèle de variance et de corrélation. Des calculs exacts sont faits pour le cas de l'échantillonnage avec probabilités

égales et des approximations sont notées pour le cas de l'échantillonnage avec probabilités inégales. De façon générique, les estimateurs directs $\tilde{f}_h$, les valeurs réelles f_h et les prédictions de modèle $\hat{f}_h$ sont reliés au moyen du modèle hiérarchique

$$\text{Niveau I}: \quad \tilde{f}_h = f_h + \in_h, \tag{17}$$

$$\text{Niveau II}: \quad f_h = \hat{f}_h + e_h, \tag{18}$$

où $\in_h \sim [0, \sigma_h^2 / \hat{R}_{S_h}]$, $e_h \sim [0, \tau^2]$, et $[\mu, \sigma^2]$ indiquent une loi d'espérance μ et de variance σ^2, mais de forme non précisée. Dans le cas de l'échantillonnage avec probabilités inégales, nous remplaçons $\hat{R}_{S_h}$ par $\hat{R}^*_{S_h}$. Ici, $\in_h$ représente l'erreur d'échantillonnage et e_h, l'erreur de modélisation. Marginalement, $\tilde{f}_h = \hat{f}_h + e_h + \in_h$, de sorte que, dans la régression, nous pondérons l'observation pour le domaine h par $w_h = (\tau^2 + \sigma_h^2 / \hat{R}_{S_h})^{-1}$, qui est l'inverse de la variance marginale. Sous échantillonnage avec probabilités égales, la variance de l'estimation directe de $\sigma_h^2 = E[\tilde{f}_h - f_h]^2$ est donnée par

$$\hat{\sigma}_h^2(\tilde{f}_h) = \frac{1}{\hat{R}_{S_h} - 1}\left\{\frac{1}{\hat{R}_{S_h}}\sum_{k \in S}(y_{h,k} - \hat{M}_h r_{h,k})^4 - (1 - \frac{3}{\hat{R}_{S_h}})\tilde{f}_h^2\right\}$$
si f est une variance (19)

et par

$$\hat{\sigma}_h^2(\tilde{f}_h) = \frac{4}{\hat{R}_{S_h} - 3} \text{ si } f \text{ est une corrélation transformée.} \tag{20}$$

Dans le cas de l'échantillonnage avec probabilités égales, l'équation (19) est exacte et ne dépend pas des hypothèses paramétriques (Seber 1977, page 14). L'approximation asymptotique (20) de la variance de la corrélation transformée Z_h (Freund et Walpole 1987, page 477) se détériore à mesure que diminuent les tailles d'échantillon et elle échoue entièrement pour $\hat{R}_{S_h} \leq 3$. Cependant, les domaines pour lesquels les échantillons sont de petite taille ont peu d'effet sur les modèles ajustés; nous excluons donc les domaines pour lesquels $\hat{R}_{S_h} \leq 3$ de la modélisation des corrélations.

Lorsque les probabilités d'échantillonnage ne sont pas égales, nous pouvons utiliser la grande contrepartie d'échantillon de (19), donnée par

$$\hat{\sigma}_h^2(\tilde{f}_h) = \sum_{k \in S}\left\{\begin{array}{l}\dfrac{(\breve{y}_{h,k} - \hat{M}_h \breve{r}_{h,k})^2}{\sum_{l \in S}\breve{r}_{h,l}^2} - \dfrac{2w_h}{\sum_{l \in S}\breve{r}_{h,l}} \\ \times(\breve{y}_{h,k} - \hat{M}_h \breve{r}_{h,k}) - \dfrac{\tilde{f}_h^2}{\sum_{l \in S}\breve{r}_{h,l}^2}\breve{r}_{h,k}^2\end{array}\right\}^2,$$

où $w_h = (\sum_S \breve{y}_{h,l}\breve{r}_{h,l}) / \sum_S \breve{r}_{h,l}^2) - \hat{M}_h$, En cas d'échantillonnage avec probabilités égales, $w_h = 0$ et l'expression susmentionnée se réduit à une version non corrigée pour le biais de (19). Si les probabilités d'échantillonnage ne sont pas égales, nous proposons de remplacer (20) par l'estimateur corrigé pour l'effet de plan.

$$\hat{\sigma}_h^2(\tilde{f}_h) = \frac{4}{\hat{R}^*_{S_h} - 3}.$$

La variance de l'erreur de modélisation τ^2 est estimée par :

$$\hat{\tau}^2 = \max\left\{0, \hat{\text{MSE}} - \frac{\sum \hat{R}_{S_h}\hat{\sigma}_h^2(\tilde{f}_h)}{\sum \hat{R}_{S_h}}\right\},$$

où $\hat{\text{MSE}} = \sum_h q_h(\tilde{V}_h - \hat{f}_h)^2$, $q_h = N\,\hat{R}_{S_h} / \sum_h \hat{R}_{S_h}$, et $N = \sum_h I(\hat{R}_{S_h} > 0)$. Nous réestimons alors les poids selon l'expression $\hat{w}_h = (\tau^2 + \hat{\sigma}_h^2(\tilde{f}_h)/\hat{R}_{S_h})^{-1}$ et nous rajustons les modèles de FVCG par itération jusqu'à la convergence. Nous suggérons de nouveau de remplacer $\hat{R}_{S_h}$ par $\hat{R}^*_{S_h}$ si les probabilités d'échantillonnage ne sont pas égales.

Nous avons comparé l'exactitude prédictive des modèles en utilisant $R^2 = 1 - \hat{\text{MSE}} / \hat{\text{MSV}}$, où $\hat{\text{MSE}}$ est l'erreur quadratique moyenne de la régression et $\hat{\text{MSV}}$ est la moyenne, pondérée par la taille de l'échantillon, des variances d'échantillonnage des estimateurs directs (variances ou corrélations transformées) pour chaque domaine. Notons que nous pourrions avoir $R^2 < 0$ pour un modèle dont l'ajustement est très médiocre.

3.5 Estimateurs combinés

Pour les domaines dont l'échantillon est petit, les estimations directes de la variance de sondage sont trop imprécises pour être utiles, tandis que, pour les domaines plus grands dans la même étude, les estimations peuvent être assez fiables. Fay et Herriot (1979), ainsi que Ghosh et Rao (1994) ont montré que réduire les estimations directes vers une valeur lissée fondée sur un modèle peut améliorer considérablement la précision. Ils proposent des estimateurs bayésiens composites ou empiriques qui sont des moyennes pondérées des estimateurs directs et d'estimateurs fondés sur un modèle. Autrement dit, au lieu d'utiliser les estimations directes ou celles obtenues par modélisation généralisée de la variance/covariance, nous utilisons une moyenne pondérée des deux estimateurs pour éventuellement obtenir d'encore meilleures estimations.

Nous pouvons construire de tels estimateurs pondérés pour les variances de domaine en utilisant le modèle spécifié en (17) et (18). Une approche naturelle consiste à appliquer aux estimateurs directs fondés sur un modèle une pondération inversement proportionnelle aux variances d'échantillonnage et d'erreur de modélisation correspondantes, respectivement (notées σ_h^2 et τ^2, respectivement, pour le

domaine *h*). L'estimateur résultant pour le domaine *h* (pour les variances et les corrélations transformées) est :

$$\tilde{f}_h = \frac{\hat{\tau}^2\,\tilde{f}_h^{\,\mathrm{dir}} + \hat{\sigma}_h^2\,\tilde{f}_h^{\,\mathrm{mod}}}{\hat{\tau}^2 + \hat{\sigma}_h^2} = \tilde{f}_h^{\,\mathrm{dir}} + \frac{\hat{\sigma}_h^2}{\hat{\tau}^2 + \hat{\sigma}_h^2}\left(\tilde{f}_h^{\,\mathrm{mod}} - \tilde{f}_h^{\,\mathrm{dir}}\right),$$

où $\tilde{f}_h^{\,\mathrm{dir}}$ et $\tilde{f}_h^{\,\mathrm{mod}}$ représentent les estimateurs directs et fondé sur un modèle. Cette formule générique s'applique aux estimations de la variance pour l'ensemble des questions et aux estimations des corrélations pour toutes les paires de questions. L'expression la plus à droite a la forme d'un estimateur bayésien empirique.

Si l'estimateur de variance direct et celui fondé sur un modèle sont indépendants, la variance de l'estimateur combiné résultant est $\tau^2\sigma_h^2/(\tau^2+\sigma_h^2) \le \min\{\tau^2, \sigma_h^2\}$. Donc, l'estimateur composite est, au moins, aussi précis que l'un ou l'autre des deux estimateurs qui le constituent, donc représente une amélioration par rapport au choix ponctuel entre la prédiction directe et celle fondée sur un modèle. Cette stratégie est utile, surtout quand les prédictions fondées sur un modèle sont meilleures que les estimations directes pour certains domaines, mais non tous.

4. Exemple : Ensemble de données de la CAHPS®

La Consumer Assessments of Health Plans Study (CAHPS®) (Goldstein, Cleary, Langwell, Zaslavsky et Heller, 2001) a été conçue principalement pour recueillir les évaluations et les déclarations des consommateurs au sujet des régimes d'assurance-maladie. Les scores moyens des régimes (peut-être après recodage) pour les diverses questions du sondage sont calculés et communiqués aux consommateurs, aux régimes d'assurance-maladie et aux acheteurs. Chaque domaine d'analyse comprend les personnes inscrites à un régime d'assurance-maladie (ou une partie géographiquement définie d'un régime) durant une année particulière; la plupart des régimes sont échantillonnés pour plusieurs années. La strate est l'unité déclarante (régime ou partie de celui-ci) durant une année donnée; les unités déclarantes correspondent à des régimes sauf dans le cas de quelques régimes très importants comptant plusieurs unités déclarantes. Par conséquent, le nombre d'unités pour l'estimation des fonctions de variance et de covariance est grand.

Nous illustrons notre méthode au moyen d'un ensemble de données de la CAHPS pour les bénéficiaires des régimes de soins américains gérés par Medicare, c'est-à-dire un système d'entités privées, mais financées par les deniers publics qui ont desservi de 5,7 à 6,9 millions de bénéficiaires âgés ou handicapés chaque année de la période couverte par l'étude (1997 à 2001). Nos données représentent 381 domaines déclarants, chacun échantillonné dans une à cinq années, pour aboutir à un total de 932 domaines distincts correspondant à une unité déclarante par année, avec 705 848 réponses. Comme les échantillons sont tirés indépendamment chaque année, les patients peuvent être échantillonnés pour plusieurs années. Cependant, l'échantillonnage répété est rare et nous pouvons omettre d'en tenir compte dans notre analyse. Par conséquent, les domaines sont des strates dans chacune desquelles est effectué un échantillonnage d'éléments avec probabilités égales. Il convient de souligner que, dans les analyses de la CAHPS, aucune correction n'est faite pour l'échantillonnage en population finie, puisque les données sont recueillies pour orienter les choix faits lors d'années ultérieures plutôt que pour enregistrer les expériences vécues par la population sondée durant une année particulière.

Les questions de la CAHPS comportent divers formats de réponse ordonnée comportant 11, 4, 3 ou 2 options de réponse. L'évaluation globale des médecins, des spécialistes, des soins et des régimes est faite au moyen d'une échelle de 0 à 10 allant du « pire possible » au « meilleur possible ». D'autres questions comportent une échelle de « fréquence » ordonnée de quatre points (jamais/parfois/habituellement/toujours), ou une échelle d'« intensité du problème » ordonnée de trois points (pas un problème/plus ou moins un problème/un gros problème), ou sont dichotomiques (non/oui). La réponse à de nombreuses questions est fournie uniquement par les répondants utilisant des services particuliers ou ayant des besoins particuliers, qui sont dépistés au moyen des questions filtres. Par exemple, une question visant à savoir si la personne a réussi à obtenir des conseils par téléphone n'est posée qu'aux répondants qui ont d'abord déclaré qu'ils avaient essayé d'obtenir des conseils de cette façon.

4.1 Statistiques descriptives

Le tableau 1 présente la répartition des réponses et celle des moyennes de domaine selon le type de question. Les observations manquantes à cause d'un enchaînement de questions structuré surviennent souvent en blocs, le nombre de questions sautées pouvant aller jusqu'à 11 sur la base d'une seule question filtre. La proportion de non-réponses n'étant pas due à un enchaînement de questions structuré est très faible (moins de 2 % pour presque toutes les questions). Dans la présente analyse, nous traitons tous les types de non-réponse de la même façon.

Les taux les plus faibles de réponse à une question (aussi faibles que 4 %) sont ceux enregistrés pour les questions à échelle d'intensité de problème, dont plusieurs ont trait à des services spécialisés, comme des traitements ou des soins de santé à domicile dont ont besoin un assez petit nombre de répondants seulement. Certaines questions à échelle de fréquence et à réponse oui/non produisent des taux de

réponse assez faibles. La variation la plus forte des proportions de questions sautées s'observe pour les questions à réponse oui/non, le taux allant de 96,7 % pour « une plainte ou un problème concernant le régime » à 12,5 % pour « l'obtention d'une prescription par l'entremise du régime ». Les moyennes de domaine sont généralement concentrées vers l'extrémité supérieure des échelles respectives, indiquant que la plupart des réponses sont favorables.

Tableau 1
Distribution des réponses et des évaluations pour des questions de même type (n = 705 848 répondants)

Statistique	Numérique	Fréquence	Problème	Oui/non
Nombre de questions	4	11	11	9
Pourcentage de réponses				
Moyenne	74,97	62,56	30,32	57,26
Minimum	50,90	27,70	4,00	12,50
Maximum	95,00	74,50	64,40	96,70
Moyenne de questions				
Moyenne	8,76	3,57	2,70	1,78
Minimum	8,57	3,09	2,49	1,62
Maximum	8,88	3,84	2,86	1,97
Distribution des évaluations (entre les questions de même type)				
0	0,5			
1	0,4	2,0	5,7	19,5
2	0,4	6,3	12,1	80,5
3	0,7	23,9	82,2	
4	0,9	67,8		
5	4,6			
6	3,0			
7	6,2			
8	16,1			
9	17,8			
10	49,5			

Les questions comportent une échelle numérique 0–10 allant de « pire possible » à « meilleur possible », une échelle de « fréquence » ordonnée de quatre points 1–4 (jamais/parfois/habituellement/ toujours), une échelle d'« intensité de problème » ordonnée de trois points 1–3 (pas un problème/plus ou moins un problème/un gros problème) ou sont dichotomiques 1–2 (non/oui).

Nous présentons aussi dans le tableau 1 la moyenne, le minimum et le maximum de domaine pour l'ensemble des questions de même type. Ces données montrent que les questions à échelle 0–10 sont celles pour lesquelles la variation totale est la plus faible (après rééchelonnement pour produire la fourchette 0–1 commune), tandis que les questions à échelle 1–2 sont celles dont la variation totale est la plus importante entre les domaines et entre les questions. Ces observations se dégagent aussi de la figure 1, où nous observons que la distribution des questions à échelle 1–2 varie considérablement d'une question à l'autre, tandis que celle des questions à échelle 0–10 est plus homogène.

Le tableau 2 donne les statistiques sommaires pour les moyennes et les écarts-types des évaluations moyennes de domaine, calculées sur l'ensemble des questions de même type. Ces données complètent celles de la figure 1 en résumant les différences entre les distributions des questions pour une échelle donnée. Les questions comportant un plus grand nombre de catégories de réponse sont concentrées vers l'extrémité supérieure de l'échelle et ont donc une variance plus faible. Par exemple, l'écart-type moyen pour les questions à échelle 1–2 (0,36) est égal au double de celui des questions à échelle 0–10 rééchelonnées (0,172). À part les questions à échelle 0–10, les distributions des évaluations moyennes de domaine varient fortement entre les questions de même type. Par exemple, l'écart-type des moyennes des questions à échelle 1–2 sur l'ensemble des questions est de 0,30 comparativement à un écart-type rééchelonné de 0,03 pour les questions à échelle 0–10.

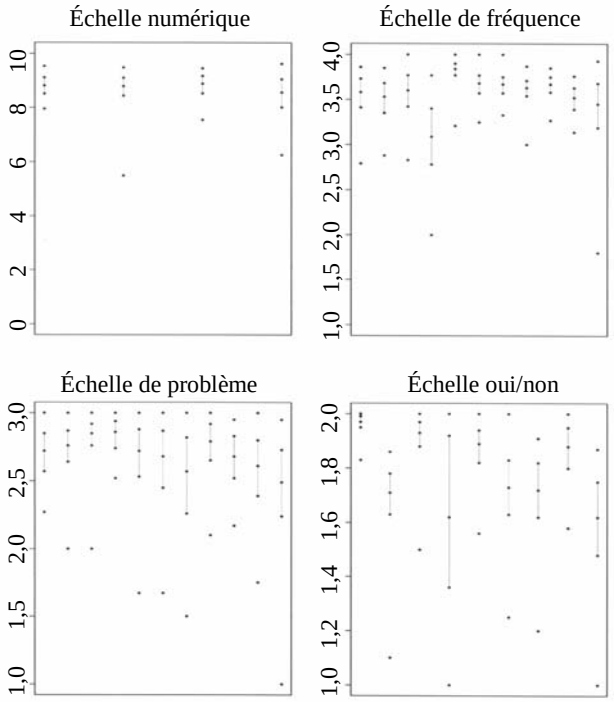

Figure 1. Sommaire de cinq points des moyennes d'échantillon de domaine pour chaque type de question. Le sommaire de cinq points comprend le minimum, le 10^{e} centile, la moyenne, le 90^{e} centile et le maximum.

Tableau 2
Statistiques sommaires des moyennes et des écarts-types de domaine évalués sur les domaines et sur les questions

Type	Statistiques sommaires pour : Moyennes des questions				É.-T. des questions	
	Min	Max	Moy.	É.-T.	Moy.	É.-T.
Numérique 0 –10	6,82	9,52	8,76	0,30	1,72	0,26
Fréquence 1– 4	2,86	3,90	3,57	0,12	0,66	0,09
Problème 1–3	1,88	2,99	2,70	0,14	0,57	0,13
Oui/non 1–2	1,34	1,96	1,78	0,08	0,36	0,06

Nota : Les colonnes 2 à 5 donnent le minimum, le maximum, la moyenne et l'écart-type des moyennes des questions de domaine sur l'ensemble des questions d'un type donné. Les colonnes 6 et 7 donnent la moyenne et l'écart-type des écarts-types des questions de domaine sur l'ensemble des questions d'un type donné.

Les corrélations d'échantillon varient également beaucoup d'une paire de questions à l'autre (figure 2), quoique la plupart soient positives. Le plus souvent, les corrélations entre questions de même type sont plus fortes que celles entre questions de types différents. Les évaluations numériques à échelle 0–10 sont celles dont les corrélations sont les plus importantes (moyenne = 0,49) et, en général, les évaluations comportant un grand nombre de catégories ont tendance à produire des corrélations plus fortes que celles comportant un moins grand nombre de catégories. Bien que la plupart des paires de questions à échelle 1–4 donnent des corrélations moyennes s'approchant de 0,5, l'une des questions est négativement corrélée aux autres (révélé par la grappe de corrélations moyennes inférieures à 0); ce résultat est dû au codage inverse d'une question dont la moyenne globale d'échantillon ne se situait pas dans la moitié supérieure de l'échelle. Les distributions des corrélations des paires de questions dichotomiques 1–2 sont centrées autour de 0, ce qui signifie que ces corrélations sont souvent négatives. L'énoncé complet des questions et des statistiques sommaires supplémentaires figure dans Zaslavsky, Beaulieu, Landon et Cleary (2000) et dans Zaslavsky et Cleary (2002).

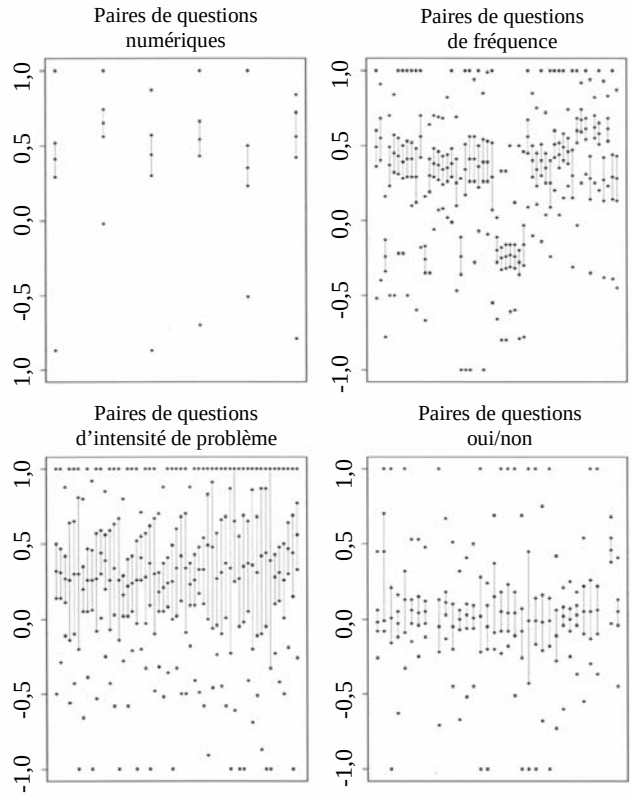

Figure 2. Sommaire de cinq points des corrélations d'échantillon de domaine entre les questions de même type. Le sommaire de cinq points comprend le minimum, le 10e centile, la moyenne, 90e centile et le maximum.

Nous présentons les modèles ajustés aux variances et aux corrélations dans la suite de la section. Des vérifications approfondies des modèles les mieux ajustés ont montré que les résidus ne présentaient aucune régularité discernable.

4.2 Fonctions de variance

Lors de travaux préliminaires non présentés ici, nous avons ajusté deux modèles pour les groupes de questions ayant la même échelle de réponse, l'un avec les mêmes paramètres de régression pour toutes les questions et l'autre avec des paramètres de régression différents pour chaque question, à l'ensemble de données comprenant toutes les questions. Les comparaisons des ajustements globaux des modèles (au moyen de critères tels que le C_p de Mallow, le R^2 et le R^2 corrigé) et les tests de signification des interactions effet-question ont montré que permettre aux paramètres de varier selon la question améliore de façon significative l'ajustement du modèle. Par exemple, pour les évaluations numériques rééchelonnées, pondérées par la taille d'échantillon du domaine, les racines des erreurs quadratiques moyennes des deux modèles étaient de 0,446 contre 0,402, et les valeurs de R^2 étaient de 0,783 contre 0,825. D'après ces résultats, nous avons décidé d'ajuster des modèles distincts pour chaque question.

Nous avons ajusté les fonctions de variance (8) à (10) à chaque question, sauf celles à réponse oui/non, qui suivent la fonction de variance binomiale dans le cas de l'échantillonnage avec probabilités égales. La procédure itérative décrite à la section 3.4 a convergé presque précisément en exactement deux itérations. Ce résultat tient au fait que les poids des observations ne varient qu'en fonction de l'estimation de τ^2, de sorte que leur variation est très faible après la première itération.

Le tableau 3 donne la variation d'échantillonnage moyenne, la variation moyenne de l'erreur de modélisation (ModErr) et R^2 pour la valeur moyenne de chaque modèle calculés sur l'ensemble des questions correspondant à chaque type d'échelle. La variation d'échantillonnage, calculée selon (19), ne dépend pas du modèle.

Pour les questions ne comportant qu'un petit nombre de catégories (celles qui ressemblent le plus à la loi binomiale), la composante quadratique de la fonction de variance a tendance à dominer la composante linéaire, ce qui produit un meilleur ajustement des modèles V2 et V3 que du modèle V1. Comme V2 impose une contrainte en un point très en dehors de la fourchette de valeurs des moyennes de domaine, il n'est pas aussi bien ajusté aux données quand le nombre de catégories augmente et que les données s'écartent par conséquent davantage de la loi binomiale. Les réponses aux questions à échelle 0–10 sont moins dispersées que celles aux questions à échelle 1–4 ou 1–3, de sorte que le modèle linéaire est mieux ajusté. Les valeurs de

R^2 pour le modèle V3 s'approchent de 0,75 pour les questions à échelle numérique (0–10), de 0,85 pour les questions à échelle de fréquence (1–4) et de 0,95 pour les questions à échelle d'intensité de problème (1–3).

Tableau 3
Statistiques de qualité d'ajustement du modèle pour les fonctions de variance

Échelle d'évaluation	0–10		1–4		1–3	
Variation d'échantillonnage	0,1460		0,3511		3,1703	
	ModErr	R^2	ModErr	R^2	ModErr	R^2
Modèle V1	0,020	0,741	0,066	0,824	0,069	0,916
Modèle V2	0,043	0,710	0,036	0,835	0,000	0,940
Modèle V3	0,016	0,750	0,024	0,847	0,000	0,947
	Prob(ModErr < Variation d'échantillonnage)					
Modèle V1	0,968		0,916		0,996	
Modèle V2	0,858		0,967		0,996	
Modèle V3	0,981		0,983		0,996	

ModErr est la composante de la variance due au manque d'ajustement, R^2 est la statistique définie à la section 3.4, Prob(ModErr < Variation d'échantillonnage) est la proportion de domaines pour lesquels l'erreur de modélisation est plus faible que la variation d'échantillonnage. Toutes les évaluations sont rééchelonnées sur une échelle de 0–1 et les erreurs de modélisation sont multipliées par 10^4.

La partie inférieure du tableau 3 donne, pour chaque question, la proportion de domaines (parmi ceux comptant au moins deux réponses à la question considérée) pour lesquels la variation d'échantillonnage est supérieure à la variation de l'erreur de modélisation. Pour plus de 90 % des domaines, la variation de l'erreur de modélisation est plus faible que la variation d'échantillonnage de l'estimation directe de la variance.

La figure 3 illustre l'ajustement de V3 pour deux questions de chacun des groupes 0–10, 1–4 et 1–3. Les illustrations pour les autres questions sont semblables, mais ne sont pas présentées ici faute d'espace. Les courbes ajustées sont contraintes de passer par la valeur 0 à l'évaluation maximale. Pour évaluer l'effet de cette contrainte sur la fonction de variance ajustée, nous ajustons également une fonction de variance quadratique (à trois paramètres) non contrainte. Celle-ci atteint une valeur très proche de 0 à l'évaluation maximale et s'approche de très près de la courbe ajustée d'après les modèles avec contraintes, ce qui appuie encore davantage le modèle V3.

Les estimations moyennes des paramètres et de leurs écarts-types sur l'ensemble des questions de même type sont présentées au tableau 4. La valeur des paramètres varie considérablement selon la question, ce qui soutient la décision d'estimer des coefficients de régression distincts.

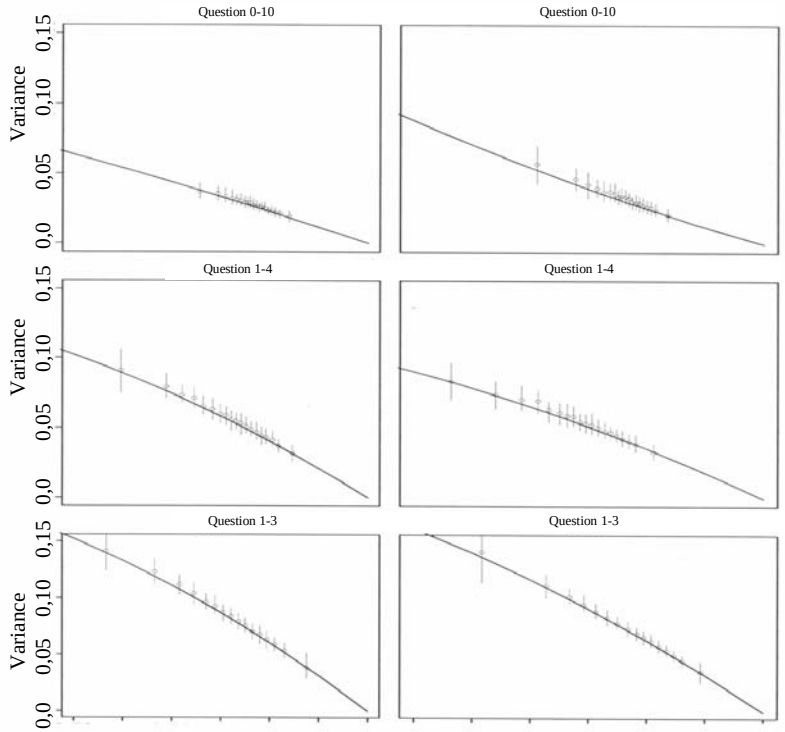

Figure 3. Fonction de variance quadratique (V3) de deux questions pour chaque type d'évaluation. Chaque point est la moyenne de 60 domaines. Les lignes verticales joignent les 10e et 90e centiles de la distribution des variances. Pour ces tracés et les suivants, la direction de l'axe horizontal a été inversée afin qu'elle concorde avec celle des variables originales.

Dans la plupart des cas, les coefficients pour les termes $p_{h,i}$ ainsi que $p_{h,i}(1-p_{h,i})$ de V3 sont significatifs, ce qui indique que ces termes sont nécessaires pour la modélisation généralisée de la variance. Dans certains cas (particulièrement pour les questions à échelle 0–10), le coefficient du terme $p_{h,i}(1-p_{h,i})$ est négatif, ce qui donne une fonction de variance estimée convexe plutôt que concave (forme de la fonction de variance binomiale). Cette situation peut se produire si les moyennes d'échantillon des évaluations sont concentrées sur une petite partie de l'échelle de réponse, sur laquelle le terme linéaire explique une grande part de la variation des données. Comme nous l'avons mentionné plus haut, l'ajout de fonctions polynomiales ou logarithmiques d'ordre plus élevé de $p_{h,i}$ n'améliore pas significativement l'ajustement du modèle.

Tableau 4
Estimations moyennes des paramètres de la fonction de variance pour chaque type de question et écart-type entre les questions (entre parenthèses)

Modèle	Type de question					
	0–10		1–4		1–3	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
V1	0,236	–	0,354	–	0,569	–
	(0,016)	–	(0,039)	–	(0,068)	–
V2	–	0,271	–	0,421	–	0,711
	–	(0,020)	–	(0,034)	–	(0,069)
V3	0,334	– 0,114	0,151	0,241	0,239	0,420
	(0,143)	(0,155)	(0,104)	(0,132)	(0,112)	(0,110)

Voir le tableau 1 pour une description des questions de type 0–10, 1–4 et 1–3.

4.3 Fonctions de corrélation

Les modèles sont classés du plus simple (C1, modèle constant) au plus complexe (C5, contenant tous les termes linéaires et quadratiques). Comme pour les modèles de variance, les tests statistiques indiquent des effets d'interaction entre questions hautement significatifs, qui sous-entendent que des modèles distincts devraient être ajustés pour chaque paire. Nous ne nous attendions pas à ce que toutes les paires de questions présentent les mêmes corrélations, puisque, intentionnellement, les questions sont réparties en groupes intérieurement cohérents, qui mesurent chacun un aspect distinct des expériences vécues par les patients, comme les interactions avec le médecin ou celles avec les agents des services à la clientèle (Hays, Shaul, Williams, Lubalin, Harris-Kojetin, Sweeny et Cleary 1999).

Les ajustements des modèles de corrélation pour les paires de questions de même type sont résumés au tableau 5. Pour la gamme de modèles considérée, les améliorations les plus importantes de l'efficacité des modèles (mesurées par R^2) sont celles observées entre C1 et C2, et entre C3 et C4. Par exemple, le R^2 moyen pour les évaluations numériques à échelle 1–10 dans les modèles C3 à C5 est de 0,0391, 0,1494 et 0,1508, respectivement, et le R^2 moyen pour les évaluations à échelle 1–4 pour les modèles C1 à C3 est de 0, 0,0700 et 0,0789, respectivement. Ces résultats donnent à penser que C2 et C4 sont les meilleurs modèles pour différentes paires de questions, affirmation qui est appuyée par les tests de vérification d'hypothèse concernant la signification des améliorations marginales de l'ajustement des modèles.

La variation d'échantillonnage la plus élevée s'observe pour les échelles d'évaluation 1–3, du moins en partie parce que les taux élevés de non-réponses dues à des enchaînements de questions ont réduit les tailles d'échantillon. L'erreur de modélisation et le R^2 des modèles de corrélation pour les questions de types différents sont semblables à ceux des modèles pour les questions de même type.

Les valeurs de R^2 des modèles de corrélation sont comprises entre 0,029 et 0,15 pour toutes les paires de questions. Bien qu'il n'existe aucune preuve que C4 soit un modèle inapproprié pour les corrélations, ces résultats indiquent qu'une variation importante des corrélations ne peut être expliquée par les moyennes de question.

Les variances d'échantillonnage des estimations directes sont souvent inférieures aux variances de l'erreur de modélisation correspondante (partie inférieure des tableaux 5 et 6, particulièrement pour les questions à échelle 0–10. Sous C4, les variances de l'erreur de modélisation ne sont plus faibles que pour 13 % des domaines pour les évaluations de type 0–10, 40 % des domaines pour les évaluations de type 1–4 et environ 81 % des domaines pour les évaluations de type 1–3 ou 1–2.

La figure 4 donne les corrélations observées et la fonction ajustée C4 pour un exemple de paire de questions pour chacune des dix combinaisons de type de question, représentant les 595 paires de questions distinctes. Pour illustrer les modèles de corrélation ajustés, nous rajustons les corrélations observées et ajustées sur la moyenne de l'une des questions et représentons graphiquement les valeurs résultantes dans un espace bidimensionnel. Nous répétons le processus pour l'autre question, ce qui nous donne deux tracés pour chaque corrélation.

La figure 4 illustre la relation généralement faible entre les corrélations et les moyennes des questions observées aux tableaux 5 et 6. L'analyse de ces deux tableaux révèle que la relation entre la corrélation et le résultat moyen est plus faible pour les questions comportant un petit nombre de catégories et pour les corrélations de questions de différents types. En particulier, les évaluations numériques à échelle 0–10 sont le seul groupe pour lequel il existe une relation claire corrélation-moyenne.

Bien que les courbes ajustées pour les fonctions de corrélation soient presque plates, la variation des estimations des paramètres sous le modèle C4 pour α_4 sont grandes et évoquent une instabilité. La très forte variabilité des estimations des paramètres est une conséquence de la

colinéarité entre les prédicteurs dans le modèle C4. Dans de nombreux cas, la valeur estimée de α_4 contrebalance les estimations des paramètres des prédicteurs linéaires, ce qui donne une courbe ajustée pratiquement plate.

4.4 Fonctions de différence de moyennes

La différence $\hat{D}_{h,ij}$ semble ne dépendre ni de la moyenne marginale ni de son carré, ce qui implique qu'un modèle analogue à V3 pourrait être approprié. Cependant, comme $\hat{D}_{h,ij}$ est habituellement suffisamment faible pour que $\hat{D}_{h,ij}\ \hat{D}_{h,ji}$ ait un effet minime sur (16), nous ajustons un modèle constant.

4.5 Estimateur composite

Le tableau 7 donne les valeurs moyennes, calculées sur l'ensemble des questions (ou paires de questions) de même type, des quantiles de la distribution des poids $\sigma_h^2 / (\tau^2 + \sigma_h^2)$ pour l'estimation fondée sur un modèle utilisée dans l'estimateur composite de la section 3.5. La proportion de domaines pour lesquels l'erreur-type des prédictions fondées sur un modèle est plus faible que celle des estimations directes est également présentée. Comme nous l'avons mentionné plus haut, les prédictions fondées sur un modèle ont plus de poids dans les estimations composites de la variance que dans celles des corrélations. La médiane moyenne (sur les questions ou les paires de questions) des poids de l'estimateur fondé sur un modèle varie de 0,892 à 1,000 pour les variances, de 0,256 à 0,709 pour les corrélations de questions de même type, et de 0,468 à 0,738 pour les corrélations de questions de types différents. En outre, tant pour les variances que pour les corrélations, le poids des prédictions fondées sur un modèle est plus important pour les questions comportant un petit nombre de catégories de réponses. Par exemple, les poids médians de l'estimateur fondé sur un modèle sont de 0,256, 0,468, 0,540 et 0,647 sur les estimations composites des corrélations quand les évaluations numériques de type 0–10 sont appariées aux évaluations de type 0–10, 1–4, 1–3 et 1–2, respectivement. Cependant, même pour les paires d'évaluation numérique à échelle 0–10, pour lesquelles l'erreur d'échantillonnage de l'estimateur direct n'est supérieur à l'erreur de modélisation que pour 3,81 % de domaines, ces résultats indiquent que le poids médian de l'estimateur fondé sur un modèle est de 0,256, valeur non triviale.

Tableau 5
Diagnostics d'ajustement de modèle pour les fonctions de corrélation pour les questions de même type, moyenne sur les paires de questions de même type

Échelle d'évaluation	0 – 10		1–4		1–3		1–2	
Variation d'échantillonnage	0,0124		0,0178		0,1482		0,0325	
	ModErr	R^2	ModErr	R^2	ModErr	R^2	ModErr	R^2
Modèle C1	0,060	0,000	0,028	0,000	0,112	0,000	0,018	0,000
Modèle C2	0,060	0,013	0,025	0,070	0,103	0,048	0,017	0,014
Modèle C3	0,057	0,039	0,024	0,079	0,102	0,054	0,017	0,018
Modèle C4	0,047	0,150	0,023	0,100	0,100	0,068	0,016	0,029
Modèle C5	0,044	0,151	0,023	0,105	0,096	0,080	0,015	0,034
	Prob(ModErr < Variation d'échantillonnage)							
Modèle C1	0,033		0,339		0,461		0,788	
Modèle C2	0,033		0,400		0,498		0,795	
Modèle C3	0,034		0,411		0,502		0,796	
Modèle C4	0,038		0,435		0,516		0,799	
Modèle C5	0,065		0,440		0,530		0,802	

Voir le tableau 1 pour une description des questions de type 0–10, 1–4, 1–3 et 1–2, et le tableau 3 pour une explication des en-têtes de colonne.

Tableau 6
Diagnostics d'ajustement de modèle pour les fonctions de corrélation pour C4 selon le type de question. Moyenne sur les questions de même type.

Type	0–10		1–4		1–3		1–2	
	ModErr	R^2	ModErr	R^2	ModErr	R^2	ModErr	R^2
0–10	0,047	0,149	0,021	0,104	0,040	0,094	0,013	0,059
1–4			0,023	0,100	0,038	0,076	0,013	0,039
1–3					0,100	0,068	0,028	0,031
1–2							0,016	0,029
Prob(ModErr < Variation d'échantillonnage)								
0–10	0,038		0,358		0,523		0,784	
1–4			0,435		0,605		0,790	
1–3					0,516		0,827	
1–2							0,799	

Voir le tableau 1 pour une description des questions de type 0–10, 1–4, 1–3 et 1–2, et le tableau 3 pour une explication des en-têtes de colonne.

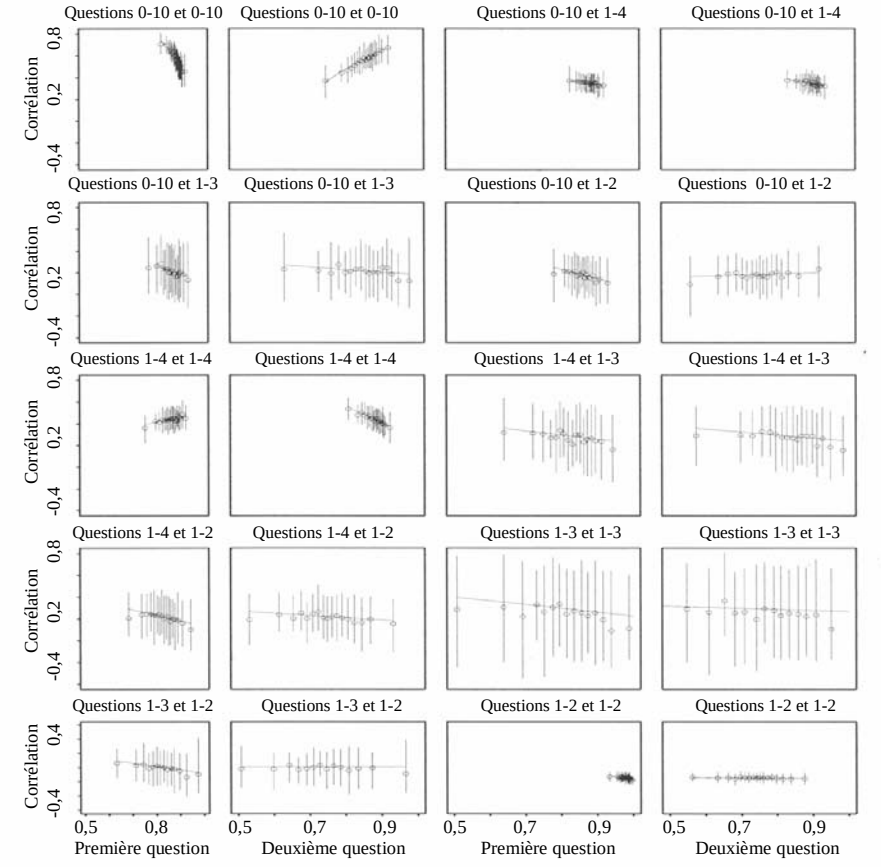

Figure 4. Fonctions de corrélation pour une paire de questions pour chaque combinaison d'échelle d'évaluation.

Nota : Les tracés pour chaque question intervenant dans la corrélation sont côte à côte. Consulter la figure 3 pour une description du contenu et des axes du tracé.

Tableau 7
Distribution des poids pour la composante fondée sur un modèle de l'estimateur composite, moyenne sur les questions de même type

Modèle	Type de question		Prob(ModErr <	Quantiles		
	1	2	Variation d'échantillonnage)	10 %	Médiane	90 %
Variance	0 – 10	–	0,981	0,778	0,892	0,948
	1 – 4	–	0,983	0,948	0,966	0,974
	1 – 3	–	0,996	1,000	1,000	1,000
Corrélation	0 – 10	0 – 10	0,038	0,141	0,256	0,335
	0 – 10	1 – 4	0,358	0,301	0,468	0,562
	0 – 10	1 – 3	0,523	0,357	0,540	0,654
	0 – 10	1 – 2	0,784	0,531	0,695	0,767
	1 – 4	1 – 4	0,435	0,324	0,497	0,591
	1 – 4	1 – 3	0,605	0,404	0,587	0,699
	1 – 4	1 – 2	0,853	0,584	0,738	0,805
	1 – 3	1 – 3	0,516	0,349	0,540	0,675
	1 – 3	1 – 2	0,827	0,584	0,737	0,817
	1 – 2	1 – 2	0,799	0,541	0,709	0,780

La distribution des poids est résumée au moyen des 10^{e}, 50^{e} et 90^{e} centiles. Voir le tableau 3 pour la définition de ModErr.

4.6 Prédiction conjointe

Comme nous avons modélisé indépendamment les corrélations pour chaque question, nos matrices de corrélations ajustées ne satisfont pas nécessairement la contrainte de définie positive, qui peut être importante pour les inférences multivariées. Dans le cadre de travaux supplémentaires, nous avons déterminé qu'à condition de limiter l'analyse multivariée aux questions de même type, les corrélations ajustées d'après les modèles C2 et C4 donnent des estimations définies positives des matrices de corrélation pour presque tous les domaines. Cependant, pour les analyses portant sur des questions de types différents (par exemple, les questions à échelle numérique 0–10 et les questions à réponse oui/non 1–2), les prédictions fondées sur C4 donnent des matrices de corrélations qui sont indéfinies pour de nombreux domaines, tandis que celles fondées sur C2 sont plus stables et donnent presque systématiquement des matrices définies positives. Ceci donne à penser que, si C4 peut être légèrement supérieur en ce qui concerne l'ajustement du modèle univarié, C2 pourrait être plus approprié pour l'inférence multivariée.

Un moyen de contourner le problème des matrices de corrélations prédites indéfinies consiste à utiliser une moyenne pondérée de la matrice de corrélations prédite pour un domaine et de la matrice de corrélations moyenne estimée (MCME) sur l'ensemble de domaines. Nous pouvons construire la MCME par pondération des estimations directes (chacune étant au moins semi-définie positive) par la taille totale de l'échantillon pour chaque domaine. Puis, nous remplaçons toute matrice de corrélations prédite indéfinie par la moyenne pondérée de la matrice de corrélations prédite et de la MCME, en accroissant le poids utilisé pour chaque domaine jusqu'à ce que nous obtenions une matrice définie positive. Comme pour un estimateur bayésien empirique, ce processus stabilise les estimations en réduisant effectivement les coefficients du modèle vers ceux d'un modèle plus simple (constant).

Lors de l'analyse simultanée des 35 questions de la CAHPS, la MCME avait un poids moyen sur l'ensemble des domaines de 0,65 pour le modèle C4, mais de 0,01 seulement pour le modèle C2, puisque les corrélations prédites sous C2 sont habituellement définies positives. Lors de l'analyse des questions de type 0–10, 1–4 ou 1–3 seulement, la MCME avait un poids moyen de 0,28 et de 0,00 pour C4 et C2, respectivement, tandis que lors de l'analyse des questions de type 0–10 ou 1–4 seulement, les poids moyens correspondants étaient de 0,06 et de 0,00. Lors de l'analyse de divers types de questions séparément, le poids moyen de la MCME avec le modèle C4 était de 0,00 pour les questions de type 0–10 ou 1–4, de 0,01 pour les questions de type 1–3 et 0,17 pour les questions de type 1-2. La MCME n'est donc pas nécessaire pour analyser les questions de type 0–10 ou 1–4, parce que les matrices de corrélations prédites sont définies positives pour chaque domaine.

5. Conclusion

Nous présentons la méthodologie pour estimer les fonctions de variance et de covariance pour les moyennes de domaine de questions d'enquête ordonnées. Notre méthodologie peut également s'appliquer aux questions d'enquête comportant une échelle de mesure continue. Nous présentons une décomposition de l'erreur de modélisation qui permet de séparer la variation due à l'échantillonnage de celle due à l'ajustement du modèle. La décomposition permet aussi d'éviter le surajustement du modèle, parce qu'elle estime la proportion de la variation des données qui peut être modélisée et, donc, le moment où les prédicteurs courants suffisent.

La procédure d'ajustement des modèles de variance et de corrélation est la même que les données contiennent ou non des enchaînements structurés de questions. L'exposé analytique de la section 3.3 montre que, s'il existe des enchaînements de questions, il faut connaître les différences moyennes entre les questions selon la situation de réponse à d'autres questions afin d'estimer la covariance d'échantillonnage. Cependant, nous soutenons que ces quantités ont vraisemblablement un effet minime sur les résultats et que, par conséquent, on pourrait utiliser un modèle constant, argument qui est appuyé par nos résultats empiriques.

Une fonction de variance quadratique dont la valeur est contrainte d'être nulle pour l'évaluation maximale et un modèle pour les corrélations transformées comportant le produit, mais non le carré des moyennes sont les meilleurs prédicteurs des estimations directes dans l'application que nous avons choisie comme exemple. En général, les erreurs-types des estimations modélisées de la variance sont beaucoup plus faibles que celles des estimations directes; toutefois, il n'en est pas ainsi des estimations des corrélations. Il est intéressant et rassurant de constater que notre fonction de variance quadratique peut être exprimée sous la forme du modèle de variance relative très répandu de Wolter (1985).

Pour nos données ordonnées, les évaluations moyennes de domaine contiennent des informations minimales concernant la corrélation entre les évaluations. Donc, la relation moyenne-covariance est principalement un artefact de la relation moyenne-variance. Cependant, pour les questions comportant un grand nombre de catégories de réponse, l'association entre les corrélations et les résultats moyens pour les questions de même type est plus forte, surtout pour les paires de questions à échelle 0–10. À part les évaluations de type 0–10 et, éventuellement, celles de

type 1–4, les corrélations pourraient tout aussi bien être modélisées par des constantes, ce qui permet aussi de garantir plus facilement le caractère défini positif de la matrice de corrélations prédite. Cependant, il est important que les paramètres du modèle de corrélation puissent varier selon la paire de questions.

Un estimateur composite, résultant de la combinaison pondérée des estimateurs direct et fondé sur un modèle proportionnellement à leur précision, a une variance plus faible que l'un et l'autre estimateur pris individuellement, surtout quand les composantes ont des poids presque égaux. L'estimateur fondé sur un modèle est celui dont l'influence sur les estimations pour les petits domaines pour lesquels ont dispose de peu d'information est la plus forte. L'influence de l'estimateur fondé sur un modèle est, par ordre décroissant d'importance, la plus forte sur les estimations de la variance, les corrélations des questions de même type et, enfin, les corrélations des questions de types différents. Les estimateurs fondés sur un modèle et les estimateurs composites peuvent les uns et les autres êtres calés (ajustement par le quotient) de sorte que les moyennes sur l'ensemble des domaines concordent avec les estimations directes, bien que cela n'ait pas été nécessaire dans notre exemple.

Les fonctions de variance et de covariance généralisées (FVCG) ont plusieurs applications dans nos travaux de recherche en cours. Nous élaborons des méthodes fondées sur la quasi-vraisemblance d'estimation des matrices de covariance pour les moyennes de domaines des questions ordonnées d'enquête, avec représentation de la covariance de deuxième niveau (structurelle) au moyen d'un modèle hiérarchique (O'Malley et Zaslavsky 2004). Les modèles de FVCG sont nécessaires pour obtenir des estimations des variances et des covariances d'échantillonnage, ainsi que pour modifier ces estimations lorsque les moyennes sont réestimées durant la procédure d'ajustement de modèle. Si la variabilité d'échantillonnage des estimations des FVCG est minimale parce que le nombre de domaines est grand, les variances et les covariances prévues des FVCG peuvent être considérées comme étant connues. Cependant, si l'erreur d'échantillonnage des estimations fondées sur les FVCG est importante, il convient d'utiliser un modèle permettant à ces erreurs de se propager tout au long de l'analyse. Dans le cadre de travaux connexes, Fay et Train (1997) ont utilisé un modèle binomial avec un effet de plan pour chaque domaine dans l'estimation bayésienne empirique des taux binomiaux. Notre étude étend cette approche à l'estimation multivariée et à des formats de réponse plus généraux.

Une autre application des FVCG est le calcul des estimations de la variance pour les combinaisons linéaires de moyennes de question afin de faciliter l'estimation de la variance de scores composites, comme ceux utilisés dans la publication des résultats de la CAHPS. Les méthodes décrites à la section 2 sont applicables à l'estimation de la variance de toute fonction de totaux, y compris les fonctions de moyennes, d'autres ratios ou des coefficients de régression.

La méthodologie des FVCG peut être étendue de plusieurs façons. Outre les mesures sommaires des résultats, les fonctions de variance et covariance généralisées (FVCG) peuvent dépendre d'autres variables indépendantes, en particulier celles qui prédiraient mieux les corrélations. Nous avons considéré des variables résumant les profils de réponse, comme la proportion de répondants dans un domaine, mais celles-ci n'ont pas amélioré le modèle. Les FVCG pourraient aussi être étendues à l'échantillonnage à plusieurs degrés.

Remerciements

Les présents travaux ont été financés par la U.S. Agency for Healthcare Research and Quality par la voie de la Consumer Assessments of Health Plans Study (subvention U18 HS09205-06) et par les U.S. Centers for Medicare and Medicaid Services (contrat 500-95-007). Nous remercions Paul D. Cleary de son appui continu durant les travaux, Matt Cioffi de la gestion des données, ainsi qu'Elizabeth Goldstein et Amy Heller des Centers for Medicare and Medicaid Services (CMS), et les autres membres de l'équipe de mise en œuvre de l'Enquête CAHPS-MMC.

Bibliographie

Cho, M.J., Eltinge, J.L., Gershunskaya, J. et Huff, L.L. (2002). Evaluation of generalized variance function estimators for the U.S. Current Employment Survey. Dans *Proceedings of the Joint Statistical Meetings* [CDROM]. Alexandra, VA: American Statistical Association, 534-539.

Eltinge, J. (2002). Use of generalized variance functions in multivariate analysis. Dans *Proceedings of the Section on Survey Research Methods*, Alexandra, VA: American Statistical Association, 904-913.

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.

Fay, R.E., et Train, G.F. (1997). Small domain methodology for estimating income and poverty characteristics for states in 1993. Dans *Proceedings of the Social Statistics Section*, Alexandra, VA: American Statistical Association, 183-188.

Freund, J.E., et Walpole, R.E. (1987). *Mathematical Statistics*. New Jersey: Prentice-Hall, Inc., 4ième Édition.

Gabler, S., Haeder, S. et Lahiri, P. (1999). Justification à base de modèle de la formule de Kish pour les effets de plan de sondage lies à la pondération et à l'effet de grappe. *Techniques d'enquête*, 25, 119-120.

Goldstein, E., Cleary, P.D., Langwell, K.M. Zaslavsky, A.M. et Heller, A. (2001). Medicare Managed Care CAHPS: A tool for performance improvement. *Health Care Financing Review*, 22, 101-107.

Ghosh, M., et Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-76.

Hays, R.D., Shaul, J.A., Williams, V.S.L., Lubalin, J.S., Harris-Kojetin, L.D., Sweeny, S.F. et Cleary, P.D. (1999). Psychometric properies of the CAHPS 1.0 survey measures. *Medical Care*, 37 (Supplément), 22-31.

Huff, L.L., Eltinge, J.L. et Gershunskaya, J. (2002). Exploratory analysis of generalized variance function models for the U.S. Current Employment Survey. Dans *Proceedings of the Joint Statistical Meetings* [CDROM], Alexandra, VA: American Statistical Association, 1519-1524.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Korn, E.L., et Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.

O'Malley, A.J. et Zaslavsky, A.M. (2004). Implementation of cluster-level covariance analysis for survey data with structured nonresponse. Dans *Proceedings of the Joint Statistical Meetings*, American Statistical Association, 1907-1914.

Otto, M.C., et Bell, W.R. (1995). Sampling error modeling of poverty and income statistics for states. Dans *Proceedings of the Section on Government Statistics*, American Statistical Association, 160-165.

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

Seber, G.A.F. (1977). *Linear Regression Analysis*. New York: John Wiley & Sons, Inc.

Spencer, B.D. (2000). Un effet de plan de sondage approximatif pour une pondération inégale en cas de corrélation possible entre les mesures et les probabilités de sélection. *Techniques d'enquête*, 26, 153-155.

Valliant, R. (1992a). Longitudinal smoothing of price index variances. Dans *Statistics Canada Symposium*. Ottawa: Statistique Canada. 113-120.

Valliant, R. (1992b). Smoothing variance estimates for price indexes over time. *Journal of Official Statistics*, 8, 433-444.

Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference*. New York: John Wiley & Sons, Inc.

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

Woodruff, S. (1992). Variance estimation for estimates of employment change in the Current Employment Statistics Survey. Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA: American Statistical Association, 626-631.

Zaslavsky, A.M., Beaulieu, N.D., Landon, B.E. et Cleary, P.D. (2000). Dimensions of consumer-assessed quality of Medicare managed-care health plans. *Medical Care*, 38, 162-174.

Zaslavsky, A.M., et Cleary, P.D. (2002). Dimensions of plan performance for sick and healthy members on the Consumer Assessments of Health Plans Study 2.0 Survey. *Medical Care*, 40, 951-964.

Techniques d'enquête, décembre 2005
Vol. 31, N° 2, pp. 201-214
Statistique Canada, N° 12-001-XPB au catalogue

Modèles spatio-temporels pour l'estimation pour petits domaines

Bharat Bhushan Singh, Girja Kant Shukla et Debasis Kundu [1]

Résumé

Nous proposons un modèle de régression spatial dans un cadre général de modèles à effets mixtes pour résoudre le problème de l'estimation pour petits domaines. L'utilisation d'un paramètre d'autocorrélation commun à l'ensemble de petits domaines permet de produire de meilleures estimations pour petits domaines. Ce paramètre s'avère fort utile dans les cas où l'utilisation de variables exogènes améliore peu ces estimations. Nous élaborons également une approximation de deuxième ordre de l'erreur quadratique moyenne (EQM) du meilleur prédicteur linéaire sans biais empirique (MPLNBE). En suivant l'approche des filtres de Kalman, nous proposons un modèle spatio-temporel. Dans ce cas également, nous obtenons une approximation de deuxième ordre de la EQM du MPLNBE. À titre d'étude de cas, nous utilisons les données de la série chronologique sur les dépenses de consommation mensuelles par habitant (DCMH) provenant de la National Sample Survey Organisation (NSSO) du ministère de la Statistique et de la Mise en œuvre des programmes du gouvernement de l'Inde pour valider les modèles.

Mots clés : Modèle linéaire à effets mixtes; autocorrélation spatiale; matrice de poids; meilleur prédicteur linéaire sans biais; meilleur prédicteur linéaire sans biais empirique, filtres de Kalman; cycles de la NSSO.

1. Introduction

La planification au niveau local nécessite des données fiables de niveau approprié. La réalisation de recensements complets ou de grandes enquêtes par sondage auprès d'un échantillon de taille adéquate est coûteuse et longue. Les recensements sont généralement réalisés une fois tous les dix ans, tandis que les enquêtes par sondage sont souvent planifiées pour fournir des estimations à un niveau beaucoup plus élevé d'agrégation. L'une de ces grandes enquêtes par sondage est l'enquête socioéconomique de la National Sample Survey Organisation (NSSO). Ici, les estimations par sondage directes sont disponibles au niveau du petit domaine (district), car la plupart des districts représentent une strate dans la procédure d'échantillonnage adoptée par la NSSO. Cependant, ces estimations sont très peu fiables, à cause d'erreurs-types inacceptablement grandes. Il est donc nécessaire de les renforcer à l'aide d'information provenant de petits domaines semblables ou de variables exogènes avec lesquelles un lien peut être établi, faciles à obtenir et reliées à la variable étudiée.

Diverses approches fondées sur un modèle ont été proposées pour améliorer les estimateurs directs. L'approche fondée sur un modèle facilite la validation au moyen de données d'échantillon. Le modèle simple caractéristique du domaine qui est proposé est le modèle à deux degrés de Fay et Herriot (1979).

$$y_i = \theta_i + \varepsilon_i,\ E(\varepsilon_i \mid \theta_i) = 0,\ \mathrm{Var}(\varepsilon_i \mid \theta_i) = \sigma_i^2, \quad (1.1)$$

$$\theta_i = X_i^T \beta + v_i z_i,\ E(v_i) = 0,\ \mathrm{Var}(v_i) = \sigma_v^2,\ i = 1, 2, \ldots, m. \quad (1.2)$$

Ici, les y_i sont des estimateurs directs par sondage des θ_i des caractéristiques étudiées. Les θ_i peuvent être les moyennes de petit domaine dans la population. Les $X_i = (X_{i1}, \ldots, X_{ip})^T$ sont des variables exogènes qui sont disponibles et que l'on suppose être étroitement liées aux θ_i et les z_i, sont des constantes positives connues. $\beta(p \times 1)$ est le vecteur des paramètres de régression.

La première équation (1.1) est le modèle fondé sur le plan de sondage, tandis que la deuxième (1.2) est le modèle de lien. Les ε_i sont les erreurs d'échantillonnage. Les estimateurs y_i sont sans biais par rapport au plan de sondage et les variances d'échantillonnage σ_i^2 sont connues. En outre, les ε_i et les v_i sont des variables aléatoires indépendantes et de même loi (iid). On suppose souvent que les erreurs et les effets aléatoires suivent une loi normale. Pour ce modèle, nous proposons le meilleur prédicteur linéaire sans biais (MPLNB) selon le modèle du meilleur estimateur linéaire sans biais (MELB). L'estimation est convergente par rapport au plan et sans biais par rapport au modèle (Ghosh et Rao 1994). Il s'agit typiquement de la moyenne pondérée de l'estimateur par sondage direct y_i et de l'estimateur synthétique par la régression $X_i^T\beta$. L'estimateur MPLNB dépend de la composante de la variance σ_v^2 qui est inconnue dans les applications pratiques. Diverses méthodes d'estimation des composantes de la variance σ_v^2 dans le modèle linéaire à effets mixtes général existent (Cressie 1992). En remplaçant σ_v^2 par un estimateur asymptotiquement convergent $\hat{\sigma}_v^2$, nous obtenons également un meilleur prédicteur linéaire sans biais empirique (MPLNBE).

1. Bharat Bhushan Singh, Girja Kant Shukla et Debasis Kundu, Department of Mathematics, I.I.T. Kanpur-208016. Courriel : drbbsingh@hotmail.com.

Dans le contexte de l'Inde, le principal problème que posent les données est l'absence de données de registre administratives ou de l'état civil au niveau du petit domaine. Souvent, il est difficile de trouver des variables exogènes étroitement liées à la variable étudiée (coefficient de corrélation multiple $R^2 > 0{,}5$).

Dans le présent article, nous envisageons l'exploitation d'une autocorrélation spatiale entre les petits domaines sous la forme d'un modèle spatial afin d'améliorer les estimateurs pour petits domaines. En outre, pour les données chronologiques, nous utilisons un modèle spatio-temporel du genre filtres de Kalman pour améliorer encore davantage les estimateurs. Nous étudions les données chronologiques sur les dépenses de consommation mensuelles par habitant (DCMH) estimées d'après une grande enquête par sondage réalisée par la National Sample Survey Organisation (NSSO). Dans le présent article, nous proposons des modèles appropriés dans le cadre du modèle linéaire à effets mixtes afin d'obtenir de meilleurs estimateurs des DCMH au niveau du petit domaine.

La présentation de la suite de l'article est la suivante. À la section 2, nous considérons un modèle spatial du genre du modèle linéaire à effets mixtes général avec introduction d'une autocorrélation spatiale entre les petits domaines. Nous présentons le meilleur prédicteur linéaire sans biais (MPLNB) et le meilleur prédicteur linéaire sans biais empirique (MPLNBE) des effets mixtes. Nous obtenons également une approximation de deuxième ordre de l'erreur quadratique moyenne (EQM) du MPLNBE et de l'estimateur de l'EQM. À la section 3, nous décrivons l'extension du modèle spatial aux séries chronologiques sous la forme d'un modèle spatio-temporel, suivant l'approche des filtres de Kalman. Nous discutons du MPLNB et du MPLNBE des effets mixtes, ainsi que d'une approximation de deuxième ordre de l'EQM du MPLNBE et de l'estimateur de l'EQM. À la section 4, nous présentons et analysons les estimations des DCMH provenant d'une grande enquête par sondage réalisée périodiquement en Inde. Enfin, à la section 5, nous présentons les conclusions de l'analyse des données. Toutes les preuves sont données en annexe.

2. Modèle spatial

Habituellement, les caractéristiques des petits domaines présentent une dépendance spatiale en ce qui concerne les similarités de quartier. Cressie (1990) a utilisé la dépendance spatiale conditionnelle pour les effets aléatoires dans le contexte de l'ajustement du sous-dénombrement au recensement. Ici, nous utilisons une dépendance spatiale simultanée (Cliff et Ord 1981) pour les effets aléatoires, qui présente certains avantages par rapport à la dépendance conditionnelle (Ripley 1981). Nous essayons donc de rendre compte d'une partie de l'erreur aléatoire non expliquée par les variables indépendantes afin d'améliorer les estimateurs par sondage directs. Le modèle proposé est un modèle à trois degrés caractéristique du domaine (Ghosh et Rao 1994).

$$y = \theta + \varepsilon, \qquad \varepsilon \sim N_m(0, R), \tag{2.1}$$

$$\theta = X\beta + u, \tag{2.2}$$

$$u = \rho W u + v, \qquad v \sim N_m(0, \sigma_v^2 I), \tag{2.3}$$

où θ est un vecteur de dimension m (correspondant au nombre de petits domaines) pour la caractéristique étudiée et y est son estimateur par sondage direct obtenu au moyen de données recueillies auprès d'un petit échantillon. Dans le modèle susmentionné, la première équation (2.1) montre le modèle fondé sur le plan de sondage (échantillonnage), la deuxième (2.2) montre le modèle de régression et la troisième (2.3), le modèle spatial sur les résidus, les deux dernières étant liées dans la première équation. Le modèle susmentionné peut s'exprimer sous la forme

$$y = X\beta + Zv + \varepsilon, \qquad Z = (I - \rho W)^{-1}, \tag{2.4}$$

où $X(m \times p)$ est la matrice d'expérience de plein rang colonne p, $\beta(p \times 1)$ est un vecteur-colonne de paramètres de régression et $Z(m \times m)$ représente les coefficients des effets aléatoires v. $W(m \times m)$ est une matrice de poids spatiaux connue qui montre le degré d'interaction dans toute paire de petits domaines. Les éléments de $W \equiv [W_{ij}]$ avec $W_{ii} = 0 \quad \forall i$ peuvent dépendre de la distance entre les centres des petits domaines ou de la longueur de leur frontière commune. À titre de solution de rechange simple, elle peut avoir des valeurs binaires $W_{ij} = 1$ (non échelonnées) si le j^{e} domaine est physiquement contigu au i^{e} domaine et $W_{ij} = 0,$ autrement. Nous avons normalisé la matrice de façon qu'elle satisfasse $\sum_{j=1}^{m} W_{ij} = 1$ pour $i = 1, 2, \ldots, m.$ La constante ρ est une mesure du niveau global d'autocorrélation spatiale et sa grandeur reflète la mesure dans laquelle W est appropriée sachant y et X. En outre, nous supposons que v et ε sont indépendantes l'une de l'autre. R est une matrice diagonale d'ordre m que l'on peut exprimer sous la forme $R = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)$ où les σ_i^2 sont les variances d'échantillonnage connues correspondant au i^{e} domaine. Le vecteur de paramètres $\psi = [\rho, \sigma_v^2]^T$ contient deux éléments.

Ce modèle est renforcé par l'emprunt de données à des petits domaines similaires grâce à deux paramètres communs, c'est-à-dire le paramètre de régression β et le paramètre d'autocorrélation ρ. Notons que le présent modèle est plus général et qu'il permet d'obtenir le modèle de Fay et Herriot (1979) en prenant $\rho = 0.$

En adoptant l'approche du modèle linéaire à effets mixtes (Henderson 1975), nous pouvons obtenir le meilleur

prédicteur linéaire sans biais (MPLNB) et $\theta = X\beta + Zv$ et l'erreur quadratique moyenne (EQM) du MPLNB comme suit

$$\hat{\theta}(\psi) = X\hat{\beta}(\psi) + \Lambda(\psi)[y - X\hat{\beta}(\psi)]$$
$$= \sigma_v^2 A^{-1}(\psi)\Sigma^{-1}(\psi)y + R\Sigma^{-1}(\psi)X\hat{\beta}(\psi), \quad (2.5)$$

$$\mathrm{EQM}[\hat{\theta}(\psi)] =$$
$$E[(\hat{\theta}(\psi) - \theta)(\hat{\theta}(\psi) - \theta)^T] = g_1(\psi) + g_2(\psi), \quad (2.6)$$

$$g_1(\psi) = \Lambda(\psi)R = R - R\Sigma^{-1}(\psi)R, \quad (2.7)$$

$$g_2(\psi) = R\Sigma^{-1}(\psi)X(X^T\Sigma^{-1}(\psi)X)^{-1}X^T\Sigma^{-1}(\psi)R, \quad (2.8)$$

$$\hat{\beta}(\psi) = [X^T\Sigma^{-1}(\psi)X]^{-1}X^T\Sigma^{-1}(\psi)y,$$
$$\Sigma(\psi) = \sigma_v^2 A^{-1}(\psi) + R,$$
$$\Lambda(\psi) = \sigma_v^2 A^{-1}(\psi)\Sigma^{-1}(\psi),\ A(\psi) = (I - \rho W)^T(I - \rho W).$$

Ici $\hat{\beta}, \Sigma$ et A sont tous des fonctions de ψ et sont habituellement exprimés sous la forme $\hat{\beta}(\psi), \Sigma(\psi)$ et $A(\psi)$, respectivement. Cependant, dans certains cas, pour abréger, le suffixe ψ est omis. Le premier terme, $g_1(\psi)$, dans l'expression de l'EQM, montre la variabilité de $\hat{\theta}$ quand tous les autres paramètres sont connus et est d'ordre $O(1)$. Le deuxième terme, $g_2(\psi)$, dû à l'estimation des effets fixes β, est d'ordre $O(m^{-1})$ pour les grandes valeurs de m. En outre, avec $\rho = 0$, le modèle susmentionné se réduit au modèle de régression linéaire à effets mixtes standard, tandis que, pour $X\beta = \mu$, nous obtenons un schéma purement spatial avec un terme d'ordonnée à l'origine uniquement.

En pratique, le paramètre ψ est inconnu et estimé d'après les données. Nous obtenons l'estimateur du maximum de vraisemblance (EMV) du paramètre ψ en maximisant la fonction de log-vraisemblance de ψ suivante

$$l = \mathrm{const} - \frac{1}{2}\log[|\Sigma(\psi)|]$$
$$-\frac{1}{2}[y - X\hat{\beta}(\psi)]^T\Sigma^{-1}(\psi)[y - X\hat{\beta}(\psi)] \quad (2.9)$$

par rapport au paramètre ψ. Nous obtenons le meilleur prédicteur linéaire sans biais empirique (MPLNBE), $\hat{\theta}(\psi)$ et l'estimateur naïf de l'EQM d'après les équations (2.5) et (2.6), respectivement, en remplaçant le vecteur de paramètres ψ par son estimateur $\hat{\psi}$.

$$\hat{\theta}(\hat{\psi}) = \sigma_v^2 A^{-1}(\hat{\psi})\Sigma^{-1}(\hat{\psi})y + R\Sigma^{-1}(\hat{\psi})X\hat{\beta}(\hat{\psi}), \quad (2.10)$$

$$\mathrm{EQM}[\hat{\theta}(\hat{\psi})] = g_1(\hat{\psi}) + g_2(\hat{\psi}), \quad (2.11)$$
$$\text{où} \quad \Sigma(\hat{\psi}) = \hat{\sigma}_v^2 A^{-1}(\hat{\psi}) + R$$
$$\text{et} \quad A(\hat{\psi}) = (I - \hat{\rho}W)^T(I - \hat{\rho}W).$$

Cette expression de la EQM du MPLNBE sous-estime gravement la EQM réelle, car nous n'avons pas tenu compte de la variabilité due à l'estimation des paramètres d'après les données. Nous obtenons une approximation de deuxième ordre de $\mathrm{EQM}[\hat{\theta}(\hat{\psi})]$ dans le cas où $\hat{\psi}$ est l'estimateur du maximum de vraisemblance (EMV) ou l'estimateur du maximum de vraisemblance restreint (EMVR) de ψ, en supposant que la valeur de m est grande et en négligeant tous les termes d'ordre $o(m^{-1})$, sous les conditions de régularité qui suivent. Nous avons calculé l'approximation en nous inspirant des méthodes de Prasad et Rao (1990) et de Datta et Lahiri (2000) qui sont de nature heuristique.

Conditions de régularité 1

a) Les éléments de X sont bornés uniformément de sorte que $X^T\Sigma^{-1}(\psi)X = [O(m)]_{p\times p}$, où $\Sigma(\psi) = [\sigma_v^2 A^{-1}(\psi) + R]$;

b) m est fini;

c) $\Lambda(\psi)X = [O(1)]_{m\times p}$, $(\partial[\Lambda(\psi)X])/(\partial\psi_d) = [O(1)]_{m\times p}$, $(\partial^2[\Lambda(\psi)])/(\partial\psi_d\partial\psi_e) = [O(1)]_{m\times m}$ pour $d, e = 1, 2$;

d) $\hat{\psi}$ est l'estimateur de ψ qui satisfait $\hat{\psi} - \psi = O_p(m^{-1/2})$, $\hat{\psi}(-y) = \hat{\psi}(y)$, $\hat{\psi}(y + xh) = \hat{\psi}(y)\ \forall h \in R^p$ et $\forall y$.

Ces conditions de régularité sont satisfaites dans ce cas. La forme normalisée spéciale de la matrice de poids W satisfait la condition (c) pour $|\rho| < 1$, car elle ne contient qu'un nombre fini d'éléments non nuls et la somme des lignes est égale à 1. Il convient de mentionner ici que la matrice $\sigma_v^2 A^{-1}\Sigma^{-1}$ contient un nombre fini d'éléments non nuls et que l'ordre de $W, (I - \rho W)$, $W(I - \rho W), \Sigma, \Sigma^{-1}$ ou toute combinaison additive ou multiplicative de ces entités et de leurs dérivées mentionnées à la condition (c) n'augmente pas. En outre l'EMV et l'EMVR satisfont la condition (d). Nous montrons au théorème A.1 qu'une approximation de deuxième ordre de l'EQM du MPLNBE est

$$\mathrm{EQM}[\hat{\theta}(\hat{\psi})] = E[\hat{\theta}(\hat{\psi}) - \theta)(\hat{\theta}(\hat{\psi}) - \theta)^T]$$
$$= g_1(\psi) + g_2(\psi) + g_3(\psi) + o(m^{-1}). \quad (2.12)$$

Ici, le troisième terme $g_3(\psi)$ provient de l'estimation du vecteur de paramètres inconnu provenant des données d'échantillon et est de même ordre $O(m^{-1})$ que $g_2(\psi)$. En outre, $g_3(\psi)$ peut s'exprimer sous la forme

$$g_3(\psi) = L^T(\psi)[I_\psi^{-1}(\psi) \otimes \Sigma(\psi)]L(\psi), \quad (2.13)$$

où

$$L(\psi) = \underset{1\le d\le 2}{\mathrm{Col}}[L_d(\psi)] = [L_\rho(\psi), L_{\sigma_v^2}(\psi)]^T,\ L_d(\psi) =$$
$$\frac{\partial\Lambda(\psi)}{\partial\psi_d},\ d = 1, 2. \quad I_\psi(\psi) = E[-\frac{\partial^2 l}{\partial\psi\partial\psi^T}]$$

est la matrice d'information et $\otimes$ représente le produit de Kronecker. En outre, $g_3(\psi)$ peut aussi s'écrire

$$g_3(\psi) = \sum_{d=1}^{2}\sum_{e=1}^{2} L_d(\psi)\Sigma(\psi)L_e^T(\psi)I_{de}^{-1}(\psi) \qquad (2.14)$$

où $I_{\psi}^{-1}(\psi) \equiv (I_{de}^{-1}(\psi))$.

En pratique, il est courant d'estimer l'EQM du MPLNBE en remplaçant les paramètres inconnus, y compris les composantes de la variance, par leurs estimateurs respectifs. Cette méthode peut donner lieu à une sous-estimation grave de l'EQM réelle (Prasad et Rao 1990; Singh, Stukel et Pfeffermann 1998). Nous obtenons l'estimateur de l'EQM du MPLNBE au théorème A.2 en annexe pour une grande valeur de *m*, en négligeant tous les termes d'ordre $o(m^{-1})$. Par conséquent, nous avons les expressions

$$E[g_1(\hat{\psi}) + g_3(\hat{\psi}) - g_4(\hat{\psi}) - g_5(\hat{\psi})] = g_1(\psi) + o(m^{-1}), \qquad (2.15)$$

$$E[g_2(\hat{\psi})] = g_2(\psi) + o(m^{-1})$$
$$\text{et} \quad E[g_3(\hat{\psi})] = g_3(\psi) + o(m^{-1}), \qquad (2.16)$$

et, enfin, l'estimateur de l'EQM de $\hat{\theta}(\hat{\psi})$ sous la forme

$$\text{eqm}[\hat{\theta}(\hat{\psi})] =$$
$$[g_1(\hat{\psi}) + g_2(\hat{\psi}) + 2g_3(\hat{\psi}) - g_4(\hat{\psi}) - g_5(\hat{\psi})] + o(m^{-1}), \qquad (2.17)$$
$$\text{où } E[\text{eqm}(\hat{\theta}(\hat{\psi}))] = \text{EQM}[\hat{\theta}(\hat{\psi})] + o(m^{-1}).$$

De toute évidence, les termes supplémentaires, $g_3(\hat{\psi})$, $g_4(\hat{\psi})$ et $g_5(\hat{\psi})$ sont les contributions dues à l'estimation du vecteur de paramètres inconnu ψ au moyen de $\hat{\psi}$. Les expressions pour $g_4(\psi)$ et $g_5(\psi)$ jusqu'à l'ordre $o(m^{-1})$ sont données par

$$g_4(\psi) = [b_{\hat{\psi}}^T(\psi) \otimes I_m]\frac{\partial g_1(\psi)}{\partial \psi},$$
$$b_{\hat{\psi}}(\psi) = \frac{1}{2} I_{\hat{\psi}}^{-1}(\psi) \operatorname*{Col}_{1\le d\le 2}\left[\text{Trace}\left[I_{\beta}^{-1}(\psi)\frac{\partial I_{\beta}(\psi)}{\partial \psi_d}\right]\right], \qquad (2.18)$$

$$g_5(\psi) = \frac{1}{2}\operatorname*{Trace}_m\left[\begin{matrix}[I_2 \otimes (R\Sigma^{-1}(\psi))] \\ \frac{\partial^2 \Sigma(\psi)}{\partial\psi\partial\psi^T}[I_{\psi}^{-1}(\psi) \otimes (\Sigma^{-1}(\psi)R)]\end{matrix}\right]. \qquad (2.19)$$

Ici, $b_{\hat{\psi}}(\psi)$ est le biais de $\hat{\psi}$, c'est-à-dire $E(\hat{\psi}) - \psi$ jusqu'à l'ordre $o(m^{-1})$ et $(\partial g_1(\psi))/(\partial\psi)$ est une matrice partitionnée $[(\partial g_1(\psi))/(\partial\rho), (\partial g_1(\psi))/(\partial\sigma_v^2)]^T$ de dimensions $(2m \times m)$ ayant deux matrices de dimension $m \times m$ dans une colonne. De la même façon, $(\partial^2\Sigma(\psi))/(\partial\psi\partial\psi^T)$ est une matrice partitionnée de dimension $(2m \times 2m)$ ayant deux partitions, en ce qui concerne les lignes et les colonnes, avec à l'intérieur, $(\partial^2\Sigma(\psi))/(\partial\psi_d\partial\psi_e)$, une sous-matrice générale de dimensions $m \times m$. $\text{Trace}(B) = \sum_{d=1}^{2} B_{dd}$, où B est une matrice carrée partitionnée en sous-matrices carrées de dimensions semblables. En outre, $g_4(\psi)$ et $g_5(\psi)$ peuvent aussi s'écrire

$$g_4(\psi) =$$
$$\frac{1}{2}\sum_{d=1}^{2}\sum_{e=1}^{2} I_{de}^{-1}(\psi)\,\text{Trace}\left[I_{\beta}^{-1}(\psi)\frac{\partial I_{\beta}(\psi)}{\partial\psi_d}\right]\frac{\partial g_1(\psi)}{\partial\psi_e}, \qquad (2.20)$$

$$g_5(\psi) =$$
$$\frac{1}{2}\sum_{d=1}^{2}\sum_{e=1}^{2}\left[R\Sigma^{-1}(\psi)\frac{\partial^2\Sigma(\psi)}{\partial\psi_d\partial\psi_e}\Sigma^{-1}(\psi)RI_{de}^{-1}(\psi)\right]. \qquad (2.21)$$

L'expression (2.17) donne la matrice de l'estimateur de l'EQM du MPLNBE, $\hat{\theta}(\hat{\psi})$ et l'EQM des estimateurs sur petit domaine individuels est donnée par les éléments diagonaux respectifs. Nous pouvons obtenir des expressions similaires pour un modèle simple sans autocorrélation spatiale. Cependant, dans ce cas, $g_5(\psi)$ devient nulle.

3. Modèle spatio-temporel

À la présente section, nous utilisons des modèles d'espace d'états, obtenus au moyen de filtres de Kalman, pour exploiter les données chronologiques, ainsi que le paramètre de régression commun et le paramètre d'autocorrélation commun en vue de renforcer les estimateurs par sondage directs en tout point dans le temps. Cette approche est particulièrement avantageuse dans le cas où les estimations par sondage antérieures sont plus fiables. Les modèles utilisés dans cette catégorie sont les suivants

$$y_t = X_t\beta + Zv_t + \varepsilon_t,\ \varepsilon_t \overset{\text{ind}}{\sim} N_m(0, R_t),\ Z = (I - \rho W)^{-1}, \quad (3.1)$$

$$v_t = kv_{t-1} + \eta_t,\ \eta_t \overset{\text{ind}}{\sim} N_m(0, \sigma_v^2 I)\ \ t = 1, 2, \ldots, T \ \text{ et}$$
$$\varepsilon_t \ \text{ et } \ \eta_t \ \text{ sont indépendantes l'une de l'autre.} \qquad (3.2)$$

Ici, les paramètres ont la signification habituelle décrite à la section précédente. La matrice de poids $W(m \times m)$ et les matrices d'expérience $X_t(m \times p)$ sont connues, $Z(m \times m)$ est une matrice de coefficients des effets aléatoires et ρ est un coefficient d'autocorrélation inconnu. R_t est une matrice diagonale de dimension *m* qui peut être exprimée sous la forme $R_t = \text{diag}(\sigma_{1t}^2, \sigma_{2t}^2, \ldots, \sigma_{mt}^2)$, où les σ_{it}^2 sont les variances d'échantillonnage connues correspondant au i^{e} petit domaine et au t^{e} point dans le temps. β est un vecteur inconnu d'effets fixes et $\psi = [\rho, \sigma_v^2, k]^T$ est un vecteur de trois paramètres inconnus. Ces paramètres sont indépendants du temps *t*. Il convient de souligner que les effets aléatoires v_t peuvent varier conformément à (3.2) et que *k* est un paramètre temporel autorégressif. Pour la stationnarité, $|k| < 1$.

Les estimateurs des effets fixes et aléatoires et de la EQM de ces estimateurs sont obtenus par étape, en partant de l'hypothèse d'une approche par modèle linéaire à effets mixtes au temps $t=1$, et en prenant $v_1 \sim N_m(0, \sigma_v^2 I)$ (Sallas et Harville 1994). Sous la forme standard, nous écrivons le modèle comme suit

$$y_t = U_t\alpha_t + \varepsilon_t,\ \alpha_t = T\alpha_{t-1} + \zeta_t,\ T = \text{diag}[I_p, kI_m], \qquad (3.3)$$

$$\zeta_t \sim N_{p+m}(0, Q), \quad Q = \text{diag}[0_p, \sigma_v^2 I_m]$$
$$U_t = [X_t, Z],\ \alpha_t = [\beta_t, v_t]^T. \qquad (3.4)$$

Ici, I_m et 0_m sont la matrice unité et la matrice nulle de dimension *m*, et $\text{diag}[I_p, kI_m]$ représente la matrice

$$\begin{bmatrix} I_{p\times p} & 0_{p\times m} \\ 0_{m\times p} & kI_{m\times m} \end{bmatrix}.$$

Si l'on suppose que β est fixe, mais qu'il dépend du temps, le modèle ne change pas, sauf que $T = \text{diag}[0_p, kI_m]$.

Les estimations initiales des effets α_t et de leur variance (basées sur $t=1$) sont obtenues comme suit

$$\hat{\beta}_1 = (X_1^T H_1^{-1} X_1)^{-1} X_1^T H_1^{-1} y_1,\ \hat{v}_1 = \sigma_v^2 Z^T H_1^{-1}(y_1 - X_1\hat{\beta}_1),$$
$$H_1 = R_1\sigma_v^2 A^{-1}, \quad \Sigma_1 = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

$$\Sigma_{11}(p\times p) = (X_1^T H_1^{-1} X_1)^{-1},$$
$$\Sigma_{12}(p\times m) = \Sigma_{21}^T = -\sigma_v^2 (X_1^T H_1^{-1} X_1)^{-1} X_1^T H_1^{-1} Z$$
$$\text{et } \Sigma_{22}(m\times m) = \sigma_v^2 I_m - \sigma_v^4 Z^T H_1^{-1} Z$$
$$+\sigma_v^4 Z^T H_1^{-1} X_1 (X_1^T H_1^{-1} X_1)^{-1} X_1^T H_1^{-1} Z.$$

Les équations des filtres de Kalman récurrentes pour la mise à jour des estimateurs aux étapes subséquentes sont

$$\Sigma_{t|t-1} = T\Sigma_{t-1}T^T + Q,\ \hat{\alpha}_{t|t-1} = T\hat{\alpha}_{t-1},\ H_t = R_t + U_t\Sigma_{t|t-1}U_t^T,$$
$$\hat{\alpha}_t = \hat{\alpha}_{t|t-1} + \Sigma_{t|t-1}U_t^T H_t^{-1}(y_t - U_t\hat{\alpha}_{t|t-1}),$$
$$\Sigma_t = \Sigma_{t|t-1} - \Sigma_{t|t-1}U_t^T H_t^{-1} U_t \Sigma_{t|t-1}$$

où les $\hat{\alpha}_{t|t-1}$ sont les estimateurs des effets α_t sachant les observations $[y_1, y_2, \ldots, y_{t-1}]$ et les $\Sigma_{t|t-1}$ sont les erreurs quadratiques moyennes de $\hat{\alpha}_{t|t-1}$. H_t représente la matrice des variances-covariances conditionnelle de y_t sachant $[y_1, y_2, \ldots, y_{t-1}]$. À l'aide des équations de filtre récurrentes, nous pouvons obtenir le meilleur prédicteur linéaire sans biais (MPLNB) de $\theta_t = X_t\beta + Zv_t$, et l'erreur quadratique moyenne (EQM) du MPLNB comme suit

$$\begin{aligned}\hat{\theta}_t(\psi) &= U_t(\psi)\hat{\alpha}_t(\psi) \\ &= y_t - R_t H_t^{-1}(\psi)[y_t - U_t(\psi)\hat{\alpha}_{t|t-1}(\psi)] \\ &= U_t(\psi)\hat{\alpha}_{t|t-1}(\psi) + \Lambda_t(\psi)e_t(\psi), \end{aligned} \qquad (3.5)$$

$$\text{EQM}[\hat{\theta}_t(\hat{\psi})] = g_{12t}(\psi) = U_t(\psi)\Sigma_t(\psi)U_t^T(\psi), \qquad (3.6)$$

où $\Lambda_t(\psi) = U_t(\psi)\Sigma_{t|t-1}(\psi)U_t^T(\psi)H_t^{-1}(\psi)$
$= I_m - R_t H_t^{-1}(\psi)$
et $e_t(\psi) = y_t - U_t(\psi)\hat{\alpha}_{t|t-1}(\psi).$

Soulignons que $g_{12t}(\psi)$ est l'analogue spatial de $g_1(\psi) + g_2(\psi)$. Comme il est fréquent en pratique, le vecteur de paramètres ψ est inconnu et ses estimateurs du maximum de vraisemblance restreint (EMVR) peuvent être obtenus en maximisant la fonction de log-vraisemblance suivante, d'après les données d'échantillon couvrant tous les points dans le temps.

$$\begin{aligned} l = \text{const.} &- \frac{1}{2}\log[|X_1^T H_1^{-1} X_1|] - \frac{1}{2}\sum_{t=1}^{T}\log[|H_t|] \\ &- \frac{1}{2}(y_1 - X_1\hat{\beta}_1)^T H_1^{-1}(y_1 - X_1\hat{\beta}_1) \\ &- \frac{1}{2}\sum_{t=2}^{T}(y_t - U_t\hat{\alpha}_{t|t-1})^T H_t^{-1}(y_t - U_t\hat{\alpha}_{t|t-1}) \end{aligned} \qquad (3.7)$$

en ce qui concerne le paramètre ψ. À l'aide de l'équation qui précède, nous obtenons l'estimateur $\hat{\psi}$, et le MPLNBE de θ_t et l'estimateur naïf de l'EQM du MPLNBE sont donnés par

$$\hat{\theta}_t(\hat{\psi}) = U_t(\hat{\psi})\hat{\alpha}_t(\hat{\psi}) = U_t(\hat{\psi})\hat{\alpha}_{t|t-1}(\hat{\psi}) + \Lambda_t(\hat{\psi})e_t(\hat{\psi}), \qquad (3.8)$$

$$\text{EQM}[\hat{\theta}_t(\hat{\psi})] = g_{12t}(\hat{\psi}) = U_t(\hat{\psi})\Sigma_t(\hat{\psi})U_t^T(\hat{\psi}). \qquad (3.9)$$

Comme nous l'avons expliqué plus haut à la section 2, l'EQM du MPLNBE sous-estime l'EQM réelle car elle ne tient pas compte de la variabilité due au remplacement des paramètres par leurs estimations. Au théorème A.3 en annexe, nous obtenons une approximation de deuxième ordre de l'$\text{EQM}[\hat{\theta}_t(\hat{\psi})]$ pour une grande valeur de *m* et en négligeant tous les termes d'ordre $o(m^{-1})$, sous les conditions de régularité qui suivent satisfaites par notre modèle. Ces conditions sont analogues aux conditions de régularité 1.

Conditions de régularité 2

a) Les éléments de $X_t, t = 1, 2, \ldots, T$ sont bornés uniformément de sorte que $X_t^T\Sigma_t^{-1}(\psi)X_t = [O(m)]_{p\times p}$, où $\Sigma_t(\psi) = [\sigma_v^2 A^{-1}(\psi) + R_t]$;

b) *m* et *T* sont finis;

c) $\Lambda_t(\psi)U_t(\psi) = [O(1)]_{m\times p}$, $(\partial[\Lambda_t(\psi)U_t(\psi)])/(\partial\psi_d) = [O(1)]_{m\times p}$, $([\partial^2\Lambda_t(\psi)])/(\partial\psi_d\partial\psi_e) = [O(1)]_{m\times m}$, $t = 1, 2, \ldots, T$ et $d, e = 1, 2, 3$;

d) $\hat{\psi}$ est l'estimateur de ψ qui satisfait $\hat{\psi} - \psi = O_p(m^{-1/2})$, $\hat{\psi}(-y) = \hat{\psi}(y)$, $\hat{\psi}(y + xh) = \hat{\psi}(y)\ \forall h \in R^p$ et $\forall$ y.

L'approximation de deuxième ordre de l'EQM du MPLNBE est

$$\begin{aligned}\text{EQM}[\hat{\theta}_t(\hat{\psi})] &= E[(\hat{\theta}_t(\hat{\psi})-\theta_t)(\hat{\theta}_t(\hat{\psi})-\theta_t)^T] \\ &= g_{12t}(\psi)+g_{3t}(\psi)+o(m^{-1}). \qquad (3.10)\end{aligned}$$

Ici, $g_{3t}(\psi)$ est le biais dû à l'estimation des paramètres d'après des données d'échantillon qui est d'ordre $O(m^{-1})$ et est donné par

$$g_{3t}(\psi) = L_t^T(\psi) I_\psi^{-1}(\psi) K_\psi(\psi) H_t I_\psi^{-1}(\psi) L_t(\psi) \qquad (3.11)$$

où $K_\psi(\psi) \equiv (K_{de}(\psi))$

et $$K_{de}(\psi) = \frac{1}{2}\sum_{i=1}^{T} \text{Trace}\left[H_i^{-1} \frac{\partial H_i}{\partial \psi_d} H_i^{-1} \frac{\partial H_i}{\partial \psi_e} \right]. \qquad (3.12)$$

En outre,

$L_t(\psi) = \underset{1\le d\le 3}{\text{Col}}[L_{td}(\psi)]$ et $L_{td}(\psi) = (\partial \Lambda_t(\psi))/(\partial \psi_d)$

pour $d = 1, 2, 3$.

Sous une forme correcte, nous pouvons écrire $g_{3t}(\psi)$ comme suit

$$g_{3t}(\psi) = \sum_{d=1}^{3}\sum_{e=1}^{3} L_{td}(\psi) \left[\begin{array}{l} \sum_{f=1}^{3}\sum_{g=1}^{3} I_{df}^{-1}(\psi) \\ \quad \times \sum_{i=1}^{T} \text{Trace}\left(H_i^{-1} \frac{\partial H_i}{\partial \psi_f} H_i^{-1} \frac{\partial H_i}{\partial \psi_g} \right) \\ \quad \times H_t I_{ge}^{-1}(\psi) \end{array} \right] L_{te}^T(\psi).$$

L'expression pour la matrice d'information intervenant ici peut être donnée par

$$\begin{aligned} I_{de}(\psi) &= E\left[-\frac{\partial^2 l}{\partial \psi_d \partial \psi_e} \right] \\ &= \frac{1}{2}\sum_{t=1}^{T} \text{Trace}\left[H_t^{-1} \frac{\partial H_t^{-1}}{\partial \psi_d} H_t^{-1} \frac{\partial H_t}{\partial \psi_e} \right] + \sum_{t=1}^{T} \left[\frac{\partial e_t^T}{\partial \psi_d} H_t^{-1} \frac{\partial e_t}{\partial \psi_e} \right] \\ &\quad - \frac{1}{2}\text{Trace}\left[\begin{array}{l} (X_1^T H_1^{-1} X_1)^{-1} X_1^T H_1^{-1} \\ \quad \times \left(\frac{\partial^2 H_1}{\partial \psi_d \partial \psi_e} - 2 \frac{\partial H_1}{\partial \psi_d} H_1^{-1} \frac{\partial H_1}{\partial \psi_e} \right) H_1^{-1} X_1 \end{array} \right] \\ &\quad - \frac{1}{2}\text{Trace}\left[\begin{array}{l} (X_1^T H_1^{-1} X_1)^{-1} X_1^T H_1^{-1} \frac{\partial H_1}{\partial \psi_d} H_1^{-1} X_1 \\ \quad \times (X_1^T H_1^{-1} X_1)^{-1} X_1^T H_1^{-1} \frac{\partial H_1}{\partial \psi_e} H_1^{-1} X_1 \end{array} \right]. \end{aligned}$$

Nous avons également obtenu l'estimateur de l'EQM du MPLNBE sous l'hypothèse d'une grande valeur de m et en négligeant tous les termes d'ordre $o(m^{-1})$ dans le théorème A.4 en annexe sous la forme

$$\begin{aligned}\text{eqm}[\hat{\theta}_t(\hat{\psi})] &= [g_{12t}(\hat{\psi}) + g_{3t}(\hat{\psi}) + g_{31t}(\hat{\psi}) \\ &\quad - g_{4t}(\hat{\psi}) - g_{5t}(\hat{\psi})] + o(m^{-1}), \qquad (3.13)\end{aligned}$$

où $g_{31t}(\psi)$, $g_{4t}(\psi)$ et $g_{5t}(\psi)$ sont donnés par

$$g_{31t}(\psi) = L_t^T(\psi)[I_\psi^{-1}(\hat{\psi}) \otimes H_t(\psi)] L_t(\psi), \qquad (3.14)$$

$$g_{4t}(\psi) = [b_{\hat{\psi}}^T(\psi) \otimes I_m] \frac{\partial g_{12t}(\psi)}{\partial(\psi)},$$

$$b_{\hat{\psi}} = \frac{1}{2} I_\psi^{-1}(\psi) \underset{1\le d\le 3}{\text{Col}} \left[\text{Trace}\left[I_\beta^{-1}(\psi) \frac{\partial I_\beta(\psi)}{\partial \psi_d} \right] \right], \qquad (3.15)$$

$$g_{5t}(\psi) = \frac{1}{2}\underset{m}{\text{Trace}}\left[\begin{array}{l} [I_3 \otimes (R_t H_t^{-1})] \\ \frac{\partial^2 H_t}{\partial \psi \partial \psi^T} [I_\psi^{-1}(\psi) \otimes (H_t^{-1} R_t)] \end{array} \right]. \qquad (3.16)$$

4. Analyse des données de la NSSO

La National Sample Survey Organisation (NSSO) du ministère de la Statistique et de la Mise en œuvre des programmes (gouvernement de l'Inde) réalise de grandes enquêtes par sondage quinquennales (EQ) sur les dépenses de consommation des ménages et l'emploi, presque tous les cinq ans en Inde. Le champ d'observation de ces enquêtes compte plus de 100 000 ménages répartis entre plusieurs villages et îlots urbains. Afin de combler les lacunes statistiques entre les EQ successives, la NSSO réalise une enquête sur les dépenses de consommation (EDC) annuelle lors de presque chaque cycle (équivalant à une période de six mois ou d'un an). La série annuelle ne couvre que de 10 000 à 30 000 ménages selon le nombre de villages et d'îlots urbains visés par l'enquête dans l'ensemble du pays. Chaque cycle de la NSSO porte généralement sur plus d'un thème. Chaque enquête annuelle porte sur un thème principal différent. Cependant, le questionnaire 1.0 de ces enquêtes est conçu en vue de recueillir des données sur les dépenses de consommation des ménages parmi d'autres caractéristiques de l'emploi.

La NSSO utilise un plan d'échantillonnage stratifié à deux degrés, où les unités de premier degré sont les villages de recensement dans le secteur rural sélectionnés par échantillonnage circulaire systématique avec probabilité proportionnelle à la taille (PPT) et les unités de deuxième degré sont les ménages sélectionnés par échantillonnage circulaire systématique avec des points de départ aléatoires indépendants. L'Inde est subdivisée en États et les districts sont les unités administratives de deuxième degré dans les États. Les enquêtes annuelles et quinquennales diffèrent peu, à l'exception du fait que, normalement, dans la série

annuelle, l'enquête est réalisée auprès d'un petit échantillon comptant quatre ménages par unité de premier degré, tandis que dans la série quinquennale, l'enquête est réalisée auprès d'un échantillon de 10 à 12 ménages par unité de premier degré. À part cela, les enquêtes de la NSSO comportent deux échantillons, c'est-à-dire un échantillon central réalisé par les chercheurs de la NSSO et un échantillon d'État effectué par les autorités de l'État. En ce qui concerne la procédure d'estimation, les unités de premier degré sont sélectionnées sous forme de deux sous-échantillons indépendants. L'estimation de la moyenne de population et de sa variance est calculée séparément d'après les deux sous-échantillons. La moyenne groupée $y_i = (\hat{y}_{1i} + \hat{y}_{2i})/2$ et $R_i = (\hat{y}_{1i} - \hat{y}_{2i})^2/4$ pour $i = 1, 2, \ldots, m$, où $\hat{y}_{1i}, \hat{y}_{2i}$ sont les moyennes de sous-échantillon, estiment, respectivement, la moyenne de population et sa variance pour un district (petit domaine) particulier. Dans le cas du cycle 55, les unités de premier degré ont été sélectionnées sous la forme de huit sous-échantillons indépendants et l'estimation de la moyenne de population et celle de sa variance ont été calculées d'après ces sous-échantillons. Étant donné les problèmes que posent les estimations des R_i avec un degré de liberté, le R_i pour chaque petit domaine a été analysé et comparé au cours du temps. Toute valeur anormale de R_i, a été lissée par calcul de la moyenne des R_i sur les points dans le temps voisins et, dans certains cas, sur les petits domaines avoisinants également. Les estimations par sondage y_i sont les estimations directes et les R_i lissés sont les éléments diagonaux de la matrice des variances-covariances d'échantillonnage R dans nos équations modélisées (2.1), (2.4) et (3.1).

Nous n'utilisons ici que les données provenant de l'échantillon central. Nous avons calculé les estimations des dépenses de consommation mensuelles par habitant (DCMH) et des erreurs-types (e.-t.) des estimateurs sous divers modèles à effets mixtes pour les 63 districts (petits domaines) ruraux d'un grand État de l'Inde, à savoir l'Uttar Pradesh. Nous avons utilisé les données provenant de six cycles de la NSSO, c'est-à-dire le cycle 50 (juillet 1993 à juin 1994), le cycle 51 (juillet 1994 à juin 1995), le cycle 52 (juillet 1995 à juin 1996), le cycle 53 (janvier à décembre 1997), le cycle 54 (janvier à juin 1998) et le cycle 55 (juillet 1999 à juin 2000). De ceux-ci, les cycles 50 et 55 sont fondés sur des enquêtes quinquennales. Les variables exogènes choisies pour être utilisées dans les modèles sont i) le nombre de ménages, ii) la superficie brute ensemencée et iii) la superficie nette ensemencée par habitant dans les district. Les données agricoles sont disponibles sur une base annuelle, tandis que les estimations des ménages et de la population ont été obtenues par une méthode d'interpolation d'après les données des recensement décennaux de 1971, 1981 et 1991. Ces variables exogènes ont été sélectionnées par analyse des covariances parmi une foule de variables allant des variables du Recensement de 1991 à celles couvertes par les données annuelles sur l'agriculture. Nous avons examiné diverses matrices de poids, comme la longueur de la frontière commune entre deux districts, la distance entre les centres de district ou les poids binaires. Comme ces derniers donnent une estimation plus grande du coefficient d'auto-corrélation spatiale, nous les avons utilisés pour poursuivre l'analyse présentée ici (après les avoir normalisés en rendant la somme des éléments de chaque ligne de la matrice des poids égale à un). Dans tout l'exercice, nous avons procédé à la maximisation de la fonction de log-vraisemblance et à l'estimation des paramètres par la méthode simplex de Nelder et Mead au moyen du logiciel MATLAB.

Divers modèles à effets mixtes utilisés pour trouver des estimations améliorées des DCMH sont présentés au tableau 1. Les paramètres de ces modèles ont la signification habituelle mentionnée aux sections 2 et 3. En outre, dans le cas de chaque modèle, nous supposons que la variance d'échantillonnage R ou R_t (dans le cas du modèle temporel) est connue.

Tableau 1
Modèles à effets mixtes

Modèle – 1	Estimations directes	
Modèle – 2	Modèle de régression	$y = X\beta + v + \varepsilon$
Modèle – 3	Modèle spatial	$y = X\beta + Zv + \varepsilon$
Modèle –3A	Modèle spatial (ordonnées à l'origine)	$y = \mu + Zv + \varepsilon$
Modèle – 4	Modèle de régression temporel	$y_t = X_t\beta + v_t + \varepsilon_t, v_t = kv_{t-1} + \eta_t$
Modèle – 5	Modèle spatio-temporel	$y_t = X_t\beta + Zv_t + \varepsilon_t, v_t = kv_{t-1} + \eta_t$

Le tableau 2 présente les estimations du modèle de régression simple et des modèles spaciaux à effets mixtes. La valeur du coefficient de corrélation multiple R^2 entre les estimations des DCMH et les variables auxiliaires est également présentée pour chaque cycle. Les erreurs-types (e.-t.) des estimations des paramètres sont indiquées entre parenthèses. Notons que $\lambda(=\lambda_1, \lambda_2)$ est la statistique du test du rapport des vraisemblances (LRT pour *likelihood ratio test*) défini comme étant $-2\log L \sim \chi_k^2$, où L est le ratio entre les vraisemblances emboîtées et les valeurs hypothétiques des paramètres pour deux modèles concurrents sous diverses hypothèses, et k est la différence entre le nombre de paramètres sous les deux modèles. Ici, λ_1 compare le modèle de régression et le modèle spatial, sous $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$ et suit la loi de χ_1^2 sous H_0, et λ_2 compare le modèle spatial et le modèle spatial (avec ordonnée à l'origine), sous $H_0 : \beta = 0$ contre $H_1 : \beta \neq 0$ [β n'inclut pas le terme d'ordonnée à l'origine β_0] et suit la loi de χ_3^2 sous H_0.

Lorsque nous comparons le modèle de régression simple (modèle 2) et le modèle spatial (modèle 3) au moyen de la statistique LRT, nous constatons que, sous $H_0(\rho = 0)$, pour le modèle 3, l'autocorrélation spatiale ρ est hautement significative pour les cycles 52 et 55; de toute évidence, pour ces deux cycles, l'utilisation du modèle spatial améliore considérablement les estimations des DCMH. Par ailleurs, dans le cas des cycles 50 et 53, et pour ceux-ci uniquement, les coefficients de régression β sont presque significatifs pour le modèle 3 comparativement au modèle 3A, ce qui indique que le modèle spatial avec le terme d'ordonnée à l'origine pourrait améliorer les estimations pour ces cycles sans qu'on ait besoin de variables exogènes.

Le tableau 3 donne les estimations des paramètres et leur erreur-type dans le cas du modèle de régression temporel et du modèle spatio-temporel.

Dans le cas du modèle 4, le processus de maximisation itérative sans contrainte a convergé pour une valeur de k supérieure à 1, ce qui est inadmissible sous l'hypothèse de stationnarité. Pour ce cas, nous avons obtenu les estimations en prenant $k = 1$ et le modèle 4 a été modifié en conséquence. Le tableau 3 donne les résultats pour $k = 1$ dans le cas du modèle de régression temporel. Le modèle spatio-temporel produit une valeur plus élevée du coefficient d'autocorrélation commun et une valeur nettement plus faible de l'estimation de σ_v^2. Au tableau 4, nous résumons les estimations moyennes par cycle des DCMH (fondées sur l'ensemble des 63 districts), leurs erreurs-types (e.-t.) estimées et le coefficient de variation (c.v.) sous chaque modèle.

Nous résumons les résultats du tableau 4 ci-après.

Les estimations directes par sondage sont moins précises et tous les modèles comprenant des termes d'effets mixtes les améliorent. Les estimations pour les cycles 50 et 55 (fondées sur de grands échantillons) sont plus précises que celles obtenues pour les autres cycles. Le modèle spatial, qui dépend de la valeur de ρ, améliore considérablement les estimations. Dans le cas des cycles 52 et 55, où l'autocorrélation s'avère significative, la réduction de l'e.-t. moyenne des estimations est importante comparativement aux modèles sans autocorrélation spatiale. Le modèle 3A avec effet spatial et sans variables auxiliaires est tout aussi bon. Le modèle spatio-temporel améliore encore davantage les estimations en tirant parti des considérations d'espace d'états. Il convient de souligner que, pour le cycle 52 (très forte autocorrélation spatiale), les estimations au moyen des modèles temporels sont moins bonnes que celles ne tenant pas compte du temps. Peut-être parce que les paramètres de régression et d'autocorrélation sont fixes, les estimations tendent vers la moyenne des cinq cycles.

Tableau 2
Estimations des paramètres pour les estimations par petits domaines des DCMH sous le modèle de régression et les modèles spatiaux

Cycle	R^2	Modèle 2	Modèle 3		LRT	Modèle 3A		LRT
		σ_v^2	ρ	σ_v^2	λ_1	ρ	σ_v^2	λ_2
50	0,27	1 724,48	0,30	1 635,70	1,80	0,59	1 724,68	6,64
		(356,19)	(0,18)	(346,45)		(0,13)	(378,66)	
51	0,27	3 424,21	0,48	3 156,90	0,66	0,67	3 022,32	4,54
		(820,89)	(0,19)	(815,24)		(0,13)	(824,54)	
52	0,17	2 150,54	0,87	714,96	13,46	0,86	768,11	0,90
		(540,23)	(0,07)	(257,15)		(0,07)	(272,27)	
53	0,13	6 312,99	–0,39	5 822,99	1,56	0,09	7 141,60	7,66
		(1 397,92)	(0,27)	(1 374,70)		(0,23)	(1 561,72)	
54	0,22	3 437,67	0,61	2 793,24	1,30	0,66	2 888,66	3,00
		(806,87)	(0,14)	(742,35)		(0,13)	(768,84)	
55	0,31	2 989,73	0,87	1 060,21	20,30	0,86	1 186,58	1,56
		(712,28)	(0,06)	(362,40)		(0,07)	(394,27)	

λ_1 et λ_2 comparent les modèles 2, 3 et les modèles 3, 3A respectivement. $\chi^2_{1,.05}$=3,841 pour λ_1 et $\chi^2_{3,.05} = 7{,}815$ pour λ_2.

Tableau 3
Estimation des paramètres de l'estimation par petits domaines des DCMH sous le modèle de régression temporel et le modèle spatio-temporel

	ρ		σ_v^2		k	
Modèle	Estimation	E.-T.	Estimation	E.-T.	Estimation	E.-T.
Modèle 4	–	–	4 715,64	431,00	–	–
Modèle 5	0,79	0,04	2 163,50	245,50	0,53	0,07

Afin d'évaluer les propriétés des estimateurs sous divers modèles comparativement au modèle le plus général (modèle spatio-temporel), nous avons simulé des données sous le modèle spatio-temporel et avons calculé les EQM réelles des estimations répétées sous chacun des modèles étudiés. Pour ce faire, nous avons exécuté la simulation en prenant les paramètres estimés d'après le modèle spatio-temporel présenté au tableau 2 et avons calculé la moyenne réelle de petit domaine répété $\theta(b)$ pour la b^{e} réplique $(b = 1, 2, \ldots, B)$ ainsi que les observations simulées $y(b)$ pour un grand nombre de répliques. Sur cet ensemble de données simulées, pour chaque réplique, nous avons appliqué divers modèles, y compris le modèle spatio-temporel et avons calculé les estimateurs de la moyenne de petit domaine sous chacun.

Lors de l'ajustement du modèle de régression temporel et du modèle spatio-temporel aux ensembles de données simulés, nous avons exécuté le processus de maximisation itérative avec la contrainte $k <= 1$. Ici, nous avons utilisé $B = 5\,000$ répétitions. Nous pouvons définir l'EQM réelle de l'estimateur pour le i^{e} petit domaine sous un modèle particulier $(k = 2 - 4)$ comme étant

$$\text{EQM}(\theta_i^k) = \frac{1}{B}\sum_{k=1}^{B}[\hat{\theta}_i^k(b) - \theta_i(b)]^2, \quad i = 1, 2, \ldots, m.$$

Nous avons évalué l'efficacité relative des estimateurs sous le modèle spatio-temporel (modèle 5) comparativement aux estimateurs sous les modèles 2 à 4 au moyen du ratio de leurs erreurs quadratiques moyennes (EQMR) donné par

$$\text{EQMR}(k, \text{Temp}) = 100\frac{\sum_{i=1}^{m}\text{EQM}(\hat{\theta}_i^k)}{\sum_{i=1}^{m}\text{EQM}(\hat{\theta}_i^{\text{Temp}})}$$

Tableau 4
MPLNBE moyens pour les DCMH (R), leur e.-t. estimée et leur c.v. sous les modèles de régression, spatial, de régression temporel et spatio-temporel

Modèle	Cycle de la NSSO 50	51	52	53	54	55
	Estimations moyennes par petits domaines					
Modèle 1	276,10	321,26	373,07	408,52	411,25	482,00
Modèle 2	272,87	312,53	354,45	397,52	400,87	471,99
Modèle 3	272,98	313,14	351,51	398,21	400,78	471,09
Modèle 3A	273,56	314,19	352,01	396,40	399,91	471,91
Modèle 4	274,13	305,62	345,54	383,53	399,56	463,32
Modèle 5	273,75	312,21	351,79	391,61	399,50	473,57
	Erreurs-types moyennes (e.-t.)					
Modèle 1	25,09	66,06	64,18	74,19	53,87	45,45
Modèle 2	17,10	33,65	29,09	39,85	32,68	30,59
Modèle 3	16,88	32,84	21,51	39,98	30,87	24,84
Modèle 3A	16,56	31,29	20,79	40,03	30,23	24,37
Modèle 4	19,51	34,91	35,19	37,79	35,14	33,15
Modèle 5	17,18	28,99	28,33	30,02	28,76	28,10
	Coefficients de variation moyens (c.v.) (%)					
Modèle 1	9,09	20,56	17,20	18,16	13,10	9,43
Modèle 2	6,27	10,79	8,21	10,01	8,15	6,48
Modèle 3	6,18	10,49	6,12	10,04	7,70	5,27
Modèle 3A	6,05	9,96	5,91	10,10	7,56	5,17
Modèle 4	7,12	11,42	10,18	9,85	8,79	7,15
Modèle 5	6,28	9,29	8,05	7,67	7,20	5,93

Tableau 5
Efficacité relative [EQMR] en pourcentage des modèles temporels comparativement aux autres modèles pour les DCMH

	Cycle de la NSSO 50	51	52	53	54	55
	Modèle spatio-temporel [Modèle 5]					
Modèle 2	123,63	170,54	193,68	203,55	204,72	169,76
Modèle 3	100,24	133,82	149,70	165,46	165,85	154,23
Modèle 4	125,81	141,50	141,93	137,55	139,11	129,88
	Modèle de régression temporel [Modèle 4]					
Modèle 2	100,71	134,50	156,35	165,30	163,13	152,56

où « Temp » dénote le modèle spatio-temporel et *k*, le modèle 2, 3 ou 4. De même, nous avons déterminé l'efficacité relative du modèle de régression temporel (modèle 4) par rapport au modèle de régression simple (modèle 2) en simulant des données au moyen des estimations des paramètres du tableau 3, sous le modèle de régression temporel. Les résultats sont présentés au tableau 5.

Pour ces paramètres, les résultats confirment la supériorité du modèle spatio-temporel comparativement aux autres. Le modèle de régression temporel s'avère également meilleur que le modèle de régression simple.

5. Conclusion

L'utilisation d'un modèle pour petits domaines, caractéristique du domaine, améliore considérablement les estimations directes par sondage, car elle permet d'exploiter l'autocorrélation spatiale entre les domaines voisins. Cependant, il ne faut appliquer le modèle qu'après avoir déterminé si la corrélation entre les petits domaines en vertu de leurs effets de voisinage est significative. Si la relation entre la variable dépendante et les variables exogènes est faible, le modèle spatial simple avec ordonnée à l'origine seulement peut améliorer tout autant les estimations. Ce modèle ne tire parti que de l'autocorrélation spatiale pour renforcer les estimations par petits domaines et ne requiert pas l'utilisation de variables exogènes. Les modèles spatiaux, grâce à l'utilisation de la matrice de poids *W* appropriée, ou d'une combinaison de matrices *W*, peuvent améliorer considérablement les estimations. La matrice de poids devrait s'appuyer sur des considérations logiques et est parfois utile dans les cas où, pour certaines raisons, on ne dispose pas de variables exogènes fiables. Cet aspect peut aussi être exploité pour obtenir les estimations par petits domaines pour les domaines qui ont été créés/délimités récemment.

Il faut faire attention à l'accroissement de l'erreur quadratique moyenne (EQM) causé par la variabilité due au remplacement des paramètres par leurs estimations. Cet aspect, que reflète l'approximation de deuxième ordre de l'EQM examinée dans le présent article, est la raison pour laquelle, très souvent, le simple modèle spatial (avec ordonnée à l'origine) donne de meilleurs résultats que le modèle spatial comprenant un plus grand nombre de paramètres. L'utilisation de données chronologiques avec des paramètres de régression constants au cours du temps améliore encore davantage les estimations par petits domaines, particulièrement pour les points dans le temps où la EQM des estimations directes par sondage est grande. Les modèles spatio-temporels présentent des avantages par rapport aux modèles temporels qui ne tiennent pas compte des effets spatiaux, grâce à l'inclusion d'une autocorrélation spatiale constante entre les petits domaines. Cependant, pour certains points dans le temps pour lesquels ρ peut être fort différente de celle des autres points, cet avantage n'est pas nécessairement vérifié, parce que les estimations tendent vers la moyenne des cinq cycles. Dans ce cas, on peut choisir un premier point dans le temps approprié pour commencer à tenir compte des effets temporels. Enfin, les variables exogènes *X* et la matrice de poids *W* se complètent par la voie du paramètre de régression β et du paramètre d'autocorrélation ρ, et l'utilisation judicieuse de ces paramètres peut donner lieu à une amélioration importante des estimations par petits domaines.

Remerciements

Les données de niveau unitaire utilisées pour l'étude nous ont été fournies par la National Sample Survey Organisation (NSSO), du ministère des Statistiques et de la Mise en œuvre des programmes aux termes d'une entente de recherche entre IIT Kanpur et la NSSO. La matrice de poids contenant la longueur de la frontière commune pour diverses paires de petits domaines (districts) a été fournie par le centre national d'informatique (NIC) du ministère des Technologies de l'information du gouvernement de l'Inde. Nous tenons à remercier les examinateurs de leurs commentaires constructifs qui nous ont permis d'améliorer beaucoup l'article.

Annexe

Théorème A.1 : Sous les conditions de régularité 1

$$\mathrm{EQM}[\hat{\theta}(\hat{\psi})] = g_1(\psi) + g_2(\psi) + g_3(\psi) + o(m^{-1}). \quad (5.1)$$

Pour prouver le théorème, nous utilisons les résultats suivants bien connus (Srivastawa et Tiwari 1976). Soit $U \sim N(0, \Sigma)$, alors pour les matrices symétriques A, B et C

$$E[U(U^T AU)U^T] = \mathrm{Trace}(A\Sigma)\Sigma + 2\Sigma A\Sigma$$

$$E[U(U^T AU)(U^T BU)U^T] = \mathrm{Trace}(A\Sigma)\mathrm{Trace}(B\Sigma)\Sigma + 2[\mathrm{Trace}(A\Sigma)\Sigma B\Sigma + \mathrm{Trace}(B\Sigma)\Sigma A\Sigma + \mathrm{Trace}(A\Sigma B\Sigma)\Sigma] + 4[\Sigma A\Sigma B\Sigma + \Sigma B\Sigma A\Sigma].$$

Preuve du théorème A.1

Kackar et Harville (1984) ont montré que $\mathrm{EQM}[\hat{\theta}(\hat{\psi})] = \mathrm{EQM}[\hat{\theta}(\psi)] + E[(\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi))(\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi))^T]$. Il est facile de montrer que $\mathrm{EQM}[\hat{\theta}(\psi)] = g_1(\psi) + g_2(\psi)$. Nous devons prouver que $g_3(\psi) = E[(\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi))(\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi))^T] + o(m^{-1})$. Le développement en série de Taylor de $\hat{\theta}(\hat{\psi})$ autour de ψ et l'utilisation de $(\hat{\psi} - \psi) = O_p(m^{-1/2})$

et de $(\partial^2\hat{\theta}(\psi))/(\partial\psi_d\partial\psi_e)|_{\psi=\hat{\psi}^*} = O_p(1)$ quand $\|\hat{\psi}^* - \hat{\psi}\| \leqq \|\hat{\psi} - \psi\|$ nous donnent

$$[\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)] = [(\hat{\psi} - \psi) \otimes I_m]^T \nabla\hat{\theta}(\psi) + O_p(m^{-1}). \quad (5.2)$$

Ici, $\nabla\hat{\theta}(\psi) = (\partial\hat{\theta}(\psi))/(\partial\psi) = [(\partial\hat{\theta}(\psi))/(\partial\rho), (\partial\hat{\theta}(\psi))/(\partial\sigma_v^2)]^T$. En utilisant

$$\frac{\partial\hat{\theta}(\psi)}{\partial\psi_d} = \sum_{\alpha=1}^{p} \frac{\partial\hat{\theta}^*(\beta,\psi)}{\partial\beta_\alpha}\Big|_{\beta=\hat{\beta}(\psi)} \frac{\partial\hat{\beta}(\psi)}{\partial\psi_d} + \frac{\partial\hat{\theta}^*(\beta,\psi)}{\partial\psi_d}\Big|_{\beta=\hat{\beta}(\psi)}$$
$d = 1, 2$

où $\hat{\theta}^*(\beta,\psi) = X\beta(\psi) + \Lambda(\psi)[y - X\beta(\psi)]$, et le fait que $(\partial\hat{\beta}_\alpha(\psi))/(\partial\psi_d) = O_p(m^{-1/2})$ (Cox et Reid 1987), nous tirons de ce qui précède

$$[\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)] = [(\hat{\psi} - \psi)^T \otimes I_m]\nabla\hat{\theta}^*(\psi) + O_p(m^{-1}) \quad (5.3)$$

où $\nabla\hat{\theta}^*(\psi) = \left[\frac{\partial\hat{\theta}^*(\beta,\psi)}{\partial\rho}, \frac{\partial\hat{\theta}^*(\beta,\psi)}{\partial\sigma_v^2}\right]^T\Big|_{\beta=\hat{\beta}(\psi)}$
$= \mathrm{L}(\psi)[y - X\hat{\beta}(\psi)].$

En utilisant les conditions de régularité 1 et le fait que $\hat{\beta}(\psi) - \beta = O_p(m^{-1/2})$, nous avons

$$[\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)]$$
$$= [(\hat{\psi} - \psi) \otimes I_m]^T L(\psi)[y - X\hat{\beta}(\psi)] + O_p(m^{-1})$$
$$= \sum_{d=1}^{2} (\hat{\psi}_d - \psi_d) L_d(\psi)[y - X\hat{\beta}(\psi)] + O_p(m^{-1}).$$

En outre, si nous utilisons le développement en série de Taylor de la vraisemblance $S(\hat{\eta}) = 0$ autour de ψ, où

$$S(\eta) = [S_\beta^T(\eta), S_\psi^T(\eta)]^T, S_\beta^T(\eta) = \underset{1\leq\alpha\leq p}{\mathrm{Col}}\left[\frac{\partial\ell}{\partial\beta_\alpha}\right]$$

et l'orthogonalité de β et ψ, il s'ensuit que

$$(\hat{\psi} - \psi) = I_\psi^{-1}(\psi)S_\psi(\eta) + O_p(m^{-1}).$$

En écrivant

$$S_\psi(\psi) = \underset{1\leq d\leq 2}{\mathrm{Col}}[S_d(\psi)] = [S_\rho(\psi), S_{\sigma_v^2}(\psi)]^T,$$

$$S_d(\psi) = \frac{\partial\ell}{\partial\psi_d} = -\frac{1}{2}\mathrm{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\right] + \frac{1}{2}[u^T B_d(\psi)u],$$

$$B_d(\psi) = \Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\Sigma^{-1}, u = y - X\beta(\psi) \text{ et}$$

$$I_{de}(\psi) = \frac{1}{2}\mathrm{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_e}\right]$$

nous obtenons

$$[\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)] = L^T(\psi)[I_\psi^{-1}(\psi) \otimes I_m][\mathrm{S}_\psi(\psi) \otimes u]$$

et, donc, l'expression

$$[\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)][\hat{\theta}(\hat{\psi}) - \hat{\theta}(\psi)]^T \text{ jusqu'à l'ordre } o(m^{-1})$$
$$= L^T(\psi)[I_\psi^{-1}(\psi) \otimes I_m] \underset{1\leq d\leq 2}{\mathrm{Col}}[uS_d(\psi)] \underset{1\leq e\leq 2}{\mathrm{Concat}}[S_e(\psi)u^T]$$
$$[I_\psi^{-1}(\psi) \otimes I_m]L(\psi)$$
$$= L^T(\psi)[I_\psi^{-1}(\psi) \otimes I_m] \underset{1\leq d\leq 2}{\mathrm{Col}} \underset{1\leq e\leq 2}{\mathrm{Concat}}[uS_d(\psi)S_e(\psi)u^T]$$
$$[I_\psi^{-1}(\psi) \otimes I_m]L(\psi). \quad (5.4)$$

Maintenant, nous pouvons écrire la vraisemblance et sa dérivée sous la forme

$$\ell = \log L = \text{const.} - \frac{1}{2}\log[|\Sigma|] - \frac{1}{2}u^T\Sigma^{-1}u$$

$$\frac{\partial\ell}{\partial\psi_d} = -\frac{1}{2}\mathrm{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\right] + \frac{1}{2}u^T B_d(\psi)u,$$

$$B_d(\psi) = \Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\Sigma^{-1}$$

$$E\left[-\frac{\partial^2\ell}{\partial\psi_d\partial\psi_e}\right] = \frac{1}{2}\mathrm{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_e}\right] = I_{de}(\psi)$$

où la matrice d'information $I_\psi(\psi) \equiv I_{de}(\psi)$.

L'espérance d'un élément type des termes les plus à l'intérieur dans l'expression (5.4) devient

$$E[uS_d(\psi)S_e(\psi)u^T] =$$
$$E\begin{bmatrix} u[u^T B_d(\psi)u][u^T B_e(\psi)u]u^T \\ -u\,\mathrm{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\right][u^T B_e(\psi)u]u^T \\ -u[u^T B_d(\psi)u]\mathrm{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_e}\right]u^T \\ +u\,\mathrm{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\right]\mathrm{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_e}\right]u^T \end{bmatrix}$$

et, en appliquant les résultats de Srivastawa et Tiwari (1976), elle devient

$$E[uS_d(\psi)S_e(\psi)u^T] =$$
$$\frac{1}{2}\mathrm{Trace}\left[\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_e}\right]\Sigma + 2\left[\frac{\partial\Sigma}{\partial\psi_d}\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_e}\right].$$

En introduisant cette expression par substitution dans (5.4) et étant donné que la deuxième expression est d'ordre $O(m^{-1})$, nous pouvons obtenir l'expression suivante jusqu'à l'ordre $o(m^{-1})$

$$[\hat{\theta}(\hat{\psi})-\hat{\theta}(\psi)][\hat{\theta}(\hat{\psi})-\hat{\theta}(\psi)]^T$$
$$= L^T(\psi)[I_\psi^{-1}(\psi)\otimes I_m]\underset{1\le d\le 2}{\text{Col}}\ \underset{1\le e\le 2}{\text{Concat}}[I_{de}(\psi)\Sigma]$$
$$[I_\psi^{-1}(\psi)\otimes I_m]L(\psi)$$
$$= L^T(\psi)[I_\psi^{-1}(\psi)\otimes I_m][I_\psi(\psi)\otimes\Sigma][I_\psi^{-1}(\psi)\otimes I_m]L(\psi)$$
$$= L^T(\psi)[I_\psi^{-1}(\psi)\otimes\Sigma]L(\psi).$$

Théorème A.2 : Sous les conditions de régularité 1

$$E[g_1(\hat{\psi})+g_3(\hat{\psi})-g_4(\hat{\psi})-g_5(\hat{\psi})]=g_1(\psi)+o(m^{-1}), \quad (5.5)$$

$$E[g_2(\hat{\psi})]=g_2(\psi)+o(m^{-1}),$$
$$E[g_3(\hat{\psi})]=g_3(\psi)+o(m^{-1}) \quad (5.6)$$

et $$E[g_5(\hat{\psi})]=g_5(\psi)+o(m^{-1}). \quad (5.7)$$

Preuve du théorème A.2

Le développement en série de Taylor de $g_1(\hat{\psi})$ autour de ψ et l'utilisation de $\hat{\psi}-\psi=O_p(m^{-1/2})$, quand $\|\hat{\psi}^*-\psi\|<=\|\hat{\psi}-\psi\|$, nous donnent

$$g_1(\hat{\psi})=g_1(\psi)+[(\hat{\psi})-(\psi)^T\otimes I_m]\nabla g_1(\psi)$$
$$+\frac{1}{2}[(\hat{\psi}-\psi)^T\otimes I_m]\nabla^2 g_1(\psi)[(\hat{\psi}-\psi)\otimes I_m]$$
$$+o_p(m^{-1})$$
$$\nabla g_1(\psi)=\left[\frac{\partial g_1(\psi)}{\partial\rho}\ \frac{\partial g_1(\psi)}{\partial\sigma_v^2}\right]^T,$$
$$\nabla^2 g_1(\psi)=\underset{1\le d\le 2}{\text{Col}}\left[\underset{1\le e\le 2}{\text{Concat}}\frac{\partial^2 g_1(\psi)}{\partial\psi_d\partial\psi_e}\right]$$
$$\frac{\partial g_1(\psi)}{\partial\psi_d}=R\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\Sigma^{-1}R$$
$$\frac{\partial^2 g_1(\psi)}{\partial\psi_d\partial\psi_e}=-2R\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_d}\Sigma^{-1}\frac{\partial\Sigma}{\partial\psi_e}\Sigma^{-1}R$$
$$+R\Sigma^{-1}\frac{\partial^2\Sigma}{\partial\psi_d\partial\psi_e}\Sigma^{-1}R.$$

En nous fondant sur le fait que $\Sigma(\psi)$ et ses dérivées sont symétriques, nous obtenons, pour le deuxième terme de l'expression, la forme

$$[(\hat{\psi}-\psi)^T\otimes I_m]\nabla^2 g_1(\psi)[(\hat{\psi}-\psi)\otimes I_m]$$
$$=-L^T(\psi)[I_\psi^{-1}(\psi)\otimes\Sigma]L(\psi)$$
$$+\frac{1}{2}\underset{m}{\text{Trace}}\left[[I_2\otimes(R\Sigma^{-1})]\frac{\partial^2\Sigma}{\partial\psi\partial\psi^T}[I_\psi^{-1}(\psi)\otimes(\Sigma^{-1}R)]\right]$$
$$=-g_3(\psi)+g_5(\psi)$$

où $I_\psi^{-1}(\psi)=\text{Var}(\psi)$ est une matrice d'information, la variance asymptotique de ψ. Le premier terme de l'expression $[(\hat{\psi}-\psi)^T\otimes I_m]\nabla g_1(\psi)$ se réduit à $g_4(\psi)$ parce que $E(\hat{\psi}-\psi)=b_{\hat{\psi}}(\psi)$ jusqu'à l'ordre $o(m^{-1})$ (Peers et Iqbal 1985).

La deuxième partie du théorème découle du développement en série de Taylor de $g_2(\hat{\psi}), g_3(\hat{\psi})$ et $g_5(\hat{\psi})$, chacune autour de ψ et de l'utilisation de $\hat{\psi}-\psi=O_p(m^{-1/2})$ et $(\partial^2 g_2(\psi))/(\partial\psi_d\partial\psi_e)|_{\psi=\hat{\psi}}=O_p(m^{-1})$, $(\partial^2 g_3(\psi))/(\partial\psi_d\partial\psi_e)|_{\psi=\hat{\psi}}=O_p(m^{-1})$, $(\partial^2 g_3(\psi))/(\partial\psi_d\partial\psi_e)|_{\psi=\hat{\psi}}=O_p(m^{-1})$ et $(\partial^2 g_5(\psi))/(\partial\psi_d\partial\psi_e)|_{\psi=\hat{\psi}}=O_p(m^{-1})$, respectivement, où $\|\hat{\psi}^*-\psi\|\le\|\hat{\psi}-\psi\|$.

Théorème A.3 : Sous les conditions de régularité 2

$$\text{EQM}([\hat{\theta}_t(\hat{\psi})]=g_{12t}(\psi)+g_{3t}(\psi)+o(m^{-1}). \quad (5.8)$$

Preuve du théorème A.3

La preuve est essentiellement du même type que celle du théorème A.1 en utilisant les résultats de Srivastawa et Tiwari (1976) qui y sont mentionnés

$$\text{EQM}([\hat{\theta}_t(\hat{\psi})]$$
$$=\text{EQM}([\hat{\theta}_t(\psi)]+E([\theta_t(\psi)-\theta_t)(\theta_t(\psi)-\theta_t)^T]$$
$$=g_{12t}(\psi)+E[(\theta_t(\psi)-\theta_t)(\theta_t(\psi)-\theta_t)^T]. \quad (5.9)$$

Par développement en série de Taylor de $\theta_t(\psi)$ autour de ψ et en utilisant $(\hat{\psi}-\psi)=O_p(m^{-1/2})$ et $(\partial^2\hat{\theta}(\psi))/(\partial\psi_d\partial\psi_e)|_{\psi=\hat{\psi}^*}=O_p(1)$ quand $\|\hat{\psi}^*-\psi\|\le\|\hat{\psi}-\psi\|$, nous obtenons

$$[\hat{\theta}_t(\hat{\psi})-\hat{\theta}_t(\psi)]$$
$$=[(\hat{\psi}-\psi)\otimes I_m]^T\nabla\hat{\theta}_t(\psi)+O_p(m^{-1})$$
$$=\sum_{d=1}^{3}[(\hat{\psi}_d-\psi_d)L_{td}(\psi)e_t(\psi)]+O_p(m^{-1}). \quad (5.10)$$

En outre, du développement en série de Taylor de l'équation de vraisemblance $S(\hat{\eta})=0$ et de l'orthogonalité de β et ψ, il découle que

$$(\hat{\psi}-\psi)=I_\psi^{-1}(\psi)S(\psi)+O_p(m^{-1}). \quad (5.11)$$

En introduisant par substitution l'expression de $(\hat{\psi}-\psi)$ dans l'équation (5.10), nous obtenons, jusqu'à l'ordre $o(m^{-1})$,

$$[\hat{\theta}_t(\hat{\psi})-\hat{\theta}_t(\psi)]=L_t^T(\psi)[I_\psi^{-1}(\psi)\otimes I_m][S_\psi(\psi)\otimes e_t] \quad (5.12)$$

et

$$[(\hat{\theta}_t(\hat{\psi})-\hat{\theta}_t(\psi))(\hat{\theta}_t(\hat{\psi})-\hat{\theta}_t(\psi))^T]$$
$$=L^T(\psi)[I_\psi^{-1}(\psi)\otimes I_m]\underset{1\le d\le 3}{\text{Col}}\ \underset{1\le e\le 3}{\text{Concat}}$$
$$[e_t S_d(\psi)S_e(\psi)e_t^T][I_\psi^{-1}(\psi)\otimes I_m]L(\psi) \quad (5.13)$$

où

$$S_\psi(\psi) = \underset{1\le d\le 3}{\text{Col}}\,[S_d(\psi)], \quad S_d(\psi) = \frac{\partial \ell}{\partial \psi_d}.$$

En utilisant l'expression pour les dérivés de la vraisemblance, nous obtenons

$$S_d(\psi) = \frac{1}{2}\left[\begin{array}{l}\text{Trace}[C_{1d}(\psi)] - \sum_{t=1}^{T}\text{Trace}\left[H_t^{-1}\frac{\partial H_t}{\partial \psi_d}\right] \\ + \sum_{t=1}^{T}[e_t^T B_{td}(\psi) e_t]\end{array}\right]$$

$$-\left[e_t^T H_t^{-1}\frac{\partial e_t}{\partial \psi_d}\right] C_{1d}(\psi) = \left[(X_1^T H_1^{-1} X_1)^{-1} X_1^T H_1^{-1}\frac{\partial H_1}{\partial \psi_d} H_1^{-1} X_1\right],$$

$$B_{td}(\psi) = H_t^{-1}\frac{\partial H_t}{\partial \psi_d} H_t^{-1}.$$

En tenant compte de ce que $e_t \sim N(0, H_t), \text{Corr}(e_i, e_j) = 0$ pour $i \neq j, \text{Corr}(e_i, (\partial e_i)/(\partial \psi_d)) = 0$ et $\text{Corr}(e_i, (\partial^2 e_i)/\text{Corr}(e_i, (\partial^2 e_i)/(\partial \psi_d \partial \psi_e)) = 0$ en raison du fait que $(\partial e_i)/(\partial \psi_d) = (\partial(y_i - U_t \hat{\alpha}_{t|t-1}))/(\partial \psi_d)$ est une fonction linéaire de $(y_1, y_2, \ldots, y_{t-1})$ et n'est donc pas corrélée à e_t, nous obtenons l'espérance des termes les plus intérieurs de l'expression (5.13) sous la forme

$$E[e_t S_d(\psi) S_e(\psi) E_t^T] = K_{de}(\psi) H_t + 2\left[\frac{\partial H_t}{\partial \psi_d} H_t^{-1}\frac{\partial H_t}{\partial \psi_d}\right]$$
$$+ \frac{1}{2}\left[\text{Trace}[B_{td}(\psi)]\frac{\partial H_t}{\partial \psi_e} + \text{Trace}[B_{te}(\psi)]\frac{\partial H_t}{\partial \psi_d}\right]$$
$$+ \frac{1}{4}\text{Trace}[B_{td}(\psi)]\,\text{Trace}[B_{te}(\psi)] H_t$$

où

$$K_{de}(\psi) = \frac{1}{2}\sum_{t=1}^{T}\text{Trace}\left[H_t^{-1}\frac{\partial H_t}{\partial \psi_d} H_t^{-1}\frac{\partial H_t}{\partial \psi_e}\right].$$

Les trois termes médians de l'expression étant d'ordre $O(1)$, ce qui, conjugué à $I_\psi^{-1}(\psi)$ dans l'expression donnée ci-dessous, les rend d'ordre $o(m^{-1})$,

$$E[(\hat{\theta}_t(\hat{\psi}) - \hat{\theta}_t(\psi))(\hat{\theta}_t(\hat{\psi}) - \hat{\theta}_t(\psi))^T] = g_{3t}(\psi)$$
$$= L^T(\psi)[I_\psi^{-1}(\psi) \otimes I_m][K_\psi(\psi) \otimes H_t]$$
$$[I_\psi^{-1}(\psi) \otimes I_m] L(\psi) + o(m^{-1})$$
$$= L^T(\psi)[I_\psi^{-1}(\psi) K_\psi(\psi) I_\psi^{-1}(\psi) \otimes H_t] L(\psi) + o(m^{-1}).$$

Théorème A.4 : Sous les conditions de régularité 2

$$E[g_{12t}(\hat{\psi}) + g_{3t}(\hat{\psi}) + g_{31t}(\hat{\psi}) - g_{4t}(\hat{\psi}) - g_{5t}(\hat{\psi})]$$
$$= g_{12t}(\psi) + o(m^{-1})$$
$$E[g_{3t}(\hat{\psi})] = g_{3t}(\psi) + o(m^{-1})$$

et

$$E[g_{5t}(\hat{\psi})] = g_{5t}(\psi) + o(m^{-1}).$$

Preuve du théorème A.4

La preuve est essentiellement du même type que celle du théorème A.2. Par développement en série de Taylor de $g_{12t}(\hat{\psi})$ autour de ψ, nous obtenons

$$g_{12t}(\hat{\psi}) = g_{12t}(\psi) + [(\hat{\psi} - \psi) \otimes I_m]^T \nabla g_{12t}(\psi)$$
$$+ \frac{1}{2}[(\hat{\psi} - \psi)^T \otimes I_m] \nabla^2 g_{12t}(\psi)[(\hat{\psi} - \psi) \otimes I_m]$$
$$+ o_p(m^{-1})$$

$$\nabla g_{12t}(\psi) = \underset{1\le d\le 3}{\text{Col}}\,[\nabla g_{12td}(\psi)], \nabla g_{12td}(\psi) = \frac{\partial g_{12t}(\psi)}{\partial \psi_d}$$

$$\nabla^2 g_{12t}(\psi) = \underset{1\le d\le 3}{\text{Col}}\left[\underset{1\le e\le 3}{\text{Concat}}\frac{\partial^2 g_{12t}(\psi)}{\partial \psi_d \partial \psi_e}\right]$$

$$\frac{\partial g_{12t}(\psi)}{\partial \psi_d} = R\Sigma^{-1}\frac{\partial R}{\partial \psi_d}\Sigma^{-1}R$$

$$\frac{\partial^2 g_{12t}(\psi)}{\partial \psi_d \partial \psi_e} = -2R\Sigma^{-1}\frac{\partial \Sigma}{\partial \psi_d}\Sigma^{-1}\frac{\partial \Sigma}{\partial \psi_e}\Sigma^{-1}R$$
$$+ R\Sigma^{-1}\frac{\partial^2 \Sigma}{\partial \psi_d \partial \psi_e}\Sigma^{-1}R.$$

Compte tenu du fait que $\Sigma(\psi)$ et ses dérivées sont symétriques, nous obtenons le deuxième terme de l'expression sous la forme

$$[(\hat{\psi} - \psi)^T \otimes I_m] \nabla^2 g_{12t}(\psi)[(\hat{\psi} - \psi) \otimes I_m]$$
$$= -L^T(\psi)[I_\psi^{-1}(\psi) \otimes \Sigma] L(\psi)$$
$$+ \frac{1}{2}\underset{m}{\text{Trace}}\left[[I_3 \otimes (R\Sigma^{-1})]\frac{\partial^2 \Sigma}{\partial \psi \partial \psi^T}[I_\psi^{-1}(\psi) \otimes (\Sigma^{-1}R)]\right]$$
$$= -g_{3t}(\psi) = g_{5t}(\psi)$$

où $I_\psi^{-1}(\psi) = \text{Var}(\psi)$ est la variance asymptotique de ψ. Le premier terme de l'expression $[(\hat{\psi} - \psi)^T \otimes I_m] \nabla g_{12t}(\psi)$ se réduit à $g_{4t}(\psi)$, parce que $E(\hat{\psi} - \psi) = b_{\hat{\psi}}(\psi)$ jusqu'à l'ordre $o(m^{-1})$ (Peers et Iqbal 1985).

La deuxième partie du théorème découle du développement en série de Taylor de $g_{3t}(\hat{\psi})$ et de $g_{5t}(\hat{\psi})$ autour de ψ et de l'utilisation de $\hat{\psi} - \psi = O_p(m^{-1/2})$ $(\partial^2 g_{3t}(\psi))/(\partial \psi_d \partial \psi_e)|_{\psi=\hat{\psi}} = O_p(m^{-1})$ et de $(\partial^2 g_{5t}(\psi))/(\partial \psi_d \partial \psi_e)|_{\psi=\hat{\psi}} = O_p(m^{-1})$, respectivement, où $\|\hat{\psi}^* - \hat{\psi}\| \le \|\hat{\psi} - \psi\|$.

Bibliographie

Cliff, A.D., et Ord, J.K. (1981). *Spatial Processes, Models and Applications*. Pion Limited, London.

Cox, D.R., et Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society*, Series B, 49, 1-39 (avec discussion).

Cressie, N. (1990). Small-Area prediction of undercount using the general linear model. *Proceedings of Statistics Canada Symposium 90, Measurement and Improvement of Data Quality*, October 1990.

Cressie, N. (1992). Estimation du maximum de vraisemblance avec contrainte (MVC) dans le lissage des taux de sous-dénombrement du recensement selon l'approche empirique de Bayes. *Techniques d'enquêtes*, 18, 83-103.

Datta, G.S., et Lahiri, P. (2000). A unified measure of uncertainty of estimates for best linear unbiased predictors in small area estimation problem. *Statistica Sinica*, 10, 613-627.

Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedure to census data. *Journal of the American Statistics Association*, 74, 267-277.

Ghosh, M., et Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, 423-447.

Kackar, R.N., et Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistics Association*, 79, 853-862.

National Sample Survey Organisation, Department of Statistics, Govt. of India (1997). Rapport No. 404(50/1.0/1), Consumption of some important commodities in India. *NSS* 50th *Round, Juillet 1993-Juin 1994*.

National Sample Survey Organisation, Department of Statistics, Govt. of India (1998). Rapport No. 436(51/1.0/1), Household consumption expenditure and employment situation in India. *NSS* 51st *Round, Juillet 1994-Juin 1995*.

National Sample Survey Organisation, Department of Statistics, Govt. of India (1998). Rapport No. 440(52/1.0/1), Household consumption expenditure and employment situation in India, *NSS* 52nd *Round, Juillet 1995-Juin 1996*.

National Sample Survey Organisation, Department of Statistics, Govt. of India (1998). Rapport No. 442(53/1.0/1), Household consumption expenditure and employment situation in India. *NSS* 53rd *Round, Janvier-Decembre 1997*.

National Sample Survey Organisation, Department of Statistics, Govt. of India (1999). Rapport No. 448(54/1.0/1), Household consumption expenditure and employment situation in India. *NSS* 54th *Round, Janvier-Juin 1998.*

National Sample Survey Organisation, Ministry of Statistics and Programme Implementaion, Govt. of India (2001). Rapport No. 472(55/1.0/1), Differences in level of consumption among socio-economic groups. *NSS* 55th *Round, Juillet 1999-Juin 2000.*

Peers, H.W., et Iqbal, M. (1985). Asymptotic expansions for confidence limits in the presence of nuisance parameters, with applications. *Journal of the Royal Statistical Society*, Series B, 47, 547-554.

Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of mean square error of small area estimates. *Journal of the American Statistics Association*, 85, 163-171.

Rao, C.R., et Toutenbourg, Heldge (1999). *Linear Models*: *Least Squares and Alternatives*. Second Edition, New York: Springer.

Rao, J.N.K. (1999). Quelques progress récents concernant l'estimation régionale fondée sur un modèle. *Techniques d'enquêtes*, 25, 199-212.

Ripley, B.D. (1981). *Spatial Statistics.* New York: John Wiley & Sons, Inc.

Sallas W.M., et Harville D.A. (1994). Noninformative priors and restricted likelihood estimation in the Kalman filter. *Bayesian Analysis of Time Series and Dynamic Models*, (Éd. James C. Spall). New York: Marcel Dekker Inc., 477-508.

Singh, A.C., Mantel, H.J. et Thomas B.W. (1994). MPLSE à données chronologiques pour petites régions évalués à l'aide de données d'enquête. *Techniques d'enquêtes*, 20, 35-46.

Singh, A.C., Stukel, D. et Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society*, Series B, 377-396.

Srivastava, V.K., et Tiwari, R. (1976). Evaluation of expectations of products of stochastic matrices. *Scandinavian Journal of Statistics*, 3, 135-138.

Techniques d'enquête, décembre 2005
Vol. 31, N° 2, pp. 215-231
Statistique Canada, N° 12-001-XPB au catalogue

Méthodes de modélisation et d'estimation de la taille du ménage en présence de non-réponse non ignorable appliquées à l'Enquête sur les dépenses de consommation de la Norvège

Liv Belsby, Jan Bjørnstad et Li-Chun Zhang [1]

Résumé

Nous considérons le problème de l'estimation, en présence de non-réponse non ignorable importante, du nombre de ménages privés de diverses tailles et du nombre total de ménages en Norvège. L'approche est fondée sur un modèle de population pour la taille du ménage, sachant la taille enregistrée de la famille. Nous tenons compte du biais de non-réponse éventuel en modélisant le mécanisme de réponse sachant la taille du ménage. Nous évaluons divers modèles, ainsi qu'un estimateur du maximum de vraisemblance et une poststratification fondée sur l'imputation. Nous comparons les résultats à ceux d'une poststratification pure avec la taille enregistrée de la famille comme variable de stratification et des méthodes d'estimation employées pour la production de statistiques officielles d'après l'Enquête sur les dépenses de consommation de la Norvège. L'étude indique que la modélisation de la réponse, la poststratification et l'imputation sont des éléments importants d'une approche satisfaisante.

Mots clés : Taille du ménage; non-réponse; imputation; poststratification.

1. Introduction

La présente étude a été motivée par le taux élevé de non-réponse à l'Enquête sur les dépenses de consommation (EDC) de la Norvège réalisée auprès des ménages privés, qui était de 32 % pour l'enquête de 1992. La non-réponse comprend les impossibilités de prendre contact et les refus de participer. Nous nous concentrons sur le problème de la non-réponse non ignorable qui se pose lorsqu'on estime le nombre de ménages de diverses tailles et le nombre total de ménages.

Nous considérerons une approche entièrement fondée sur un modèle, à savoir la modélisation et l'estimation de la distribution de la taille des ménages sachant la taille enregistrée de la famille et du mécanisme de réponse sachant la taille du ménage. Ce modèle tient compte du fait que le mécanisme de non-réponse pourrait être non ignorable, en ce sens que la probabilité de réponse peut dépendre de la taille du ménage. Le modèle de réponse sert à corriger pour la non-réponse. Les approches fondées sur un modèle avec non-réponse incluse, parfois appelées approches prédictives, ont été considérées, entre autres, par Little (1982), Greenlees, Reece et Zieschang (1982), Baker et Laird (1988), Bjørnstad et Walsøe (1991), Bjørnstad et Skjold (1992), ainsi que Forster et Smith (1998).

Pour divers modèles de taille du ménage et de réponse, nous examinons principalement deux approches fondées sur un modèle, c'est-à-dire un estimateur du maximum de vraisemblance et la poststratification basée sur l'imputation en fonction de la taille enregistrée de la famille. Nous comparons ces méthodes à la poststratification pure et aux méthodes utilisées à l'heure actuelle pour l'EDC.

La grande question ici est de comparer des modèles et des méthodes dont le problème principal est le biais d'estimation. En outre, nous estimons par une méthode bootstrap les erreurs-types des estimations et des différences entre les estimations, sachant les tailles des strates a posteriori déterminées d'après la taille de la famille. En plus d'évaluer l'incertitude statistique des estimateurs, ceci nous permet de déterminer dans quelle mesure les différences entre les estimateurs proposés sont attribuables à l'erreur d'échantillonnage, au biais de non-réponse ou aux deux. Cependant, dans cette évaluation, nous gardons à l'esprit la citation qui suit, tirée de Little et Rubin (1987, page 67) : [traduction] « Il importe d'insister sur le fait que, dans de nombreuses applications, le problème du biais de non-réponse est souvent plus crucial que celui de la variance. En fait, d'aucuns ont soutenu que fournir une estimation valide de la variance d'échantillonnage est pire que ne pas fournir d'estimation si l'estimateur présente un biais important, qui domine l'erreur quadratique moyenne. »

À la section 2, nous décrivons la structure des données et le plan de sondage de l'EDC, et à la section 3, nous examinons les problèmes de modélisation. À la section 3.1, nous présentons les divers modèles de taille du ménage et de réponse à prendre en considération pour l'EDC de 1992, à la section 3.2, nous décrivons la méthode du maximum de vraisemblance pour l'estimation des paramètres et à la

1. Liv Belsby, Statistique Norvège, Division des méthodes statistiques et des normes, C.P. 8131 Dep., N-0033 Oslo. Courriel : lbe@ssb.no; Jan F. Bjørnstad, Statistique Norvège, Division des méthodes statistiques et des normes, C.P. 8131 Dep., N-0033 Oslo. Courriel : jab@ssb.no et Li-Chun Zhang, Statistique Norvège, Division des méthodes statistiques et des normes, C.P. 8131 Dep., N-0033 Oslo. Courriel : lcz@ssb.no.

section 3.3, nous évaluons les modèles. Le meilleur ajustement des modèles considérés s'obtient avec un modèle de groupe de tailles de la famille pour la taille du ménage et un lien logistique pour la probabilité de réponse en utilisant la taille du ménage comme variable nominale. À la section 3.4, nous donnons les distributions estimées de la taille du ménage pour diverses tailles de famille et les probabilités de réponse estimées pour diverses tailles de ménage.

À la section 4, nous examinons l'estimation fondée sur un modèle, la méthode d'imputation, les estimateurs basés sur l'imputation et la méthode d'estimation de la variance. Nous montrons que, pour le modèle choisi pour la taille du ménage à la section 3.3, l'estimateur du maximum de vraisemblance et l'estimateur poststratifié basé sur l'imputation sont identiques.

À la section 5, nous abordons l'objectif principal, c'est-à-dire l'estimation des nombres totaux de ménages de diverses tailles d'après l'EDC de 1992 en utilisant les estimateurs décrits à la section 4. Le modèle produisant le meilleur ajustement semble donner de bons résultats pour notre problème d'estimation. Nous concluons que la poststratification, la modélisation de la réponse et l'imputation sont des éléments essentiels à l'élaboration d'une méthode satisfaisante.

2. Enquête sur les dépenses de consommation de la Norvège

Les totaux de population selon la catégorie de taille du ménage fournissent des nombres plus corrects de logements que les totaux pour les catégories de taille de la famille déterminées d'après le Registre norvégien des familles. En outre, les autorités chargées de l'évaluation des interventions publiques dans le domaine de la construction de logements se basent sur le nombre estimé de ménages. Par conséquent, estimer les totaux selon la taille du ménage est une question importante en planification sociale. L'estimation est invariablement affectée par la non-réponse non ignorable, quel que soit le genre d'enquête utilisé, de sorte qu'il s'agit d'une bonne illustration de la façon de traiter le biais de non-réponse. Nous fonderons notre estimation sur les données de l'Enquête sur les dépenses de consommation (EDC) de la Norvège, pour laquelle il est important d'obtenir des renseignements sur la composition des ménages, puisque la taille du ménage influence la consommation.

L'EDC réelle, c'est-à-dire l'enquête sur les variables de dépenses, est réalisée auprès d'un échantillon de ménages privés représentatifs de l'ensemble des ménages privés de la Norvège. Cet échantillon est obtenu en sélectionnant un échantillon de personnes et en incluant tous les membres des ménages auxquels elles appartiennent. Les personnes de plus de 80 ans sont exclues, car elles vivent souvent en établissement. Aux fins de l'étude, les unités d'intérêt de l'enquête sont les *personnes* de 16 à 80 ans vivant dans les ménages privés et la variable d'intérêt est la taille du *ménage* auquel la personne appartient, qui est observée uniquement dans l'échantillon de personnes sélectionnées répondantes.

Le plan de sondage est un plan d'échantillonnage de personnes autopondéré à trois degrés. Autrement dit, chaque personne faisant partie de la population a la même probabilité d'inclusion dans l'échantillon total. Aux deux premiers degrés, des régions géographiques sont sélectionnées de façon stratifiée, tandis qu'au troisième, des personnes sont sélectionnées aléatoirement à partir des régions géographiques sélectionnées. Au premier degré, les unités primaires d'échantillonnage (UPE) sont les municipalités de la Norvège. Celles comptant moins de 3 000 habitants sont regroupées, de sorte que chaque UPE soit constituée d'au moins 3 000 personnes. Les UPE sont d'abord regroupées en dix régions, puis, dans chaque région, elles sont stratifiées d'après la taille (nombre d'habitants) et le type de municipalité (c'est-à-dire, structure industrielle et centralité). En tout, nous avons obtenu 102 strates. Les villes de plus de 30 000 habitants représentent leur propre strate et, par conséquent, sont sélectionnées avec certitude au premier degré. Pour les autres strates, une UPE est sélectionnée avec probabilité proportionnelle à la taille. Au deuxième degré, les UPE sélectionnée sont réparties en trois régions plus petites (unités secondaires d'échantillonnage, USE) et l'une de celles-ci est sélectionnée au hasard. Enfin, au troisième degré, dans chacune des USE sélectionnée, nous tirons un échantillon aléatoire de personnes. Pour chaque USE sélectionnée, la taille de l'échantillon est déterminée de façon que l'échantillon total résultant de personnes soit autopondéré.

Notre application est fondée sur les données de l'EDC de 1992. L'EDC est une enquête annuelle et, depuis 1992, on utilise un estimateur d'Horvitz-Thompson modifié, comportant une correction pour la non-réponse par estimation des probabilités de réponse sachant la taille du ménage (voir Belsby 1995). Les poids sont égaux à l'inverse de la probabilité de sélection multiplié par la probabilité conditionnelle de réponse sachant que l'unité est sélectionnée. Depuis 1993, la probabilité de réponse est estimée au moyen d'un modèle logistique dont les variables auxiliaires sont le lieu de résidence (région rurale/urbaine) et la taille du ménage. Pour la plupart des non-répondants, on utilise la taille de la famille comme substitut de la taille du ménage.

Par ménage, nous entendons les personnes ayant un logement commun et partageant au moins un repas par jour (c'est-à-dire logeant sous le même toit). Pour une description complète de l'EDC, consulter Statistics Norway (1996). Dans l'EDC, les variables auxiliaires connues pour

l'échantillon total, y compris les non-répondants, sont la taille de la famille, le moment de l'enquête (été/pas l'été) et le lieu de résidence (région urbaine/rurale). Les familles sont inscrites dans le Registre norvégien des familles (RNF) et peuvent différer du ménage auquel appartiennent les membres de la famille, par définition ou à cause de changements qui n'ont pas encore été enregistrés. Donc, la taille enregistrée de la famille selon le RNF diffère, dans une certaine mesure, de la taille du ménage. Au départ, d'après l'expérience des enquêtes antérieures, nous supposons que toutes les variables auxiliaires et la taille du ménage influent sur le taux de réponse.

Le tableau 1 montre les données pour l'EDC de 1992 avec un échantillon total de 1 698 personnes. Les ménages dont la taille est égale ou supérieure à cinq sont regroupés à cause de leur faible fréquence dans l'échantillon de ménages. Nous basons notre modélisation et notre estimation sur deux tableaux correspondants, l'un pour les personnes vivant dans les régions rurales et l'autre pour celles vivant dans les régions urbaines. Les données sont présentées au tableau A1 à l'annexe A1.

Par exemple, le chiffre 48 dans la cellule (1,2) signifie que, des 162 personnes enregistrées comme vivant seules dans l'échantillon répondant, 48 vivent effectivement dans un ménage de deux personnes. Cette situation tient en grande partie au fait que les jeunes gens ont tendance à cohabiter sans être mariés; voir Keilman et Brunborg (1995).

3. Modélisation de la taille du ménage et de la non-réponse

Nous considérons un modèle de population hypothétique pour la taille du ménage, sachant les variables auxiliaires, autrement dit, nous modélisons la probabilité conditionnelle. Pour tenir compte de la non-réponse dans l'analyse statistique, nous devons modéliser le mécanisme de réponse, c'est-à-dire la distribution de la réponse sachant la taille du ménage et les variables auxiliaires. Le mécanisme d'échantillonnage des personnes est ignorable pour l'enquête que nous envisageons, autrement dit, est indépendant du vecteur population de tailles du ménage. Par conséquent, nous faisons l'analyse statistique sachant l'échantillon total, suivant le principe de vraisemblance (voir Bjørnstad 1996), de sorte que les considérations relatives aux probabilités fondées sur le plan d'échantillonnage sont sans importance. Il s'agit de l'approche dite de prédiction. Cependant, lorsque nous évaluons l'incertitude statistique des méthodes d'estimation, nous le faisons sous un angle commun de randomisation décrit à la section 4.3.

Pour l'EDC, le vecteur de variables auxiliaires comprend la taille de la famille, le lieu de résidence subdivisé en régions rurale et urbaine et le moment de la collecte des données.

3.1 Les modèles

Considérons d'abord un modèle simple de la taille du ménage, noté Y. Soit $\mathbf{x}$ toutes les variables auxiliaires. Nous supposons que la taille du ménage dépend uniquement de la taille de la famille x et, par conséquent, qu'il s'agit d'un modèle avec fonction de lien paramétrique contrainte, mais sans autres hypothèses,

$$P(Y_i = y \mid \mathbf{x}_i) = P(Y_i = y \mid x_i) = p_{y,\, x_i}, \qquad (3.1)$$

où

$$\sum_y p_{y,\, x_i} = 1, \text{ pour chaque valeur possible de } x_i.$$

Le modèle (3.1) est souple en ce sens qu'il ne comprend aucune contrainte sur la fonction hypothétique de x_i. L'inconvénient est le nombre élevé de paramètres comparativement à un modèle de type logistique avec fonction de lien linéaire en $\mathbf{x}$ (la fonction reliant $P(Y = y)$ à $\mathbf{x}$). Si l'on ignore la non-réponse, dans le cas de ce modèle, les estimations seront simplement les taux observés.

La taille du ménage définit des catégories ordonnées. Donc, un choix naturel est le modèle logit cumulatif, appelé modèle à odds proportionnels (voir McCullagh et Nelder 1991), en supposant (avec θ_y croissant en y) que

$$P(Y_i \le y \mid \mathbf{x}) = \begin{cases} \dfrac{1}{1 + \exp(-\theta_y + \beta^t \mathbf{x})} & \text{pour } y = 1, 2, 3, 4 \\ 1 & \text{pour } y \ge 5. \end{cases}$$

Tableau 1
Tailles de la famille et du ménage pour l'Enquête sur les dépenses de consommation de la Norvège de 1992

	Taille du ménage							
Taille de la famille	1	2	3	4	≥ 5	Total	Non-réponse	Taux de réponse
1	83	48	20	9	2	162	153	0,514
2	9	177	37	4	3	230	160	0,590
3	10	25	131	40	6	212	91	0,700
4	2	13	37	231	17	300	123	0,709
≥ 5	1	4	4	17	181	207	60	0,775
Total	105	267	229	301	209	1 111	587	0,654

Cependant, un test de qualité de l'ajustement, avec $\mathbf{x}$ représentant la taille de la famille et le lieu de résidence, indique que ce modèle est mal ajusté aux données. Nous choisissons donc de le rejeter.

Nous supposons que la probabilité de non-réponse peut dépendre de la taille du ménage. Par exemple, les ménages d'une personne sont moins susceptibles de répondre que les ménages de plus grande taille, puisqu'il est plus facile de « trouver quelqu'un à la maison » dans le cas de ces derniers. La non-réponse est indiquée par la variable R, où $R_i = 1$ si la personne i répond et 0 autrement. Soit R_s le vecteur de ces indicatrices dans l'échantillon total. D'après Bjørnstad (1996), nous définissons le mécanisme de réponse (MR), c'est-à-dire la loi conditionnelle de R_s sachant les valeurs de $\mathbf{x}$ dans la population et les valeurs de y dans l'échantillon total, comme étant ignorable s'il peut être écarté dans une analyse fondée sur la vraisemblance. Cela signifie que le mécanisme de réponse est ignorable si cette loi conditionnelle de R_s ne dépend pas des valeurs inobservées de y, ce qui coïncide avec la définition utilisée par Little et Rubin (1987, pages 90, 218). Ici, nous supposons que toutes les paires $(Y_i,\ R_i)$ sont indépendantes. Alors, le mécanisme de réponse MR est ignorable si Y_i et R_i sont indépendants. Par conséquent, le mécanisme de réponse non ignorable est équivalent à

$$P(Y_i = y_i \mid \mathbf{x}_i, r_i = 0) \neq P(Y_i = y_i \mid \mathbf{x}_i, r_i = 1)$$

et, alors, tous deux sont différents de $P(Y_i = y_i \mid \mathbf{x}_i)$.

Donc, estimer les paramètres du modèle pour $P(Y = y \mid \mathbf{x})$ en utilisant uniquement l'échantillon de répondants et en ne tenant pas compte du fait que la probabilité de réponse dépend de la taille du ménage produirait fort probablement des estimations biaisées des paramètres inconnus. En outre, l'estimateur par poststratification produirait des estimations biaisées, parce qu'il repose sur l'hypothèse que la distribution de *R* dépend uniquement des variables auxiliaires $\mathbf{x}$. Par exemple, le taux de réponse observé plus faible chez les familles d'une personne indique qu'il pourrait en être de même des ménages d'une personne. Le cas échéant, la probabilité estimée pour un ménage de taille unitaire, fondée uniquement sur les répondants, serait trop faible. La poststratification en fonction de la taille de la famille ne corrigerait fort probablement qu'une partie de ce biais.

Nous supposons que le modèle de la probabilité de réponse, sachant les variables auxiliaires et la taille du ménage y_i, est logistique. Il dépend des variables auxiliaires $\mathbf{z}_i$, qui incluent une partie de $\mathbf{x}_i$, ce qui est exprimé par

RM1$(y, \mathbf{z})$:

$$P(R_i = 1 \mid y_i, \mathbf{z}_i) = \frac{1}{1 + \exp(-\alpha - \gamma y_i - \psi^t \mathbf{z}_i)}. \qquad (3.2)$$

Ici, α et γ sont des paramètres scalaires et ψ est un vecteur. La variable y_i possède un ordre. Motivé par ce fait, et pour éviter d'introduire de nombreux paramètres, nous utilisons y_i dans (3.2) comme variable ordinale plutôt que comme variable de classe. Donc, la fonction logit,

$$\log\{P(R_i = 1 \mid y_i, \mathbf{z}_i) / P(R_i = 0 \mid y_i, \mathbf{z}_i)\} = \alpha + \gamma y_i + \psi^t \mathbf{z}_i,$$

est linéaire en y_i. Pour éviter l'hypothèse de fonction logit linéaire en y_i, nous considérons également un modèle où y_i est une variable nominale, c'est-à-dire,

RM2$(y, \mathbf{z})$: $P(R_i = 1 \mid y_i, \mathbf{z}_i) =$

$$\frac{1}{1 + \exp\begin{pmatrix} -\alpha_0 - \alpha_1 I_1(y_i) - \alpha_2 I_2(y_i) \\ -\alpha_3 I_3(y_i) - \alpha_4 I_4(y_i) - \psi^t \mathbf{z}_i \end{pmatrix}}, \qquad (3.3)$$

où la variable indicatrice $I_y(y_i)$ est égale à 1 si $y_i = y$ et 0 autrement. L'inconvénient de ce modèle est qu'il comprend trois paramètres de plus que le modèle (3.2).

3.2 Estimation des paramètres par la méthode du maximum de vraisemblance

Toutes les personnes sélectionnées dans l'échantillon proviennent de ménages différents (les unités d'échantillonnage en double ont été éliminées), de sorte que le modèle de population suppose que les tailles de ménage Y_i sont statistiquement indépendantes. Pour *cette* variable, l'effet d'interviewer ou de mise en grappes ne joue aucun rôle.

Considérons la fonction de vraisemblance pour estimer les paramètres inconnus, en supposant que toutes les paires $(Y_i,\ R_i)$ sont indépendantes et que le modèle de réponse RM1 est donné par (3.2). Pour simplifier la notation, nous ré-annotons les observations de sorte que les observations 1 à n_r soient les répondants et les observations $n_r + 1$ à n soient les non-répondants. Dans le cas du modèle de réponse RM2, l'expression de la vraisemblance est de la même forme avec (3.2) remplacé par (3.3).

Pour les répondants, soit $L_i = P(Y_i = y_i \cap R_i = 1 \mid \mathbf{x}_i)$. Alors, pour le modèle (3.1)

$$L_i = \frac{1}{1 + \exp(-\alpha - \gamma y_i - \psi^t \mathbf{z}_i)} \cdot p_{y_i, x_i},\ i = 1, ..., n_r \qquad (3.4)$$

Pour les non-répondants, soit $L_i = P(R_i = 0 \mid \mathbf{x}_i)$. Alors

$$L_i = \sum_{y=1}^{5} \frac{1}{1 + \exp(\alpha + \gamma y + \psi^t \mathbf{z}_i)} \cdot p_{y, x_i},\ i = n_r + 1, ..., n. \qquad (3.5)$$

La fonction de vraisemblance pour l'échantillon complet de personnes pour les divers ménages est donnée par

$$L(\theta, \beta, \alpha, \gamma, \psi) = \prod_{i=1}^{n} L_i. \qquad (3.6)$$

Pour $i = 1, ..., n_r$, L_i est donnée par (3.4) et pour $i = n_r + 1, ..., n$, L_i est donnée par (3.5).

Nous obtenons les estimations en maximisant la fonction de vraisemblance (3.6). La maximisation a été faite numériquement à l'aide du logiciel TSP (1991) (voir Hall, Cummins et Schnake 1991). L'algorithme d'optimisation est celui de la méthode type du gradient utilisant les dérivés première et seconde analytiques. Celles-ci sont calculées par le logiciel, ce qui nous évite un travail de programmation considérable. L'ajustement du modèle est fondé sur la statistique du chi carré et sur les valeurs de t, fournies par TSP, où les erreurs-types sont calculées d'après les dérivées secondes analytiques. Les valeurs de t doivent être interprétées avec une certaine prudence, puisque l'absence de biais dans les erreurs-types estimées dépend de la qualité de la spécification du modèle, ainsi que du nombre d'observations comparativement au nombre de paramètres.

3.3 Évaluation des modèles de la taille du ménage et de la réponse

Nous présentons l'évaluation de l'ajustement des modèles au moyen de la statistique de qualité de l'ajustement de Pearson. L'étude de modélisation est fondée sur les données de l'EDC de 1992. Nous considérons que les paramètres sont significatifs si la valeur absolue de t est supérieure à 2. Cependant, nous ne voulons pas d'un modèle trop restrictif et, par conséquent, nous gardons certaines variables même si leur valeur de t absolue est inférieure à 2.

Dans les modèles de réponse RM1 et RM2, nous utilisons la variable $\mathbf{z} = z$, lieu de résidence. Soit $z = 0$ s'il s'agit d'une région rurale et $z = 1$, s'il s'agit d'une région urbaine. On a observé lors de l'EDC de 1986–1988 et de l'EDC de 1992–1994, voir Statistics Norway (1990, 1996), que la non-réponse est plus fréquente durant l'été. Par conséquent, nous avons également inclus le moment de l'enquête dans le modèle, c'est-à-dire une variable indiquant si les données ont été recueillies ou non entre le 21 mai et le 12 août. Toutefois, nous avons constaté que le moment de la réalisation de l'enquête n'était pas significatif, la valeur de *t* étant clairement inférieure à 2. Nous avons également noté que l'effet de la taille de la famille n'était pas significatif. Par contre, si l'on omet la variable de taille du ménage dans le modèle de réponse, alors l'effet de la taille de la famille devient significatif.

Idéalement, nous aimerions examiner la fonction logit empirique de la réponse en fonction de la taille du ménage. Cependant, cette dernière est inconnue pour les non-répondants. Par conséquent, nous représentons graphiquement la fonction logit en fonction de la taille de la famille; voir la figure 1. Si la taille de la famille passe de un à deux, les fonctions pour les familles rurales et urbaines augmentent à peu près parallèlement. Cependant, pour une taille de famille de trois et de quatre, les fonctions logit cessent d'être linéaires et parallèles. Donc, nous pensons que le codage de la taille du ménage sous forme de variable nominale, comme dans le modèle RM2, donnera un meilleure ajustement que la contrainte obligeant les fonctions logit à être parallèles pour les régions rurale et urbaine et linéaires en fonction de la taille du ménage, comme dans le modèle RM1.

Afin de tester la qualité de l'ajustement des modèle, nous considérons la statistique du chi carré de Pearson, sachant les variables auxiliaires x, z. Compte tenu du type rural ou urbain du lieu de résidence et de la taille enregistrée de la famille, il existe six résultats possibles, c'est-à-dire des ménages de taille 1, ..., 5 et la non-réponse. En tout, nous avons dix essais multinomiaux et soixante cellules. Pour les tailles de familles (1,2) et (4,5), les tailles de ménage extrêmes (4,5) et (1,2), respectivement, sont combinées, parce que les tailles attendues sous les modèles sont trop petites. Ceci réduit le nombre de cellules à 52. Le nombre de degrés de liberté (d.d.l.) est calculé comme suit : nombre de cellules – nombre d'essais – moins nombre de paramètres. Pour le modèle (3.1) et RM1(y, z), d.d.l. = 52 – 10 – (20 + 3) = 19, et pour (3.1) et RM2(y, z), d.d.l. = 52 – 10 – (20 + 6) = 16. Pour le modèle (3.1) et RM1(y, z), la statistique χ^2 de Pearson est égale à 26,35 et la valeur p est 0,121. Pour le modèle (3.1) et RM2(y, z), χ^2 est 21,77 et la valeur p est 0,151.

En étudiant les résidus standardisés, (observé-attendu) / $\sqrt{\hat{\mathrm{V}}\mathrm{ar}(\text{observé})}$, nous constatons que la raison principale du meilleur ajustement est que le modèle (3.1) avec RM2(y, z) prédit mieux les dénombrements observés pour la région urbaine où le taux de réponse est le plus faible (voir l'annexe A1). Donc, les données indiquent que le codage de la taille du ménage sous forme de variable nominale, comme dans RM2, améliore l'ajustement comparativement à l'utilisation d'une variable ordinale. Le modèle (3.1), avec la fonction de lien paramétrique contrainte, combiné à RM2 est le meilleur des modèles que nous avons examinés jusqu'à présent.

3.4 Distribution estimée de la taille du ménage et probabilités de réponse

Le tableau 2 donne les estimations pour le modèle de population (3.1) combiné au modèle de réponse logistique RM2 donné par (3.3).

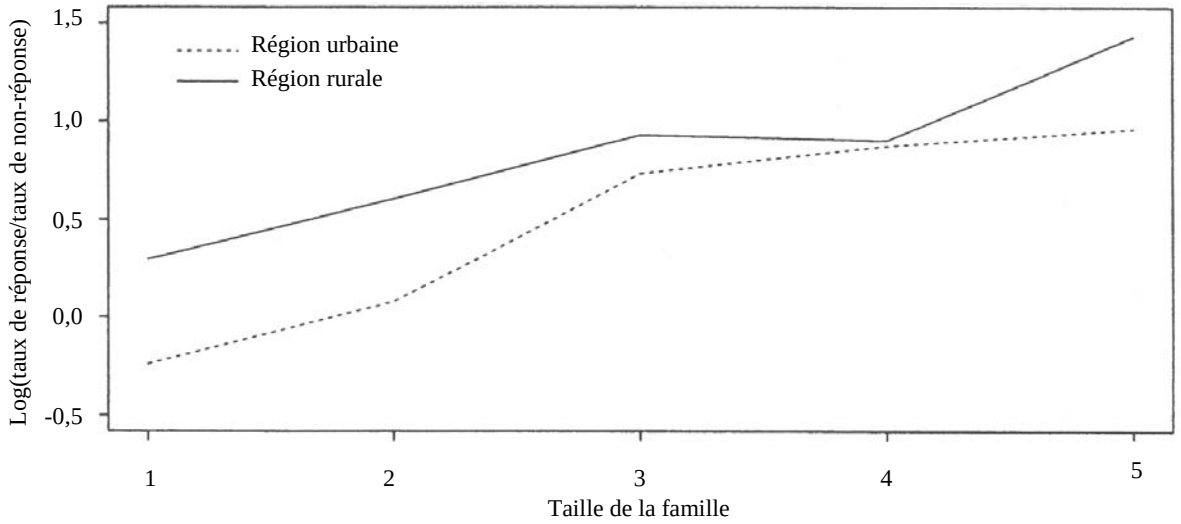

Figure 1. Fonction logit pour le taux de réponse empirique par rapport à une taille de la famille de 1, ..., 5 dans les régions urbaine et rurale, respectivement. Le calcul est fondé sur les répondants et les non-répondants du tableau 1 à l'annexe A1.

Tableau 2
EDC de 1992. Estimation des paramètres, en pourcentage, pour le modèle de population avec fonction de lien paramétrique contrainte, $p_{y,x}$, combinée au modèle de réponse logistique RM2 (y, z). Les estimations pour le modèle de population en ignorant le mécanisme de réponse figure entre parenthèses

	Taille du ménage				
Taille de la famille, x	1	2	3	4	5 ou plus
1	60,01 (51,23)	26,75 (29,63)	8,35 (12,35)	4,09 (5,56)	0,80 (1,23)
2	5,27 (3,91)	79,80 (76,98)	12,48 (16,09)	1,47 (1,74)	0,98 (1,30)
3	7,53 (4,72)	14,45 (11,79)	56,67 (61,79)	18,85 (18,87)	2,50 (2,83)
4	1,06 (0,67)	5,31 (4,33)	11,38 (12,33)	77,20 (77,00)	5,05 (5,67)
5 ou plus	0,84 (0,48)	2,60 (1,93)	1,96 (1,93)	9,05 (8,21)	85,55 (87,44)

Interprétons certaines valeurs obtenues au moyen du modèle fondé sur le ménage. C'est pour les familles à une seule personne que la prise en compte du mécanisme de réponse a l'effet le plus important sur la distribution estimée de la taille du ménage. La probabilité estimative que la taille d'un ménage soit égale à un, sachant que la taille de la famille est de un, est égale à 60,01 %. L'estimation fondée sur l'approche habituelle, en ne tenant pas compte de la non-réponse, est de 51,23 %. Le modèle de réponse « ajuste » le taux observé chez les répondants pour produire une valeur plus élevée, ce qui semble raisonnable, puisque le taux de non-réponse est plus élevé pour les ménages de petite taille. La probabilité estimée que la taille d'un ménage soit de cinq ou plus, sachant que la taille de la famille est égale ou supérieure à cinq, est de 85,55 %, valeur qui diffère peu du taux observé chez les répondants, soit de 87,44 %. Ce résultat indique que, pour une taille de la famille égale ou supérieure à cinq, la distribution de la taille du ménage est à peu près la même chez les répondants que chez les non-répondants.

Le tableau 3 donne les probabilités de réponse estimées en combinant le modèle de réponse RM2 au modèle de population (3.1). En outre, nous présentons les probabilités de réponse estimées basées sur un modèle saturé, avec ajustement parfait, décrit à la section 4.2. Le modèle, défini par (4.9), suppose que la probabilité de réponse des personnes faisant partie d'un ménage de même taille dans une région rurale ou urbaine, respectivement, est identique pour différentes tailles de familles. De surcroît, le modèle de la taille du ménage dépend du lieu de résidence et de la taille de la famille, mais sans contraintes sur la fonction de lien. Nous notons que RM2 (y, z) satisfait (4.9b), mais est plus restrictif. Le modèle (4.9) offre plus de liberté que le modèle (3.1) combiné à RM2 (y, z).

Tableau 3
Probabilité estimée de réponse basée sur le modèle logistique RM2 combiné à (3.1), et sur le modèle saturé (4.9). Les estimations sont données en pourcentage

	Taille du ménage				
Lieu de résidence	1	2	3	4	5 ou plus
	Probabilités de réponse estimées pour le modèle RM2				
Région rurale	47,77	60,90	79,16	73,26	81,52
Région urbaine	38,92	52,04	72,44	65,62	75,46
	Probabilités de réponse estimées pour le modèle saturé				
Région rurale	50,79	62,37	76,90	70,57	83,07
Région urbaine	35,17	50,85	74,79	70,68	72,89

Les probabilités de réponse estimées reflètent le taux de réponse plus faible des ménages d'une seule personne et le taux de réponse plus faible dans les régions urbaines. Les ménages de cinq personnes et plus sont ceux dont le taux de réponse est le plus élevé. D'après les modèles, la probabilité estimée de réponse est, curieusement peut-être, plus élevée pour les ménages de trois personnes que pour ceux de quatre personnes. Ce résultat pourrait tenir au fait que les femmes choisissent souvent d'avoir deux enfants et que les ménages de trois personnes sont, pour la plupart, constitués d'une mère, d'un père et d'un *petit* enfant. Ce genre de famille a tendance à rester à la maison et, donc, à être plus accessible qu'une famille de quatre personnes typiques avec deux enfants plus âgés.

Le taux estimé de réponse plus élevé pour les ménages de trois personnes que pour ceux de quatre personnes équivaut à ce que le ratio $P(Y=3|R=1)/P(Y=3|R=0)$ soit plus grand que le ratio $P(Y=4|R=1)/P(Y=4|R=0)$. Ceci concorde avec la distribution de la taille des ménages du tableau 2, où nous estimons que $P(Y=4) \approx P(Y=4|R=1)$, c'est-à-dire $P(Y=4|R=0) \approx P(Y=4|R=1)$. Par ailleurs, les estimations du tableau 2 indiquent que $P(Y=3|R=1) > P(Y=3)$, ce qui signifie que $P(Y=3|R=1) > P(Y=3|R=0)$.

Nous voyons que le modèle logistique RM2 combiné au modèle de population avec fonction de lien paramétrique $p_{y,x}$ contrainte a un effet de lissage sur les estimations fondées sur le modèle saturé donné par le modèle (4.9), à cause de l'hypothèse supplémentaire de parallélisme des fonctions logit des probabilités de réponse pour les régions urbaine et rurale.

4. Estimateurs des totaux selon la taille du ménage

À la présente section, nous présentons les estimateurs des totaux selon la taille du ménage, ainsi que la méthode d'estimation de la variance. Nous utilisons un estimateur du maximum de vraisemblance avec la fonction de lien paramétrique contrainte donnée par (3.1) comme modèle de population. Nous montrons que cet estimateur est identique à un estimateur poststratifié basé sur l'imputation, qui de nouveau, s'avère être un estimateur par poststratification standard si l'on ne tient pas compte du mécanisme de réponse. De surcroît, nous présentons un estimateur poststratifié imputé, basé sur un modèle saturé pour la taille du ménage et la probabilité de réponse.

4.1 Estimateurs fondés sur une fonction de lien paramétrique contrainte comme modèle de population

Si N_y représente le nombre total de personnes vivant dans un ménage de taille y, le nombre de ménages de taille y est égale à $H_y = N_y/y$. Le nombre total de ménages est représenté par H, $H = \sum_y H_y$.

Le problème statistique consiste à estimer H_y pour $y = 1, ..., J$ et H. Nous choisissons la taille la plus grande J de façon que les ménages de taille supérieure à J soient peu nombreux. Strictement parlant, H_J est le nombre de ménages de taille égale ou supérieure à J, et il en est de même pour N_J. Pour notre application, nous choisissons $J = 5$ car la fréquence des ménages de plus de cinq personnes est faible dans l'échantillon. Nous pouvons écrire $N_y = \sum_{i=1}^{N} I(Y_i = y)$, où la fonction indicatrice $I(Y_i = y) = 1$ si $Y_i = y$, et 0 autrement. Donc, avec $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_N)$,

$$E(H_y|\mathbf{x}) = \frac{1}{y}\sum_{i=1}^{N} P(Y_i = y|\mathbf{x}_i).$$

Nous pouvons obtenir un estimateur par la méthode du maximum de vraisemblance pour H_y en estimant $E(H_y|\mathbf{x})$, c'est-à-dire en remplaçant $P(Y_i = y|\mathbf{x}_i)$ par l'estimateur du maximum de vraisemblance $\hat{P}(Y_i = y|\mathbf{x}_i)$. Les données sont stratifiées en fonction de la taille de la famille 1, ..., K, où la dernière catégorie contient les personnes appartenant aux familles de taille $\geq K$. Si nous utilisons le modèle avec la fonction de lien paramétrique contrainte définie par (3.1), nous supposons que Y dépend uniquement de la taille de la famille x, et l'estimateur prend la forme

$$\hat{H}_y = \frac{1}{y}\sum_{x=1}^{K} M_x \hat{P}(Y = y|x) \qquad (4.1)$$

où M_x (M_K) représente le nombre de personnes dans la population dont la taille enregistrée de la famille est $x (\geq K)$. Les M_x sont des données auxiliaires connues provenant du Registre norvégien des familles.

Une approche courante pour corriger pour la non-réponse consiste à imputer les valeurs manquantes dans l'échantillon. En nous fondant sur la distribution estimée de Y pour une taille de famille et un lieu de résidence donnés pour les non-répondants, $\hat{P}(Y=y|x, z, r=0)$, nous affectons les non-répondants aux valeurs 1, ..., 5 dans les proportions données par $\hat{P}(Y = y|x, z, r = 0)$ pour $y = 1, ..., 5$. Soit $n^*_{xy}(0)$ ($n^*_{xy}(1)$) le nombre de valeurs imputées pour la taille de famille x et la taille de ménage y, pour les régions rurales (urbaines) et soit $m_{xu}(0)$ ($m_{xu}(1)$) le nombre d'observations manquantes pour les personnes dans les régions rurales (urbaines) dont la taille de la famille est x. Alors

$$n^*_{xy}(z) = m_{xu}(z) \cdot \hat{P}(Y = y|x, z, r = 0), z = 0, 1. \quad (4.2)$$

et

$$n^*_{xy} = n^*_{xy}(0) + n^*_{xy}(1)$$

est le nombre total de valeurs imputées pour lesquelles la taille de la famille est x et la taille du ménage est y, c'est-à-dire, n^*_{xy} est le nombre prévu estimé de ménages de taille y, *sachant* la taille de la famille x et $r = 0$.

Le résultat général suivant tient, ce qui montre qu'avec le modèle de population (3.1), l'estimateur du maximum de vraisemblance (4.1) est identique à un estimateur poststratifié basé sur l'imputation.

Théorème. Supposons que Y est donné par le modèle (3.1). Autrement dit, $P(Y=y\,|\,x,\,z)=p_{y,\,x}$ est indépendant de z, mais par ailleurs les $p_{y,\,x}$ sont complètement inconnues, la seule contrainte étant $\sum_y p_{y,\,x}=1$, pour toutes les valeurs de x. Le mécanisme de réponse est paramétrisé arbitrairement, c'est-à-dire qu'aucune hypothèse n'est faite au sujet de $P(R=1\,|\,Y=y,\,x,\,z)$. Alors, les estimations du maximum de vraisemblance pour $p_{y,\,x}$ sont données, pour $x=1,\,...,\,K$, par

$$\hat{p}_{y,\,x}=\frac{n_{xy}+n^{*}_{xy}}{m_x+m_{xu}},$$

où n_{xy} est le nombre de répondants appartenant à une famille de taille x et à un ménage de taille y, $m_x\,(m_K)$ est le nombre de répondants appartenant à une famille de taille $x(\geq K)$, et $m_{xu}=m_{xu}\,(0)+m_{xu}\,(1)$.

Preuve. Voir l'annexe A2.

Le théorème implique que l'estimateur peut s'écrire sous forme de l'estimateur poststratifié basé sur l'imputation, en utilisant la taille de la famille comme variable de stratification,

$$\hat{H}^{\,I}_{y,\,\text{post}}=\frac{1}{y}\sum\nolimits_{x=1}^{K}M_x\,\frac{n_{xy}+n^{*}_{xy}}{m_x+m_{xu}}. \qquad (4.3)$$

Si nous supposons que le mécanisme de réponse est ignorable et que nous utilisons le modèle (3.1), la fonction de vraisemblance est donnée par $\prod_{i=1}^{n_r}P(Y_i=y_i\,|\,x_i)$. Alors, la fonction du maximum de vraisemblance $\hat{P}(Y=y\,|\,x)$ est simplement le taux observé chez les répondants dont la taille du ménage est y, sachant la taille x de la famille. Donc, l'estimateur du maximum de vraisemblance s'avère être identique à l'estimateur poststratifié standard, avec la taille de la famille comme variable de stratification,

$$\hat{H}_{y,\,\text{post}}=\frac{1}{y}\sum\nolimits_{x=1}^{K}M_x\,\frac{n_{xy}}{m_x}. \qquad (4.4)$$

Pour une étude générale de la poststratification, voir, par exemple, Holt and Smith (1979) et Särndal, Swensson et Wretman (1992, chapitre 7.6).

Afin d'illustrer les effets de la modélisation de la non-réponse et de la poststratification, nous présentons aussi des estimations fondées sur l'estimateur par expansion ordinaire, donné par

$$\hat{H}_{y,\,e}=\frac{1}{y}\cdot N\frac{n_y}{n_r} \qquad (4.5)$$

et par l'estimateur par expansion basé sur l'imputation donné par

$$\hat{H}^{\,I}_{y,\,e}=\frac{1}{y}\cdot N\frac{n_y+n^{*}_{y}}{n}. \qquad (4.6)$$

Ici, n_y est le nombre de répondants dans les ménages de taille y, n_r est le nombre total de répondant, et $n^{*}_{y}=\sum_x n^{*}_{xy}$. L'estimateur (4.5) ne vise pas à corriger pour la non-réponse ni n'utilise la distribution de la population de familles comme outil de poststratification pour améliorer l'estimation, tandis que l'estimateur (4.6) essaye de tenir compte du mécanisme de non-réponse, mais ne peut fournir une correction pour les échantillons non représentatifs.

4.2 Poststratification basée sur l'imputation avec un modèle saturé

Nous passons maintenant à une méthode intuitive d'imputation qui a été utilisée pour estimer les probabilités de réponse pour un estimateur d'Horvitz-Thompson modifié pour la production des statistiques officielles d'après les données de l'EDC de 1992 (décrite dans Belsby 1995). Nous utilisons cette méthode d'imputation pour l'estimateur poststratifié (4.3).

La méthode d'imputation consiste à répartir, dans une région rurale/urbaine, les $m_{xu}(z)$ unités non-répondantes sur les ménages de taille 1, ..., 5 de telle façon que, sachant la taille du ménage, le taux de non-réponse soit le même pour toutes les tailles de familles. Cette méthode repose sur l'hypothèse implicite que la probabilité de réponse des personnes dont la taille du ménage est la même dans une région rurale/urbaine est identique pour diverses tailles de familles. Soit $h_{xy}(z)$ le nombre de non-répondants dont la taille de la famille est x, la taille du ménage est y et le lieu de résidence est z. Le nombre correspondant de répondants est $n_{xy}(z)$. Les valeurs de $h_{xy}(z)$ sont déterminées au moyen des équations

$$\frac{h_{xy}(z)}{h_{xy}(z)+n_{xy}(z)}=\frac{h_{iy}(z)}{h_{iy}(z)+n_{iy}(z)},\quad z=0,1. \qquad (4.7)$$

Si $n_{xy}(z)=0$, nous posons que $h_{xy}(z)=0$. L'équation (4.7) est résolue sous les conditions

$$\sum_y h_{xy}(z)=m_{xu}(z);\,x=1,2,3,4,5 \text{ et } z=0,1. \qquad (4.8)$$

La résolution des équations (4.7) et (4.8) nécessite, pour chaque valeur de z, une ligne $(n_{x1}(z),\,n_{x2}(z),...,\,n_{x5}(z))$ de valeurs non nulles, ce qui est vérifié dans notre cas. Les valeurs imputées $h_{xy}(z)$ déterminées par (4.7) et (4.8) correspondent à la méthode d'imputation décrite par (4.2) pour le modèle suivant :

$$P(Y=y\,|\,x,z)=p_{y,\,x,\,z} \text{ sans contraintes} \qquad (4.9a)$$

$$P(R=1\,|\,Y=y,x,z)=q_{y,\,z}\text{, indépendant de } x. \qquad (4.9b)$$

Nous pouvons le montrer comme suit :

Pour les dix essais multinomiaux déterminés par les différentes valeurs (x, z), nous avons 50 probabilités de cellule inconnues $\pi_{yx,z} = P(Y = y, R = 1 | x, z)$. En l'absence de contraintes sur les probabilités de cellule, les estimations du maximum de vraisemblance (evm) sont données par les fréquences relatives observées,

$$\hat{\pi}_{yx,z} = \frac{n_{xy}(z)}{m_x(z) + m_{xu}(z)}.$$

Ce résultat tient aussi quand $n_{xy}(z) = 0$. Maintenant, nous pouvons montrer qu'il existe une correspondance biunivoque entre $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1)$ et $(\boldsymbol{p}_0, \boldsymbol{q}_0, \boldsymbol{p}_1, \boldsymbol{q}_1)$, où $\boldsymbol{\pi}_z = (\pi_{yx,z} : y = 1, ..., 5; x = 1, ..., 5)$, $\boldsymbol{p}_z = (p_{yx,z} : y = 1, ..., 5; x = 1, ..., 5)$ et $\boldsymbol{q}_z = (q_{1,z}, ..., q_{5,z})$. Puisque $\pi_{yx,z} = p_{y,x,z} \cdot q_{yz}$, les emv de $p_{yx,z}$ et de $q_{y,z}$ doivent satisfaire

$$\hat{p}_{yx,z} \cdot \hat{q}_{y,z} = \frac{n_{xy}(z)}{m_x(z) + m_{xu}(z)} \qquad (4.10)$$

et sont déterminées de façon unique par $\hat{\pi}_{yx,z}$.

Considérons $h_{xy}(z)$, donné par (4.5) et (4.6). Soit $h_y(z) = \sum_x h_{xy}(z)$ et $n_y(z) = \sum_x n_{xy}(z)$. D'après (4.7),

$$\frac{h_j(z)}{h_j(z) + n_j(z)} = \frac{h_{xj}(z)}{h_{xj}(z) + n_{xj}(z)}, \text{ si } n_{xj}(z) > 0. \qquad (4.11)$$

De (4.10) et (4.11), il découle que les estimations intuitives suivantes sont aussi des emv.

$$\hat{q}_{y,z} = \frac{n_y(z)}{n_y(z) + h_y(z)} \qquad (4.12)$$

$$\hat{p}_{y,x,z} = \frac{n_{xy}(z) + h_{xy}(z)}{m_x(z) + m_{xu}(z)} \qquad (4.13)$$

(aussi quand $n_{xy}(z) = h_{xy}(z) = 0$).

(Nous pouvons aussi montrer (4.12) et (4.13) en maximisant directement la log-vraisemblance.) Puis, nous montrons que les valeurs imputées (4.2) pour le modèle (4.9) sont égales à $h_{xy}(z)$. D'après (4.2), nous avons $n^*_{xy}(z) = m_{xu}(z) \cdot \hat{P}(Y = y | x, z, r = 0)$. Sous le modèle (4.9) et les estimations (4.12) et (4.13), nous trouvons que

$$\begin{aligned}\hat{P}(Y = y | x, z, R = 0) &= \\ &\frac{\hat{P}(Y = y | x, z) - \hat{P}(Y = y, R = 1 | x, z)}{\hat{P}(R = 0 | x, z)} \\ &= \frac{\hat{p}_{yx,z} - \hat{\pi}_{yx,z}}{1 - \sum_y \hat{\pi}_{yx,z}} \\ &= \frac{n_{xy}(z) + h_{xy}(z) - n_{xy}(z)}{m_{xu}(z)} = \frac{h_{xy}(z)}{m_{xu}(z)},\end{aligned}$$

et il s'ensuit que $n^*_{xy}(z) = h_{xy}(z)$. Si $n_{xy}(z) = 0$, alors $\hat{p}_{y,x,z} = \hat{\pi}_{yx,z} = 0$, et $n^*_{xy}(z) = 0$. Nous notons que le modèle (4.9) est saturé et donnera, d'après (4.10), un ajustement parfait.

Les estimations par expansion basée sur l'imputation (4.6), avec le modèle (4.9), sont identiques aux estimations d'Horvitz-Thompson modifiées avec $\hat{q}_{y,z} = n_y(z) / [n_y(z) + n^*_y(z)]$ (provenant de (4.12)) pour les probabilités estimées de réponse utilisées pour la production des statistiques officielles d'après les données de l'EDC de 1992. Ceci découle du fait que l'estimateur d'Horvitz-Thompson modifié de N_y est donné par

$$\hat{N}_{y,\mathrm{HT}} = \sum_{i \in s_r} \frac{I(Y_i = y)}{\pi_i},$$

où $\pi_i = P$ (que la personne i soit sélectionnée dans l'échantillon et réponde). Donc,

$$\pi_i = \frac{n}{N} \hat{P}(R_i = 1 | x_i, z_i, Y_i = y) = \frac{n}{N} \hat{q}_{y,z_i}$$

et

$$\hat{N}_{y,\mathrm{HT}} = \frac{N}{n} \left(\frac{n_y(0)}{\hat{q}_{y,0}} + \frac{n_y(1)}{\hat{q}_{y,1}} \right). \qquad (4.14)$$

Ici,

$$\begin{aligned}&\hat{N}_{y,\mathrm{HT}} = \\ &\frac{N}{n} \left(\frac{n_y(0)}{n_y(0)/(n_y(0) + n^*_y(0))} + \frac{n_y(1)}{n_y(1)/(n_y(1) + n^*_y(1))} \right) \\ &\quad = N \frac{n_y + n^*_y}{n}.\end{aligned}$$

Donc, cet estimateur d'Horvitz-Thompson modifié souffre de la même caractéristique négative que l'estimateur par expansion basé sur l'imputation (4.6); il ne peut corriger le biais dans un échantillon non représentatif. Pour une description générale de la méthode d'Horvitz-Thompson modifiée, consulter, par exemple, Särndal et coll. (1992, chapitre 15).

4.3 Estimation de la variance

Nous estimons la variance des diverses estimations par la méthode du bootstrap. Cette estimation peut se faire sous le cadre de référence de la modélisation ou de la quasi-randomisation (Little et Rubin 1987). Par exemple, pour estimer la variance sous le modèle (3.1) et le mécanisme de réponse RM1 (3.2), nous pouvons appliquer le bootstrap paramétrique avec les paramètres estimés (Efron et Tibshirani 1993). Cependant, la façon de comparer les variances estimées sous les divers modèles n'est pas claire. Nous avons, par conséquent, choisi d'estimer les variances des divers estimateurs sous un cadre commun de quasi-randomisation. Nous supposons que nous avons affaire à un

échantillonnage aléatoire simple conditionnel, sachant la taille de la famille, la seule hypothèse que nous faisons pour l'estimation de la variance. Inconditionnellement, nous avons un échantillon autopondéré, mais pas aléatoire simple et, par conséquent, il s'agit d'une approximation assez grossière du plan de sondage conditionnel réel. Cependant, pour une étude comparative des estimateurs, l'approximation suffira. L'indicateur de non-réponse r_i est considéré comme étant une constante associée à la personne i. Nous tirons l'échantillon bootstrap, en rééchantillonnant $(y_i, z_i, r_i = 1)$, $(z_i, r_i = 0)$ aléatoirement avec remise, comme il l'est décrit dans Shao et Sitter (1996, section 5), dans chaque poststrate de $\{i; x_i = x\}$. Bien que les tailles des poststrates de l'échantillon soient fixes, aussi bien le nombre de non-répondants que le nombre de personnes provenant d'une région urbaine ou rurale varie d'un échantillon bootstrap à l'autre. Nous calculons les estimations bootstrap de la même façon que celles fondées sur les données observées. En particulier, nous imputons les données bootstrap de la même façon que les données originales si l'estimateur est basé sur l'imputation. Enfin, nous obtenons les variances et les erreurs-types estimées par l'approximation habituelle de Monte Carlo fondée sur 500 échantillons bootstrap indépendants.

5. Nombres estimés de ménages de tailles différentes d'après l'Enquête sur les dépenses de consommation de la Norvège de 1992

Nous présentons ici les nombres estimés de ménages de taille un à cinq et plus, et le nombre total de ménages dans la population norvégienne de moins de 80 ans. L'estimation est faite d'après les données de l'EDC de 1992 et fondée sur les estimateurs décrits à la section 4. Pour calculer les estimations, nous devons connaître les nombres de familles des différentes tailles dans la population, c'est-à-dire M_x, au moment de l'enquête de 1992. Le nombre réel au moment de l'enquête n'a pas été enregistré. À titre d'approximation, nous utilisons les nombres au 1er janvier 1993. Ils sont donnés au tableau 4.

Tableau 4
Nombre de familles et de personnes de moins de 80 ans en Norvège le 1er janvier 1993

Nombre de personnes dans la famille	Familles	Personnes
1 personne	793 869	793 869
2 personnes	408 440	816 880
3 personnes	261 527	784 581
4 personnes	266 504	1 066 016
5 personnes ou plus	127 653	670 528
Total	1 857 993	4 131 874

Notons que la taille moyenne de la famille pour les familles de 5 personnes ou plus est égale à 670 528/ 127 653 = 5,25. Nous utilisons 5,25 comme estimation de la taille moyenne du ménage pour les ménages de 5 personnes ou plus et nous divisons par 5,25 au lieu de 5 dans toutes les estimations de H_5.

5.1 Estimation du maximum de vraisemblance et poststratification

Les distributions estimées des ménages sont présentées au tableau 5. Les estimations sont fondées sur l'estimateur du maximum de vraisemblance (emv) (4.1) en utilisant le modèle de population avec la fonction de lien paramétrique contrainte $p_{y,x}$ combinée aux modèles de réponse RM1(y, z) et RM2(y, z). Pour illustrer l'effet de la modélisation de la non-réponse par opposition à celui de la poststratification, nous présentons aussi l'estimateur poststratifié standard (4.4). Rappelons qu'il s'agit de l'estimateur du maximum de vraisemblance lorsqu'on ne tient pas compte du mécanisme de non-réponse. En outre, nous présentons la distribution estimée de la taille des ménages d'après la poststratification basée sur l'imputation (4.3) avec le modèle saturé (4.9). Pour évaluer la variabilité d'échantillonnage des divers estimateurs, nous incluons aussi les estimations des erreurs-types.

Les trois modèles qui tiennent compte du mécanisme de réponse donnent des nombres totaux de ménages plus élevés. Ils produisent aussi des nombres considérablement plus élevés de ménages d'une seule personne. Ce résultat nous paraît raisonnable, puisque nous nous attendons à ce que le taux de non-réponse le plus élevé soit celui observé pour cette catégorie de ménages. Et, par conséquent, c'est la prise en compte du mécanisme de non-réponse qui a le plus d'influence sur ces estimations. Nous notons que le modèle à fonction de lien paramétrique contrainte (3.1) conjugué au modèle de réponse logistique RM2(y, z) donne pratiquement les mêmes estimations poststratifiées que le modèle (4.9), ainsi qu'approximativement les mêmes erreurs-types. Étant donné le niveau de liberté du modèle (4.9), avec un ajustement parfait, il semble que le modèle (3.1) et le mécanisme de réponse RM2(y, z) donnent de bons résultats pour estimer le nombre de ménages de diverses tailles. En ce qui concerne l'incertitude des estimations, nous constatons, comme on pourrait s'y attendre, que les erreurs-types semblent typiquement augmenter avec le nombre de paramètres inconnus dans le modèle sous-jacent. En outre, le nombre total de ménages est estimé de façon assez précise, si l'on ne tient pas compte du biais éventuel, tandis qu'il est manifestement plus difficile d'estimer le nombre de ménages d'une personne.

Afin de déterminer dans quelle mesure les différences entre les estimations sont dues à l'erreur d'échantillonnage

ou au biais de non-réponse, nous examinons les erreurs-types estimées des différences entre les estimations ponctuelles. Certaines de celles-ci sont données au tableau 6, en utilisant principalement la poststratification basée sur l'imputation avec le modèle saturé comme référence. Brièvement, nous utilisons les termes Est1 à Est4 pour les estimations définies de la façon dont elles figurent au tableau 5 :

Est1 : Estimateur du maximum de vraisemblance fondé sur le modèle de population $p_{y,x}$ et le modèle de réponse RM1

Est2 : Estimateur du maximum de vraisemblance fondé sur le modèle de population $p_{y,x}$ et le modèle de réponse RM2

Est3 : Poststratification basée sur l'imputation fondée sur le modèle saturé (4.9)

Est4: Estimateur poststratifié sans imputation.

L'examen des tableaux 5 et 6 nous permet de conclure que les valeurs prévues de Est4 et Est3 diffèrent lorsqu'on estime H_1, H_3, H_5 et H. En ce qui concerne les autres comparaisons, nous constatons que, lors de l'estimation de H_3, l'écart entre Est1 et Est2/Est3 est significatif et nous notons, d'après les discussions antérieures présentées à la section 3.3, que RM2 donne un meilleur ajustement aux données que RM1.

Les estimations basées sur l'estimateur par expansion $\hat{H}_{y,e}$ donné par (4.5), exprimées en centaines, sont 390 500, 496 500, 283 900, 279 900, 148 000 et 1 598 800, et les erreurs-types estimées sont égales à 33 100, 21 700, 14 600, 11 600, 6 100 et 23 700 pour H_1, ..., H_5 et H, respectivement. Les erreurs-types des différences entre ces estimations et les estimations Est3 sont 52 800, 30 900, 19 100, 10 800, 5 400 et 32 000 pour H_1, ..., H_5 et H, respectivement. Ces estimations par expansion témoignent d'un biais important dû à la non-réponse, particulièrement celles de H_1, H_5 et H, avec correction d'une part du biais par poststratification (probablement de l'ordre de 50 % pour les estimations de H_1 et H). Nous notons aussi que les erreurs-types de l'estimateur poststratifié et de cet estimateur par expansion simple sont à peu près les mêmes. Donc, en réduisant le biais par poststratification, on réduit simultanément l'erreur totale.

La poststratification corrige pour le biais causé par la divergence entre les distributions de la taille de la famille dans l'échantillon de répondants et dans la population. D'après les tableaux 1 et 4, nous voyons que ces distributions de la taille de la famille (en pourcentage), pour $x = 1, \ldots, 5$, sont

Éch. répondants : 14,6 – 20,7 – 19,1 – 27,0 – 18,6
Population : 19,2 – 19,8 – 19,0 – 25,8 – 16,2.

Puisque le nombre de familles d'une seule personne est beaucoup trop faible dans l'échantillon de répondants, il en sera de même de l'estimation par expansion de H_1. Si les poststrates sont déterminées d'après la taille de la famille, la poststratification corrige le biais lié à la taille de la famille dans l'échantillon de répondants, mais ne suppose pas implicitement que la distribution des tailles de ménage est la même pour les non-répondants que pour les répondants, pour une taille de famille donnée. Autrement dit, les répondants sont traités comme un sous-échantillon aléatoire d'unités échantillonnées ayant la même taille de famille, comme le mentionne Little (1993), ce qui n'est fort probablement pas le cas en réalité. Rappelons que l'effet de la variable de taille de la famille n'est pas significatif quand la variable de taille du ménage est incluse dans les modèles de réponse. Donc, il semble raisonnable de supposer, comme dans nos modèles de réponse, que les taux de réponse varient en fonction de la taille réelle du ménage plutôt qu'en fonction de la taille enregistrée de la famille. Habituellement, les estimations du nombre de ménages d'une personne seront biaisées si l'on ne tient pas compte des non-répondants.

Après correction pour le biais de non-réponse en complétant l'échantillon au moyen de valeurs imputées, l'échantillon proprement dit peut être asymétrique comparativement à la population. Afin d'illustrer l'effet de la poststratification pour corriger cette asymétrie, nous comparons, en utilisant le modèle saturé (4.9), les estimations poststratifiées basées sur l'imputation Est3 aux estimations par expansion basées sur l'imputation données par (4.6) : 583 900, 567 700, 244 300, 259 300, 122 400 et 1 777 600 pour H_1, ..., H_5 et H, respectivement. Comme nous l'avons mentionné à la section 4.2, voir (4.14), ces estimations sont identiques aux estimations d'Horvitz-Thompson modifiées. Leurs erreurs-types sont pratiquement les mêmes que celles de Est3. Donc, les méthodes d'estimation poststratifiées de rechange fondées sur des modèles de réponse non ignorables donnent des erreurs-types qui ne sont, au moins, pas pires que celles données par l'estimateur d'Horvitz-Thompson modifié. Donc, si l'on réduit le biais en recourant aux méthodes de rechange, on réduit aussi l'erreur totale. Les erreurs-types des différences entre Est3 et cet estimateur d'Horvitz-Thompson modifié pour les estimations de H_1, ..., H_5 et H sont 3 500, 2 200, 1 100, 600 200 et 2 100, respectivement. Manifestement, ces deux méthodes donnent des estimations significativement différentes pour tous les totaux selon la taille du ménage. Dans cette comparaison, une caractéristique ressort. L'estimation par expansion du nombre de ménages de deux personnes, 567 700, est clairement trop élevée, comme le montre la comparaison des distributions de la taille de la famille dans l'échantillon total et dans la population (en pourcentage), pour $x = 1, \ldots, 5$:

Tableau 5
Nombres totaux estimés de ménages pour les personnes de moins de 80 ans en Norvège au 1er janvier 1993, exprimés en centaines. Erreur-type estimée des estimations entre parenthèses

	Estimateur du maximum de vraisemblance avec mécanisme de réponse non ignorable				Poststratification basée sur l'imputation		Mécanisme de réponse ignoré	
Taille du ménage, y	Modèle de population $p_{y,x}$ et modèle de réponse RM1 (y, z)	%	Modèle de population $p_{y,x}$ et modèle de réponse RM2 (y, z)	%	Modèle de population et de réponse saturé	%	Estimateur poststratifié	%
1	558 800 (38 900)	32	595 400 (48 000)	34	596 600 (53 500)	34	486 000 (35 800)	29
2	520 200 (20 600)	30	525 800 (27 400)	30	523 600 (29 800)	30	507 800 (20 000)	30
3	278 900 (13 800)	16	249 100 (20 300)	14	250 000 (19 800)	14	286 200 (14 100)	17
4	258 900 (9 800)	15	269 000 (11 600)	15	268 900 (11 500)	15	270 600 (10 100)	16
≥ 5	125 800 (4 700)	7	126 000 (5 100)	7	126 200 (5 000)	7	131 300 (4 700)	8
Total	1 742 600 (25 600)	100	1 765 300 (29 700)	100	1 765 300 (31 900)	100	1 681 900 (23 300)	100

Tableau 6
Erreurs-types estimées des différences entre les estimations ponctuelles du tableau 5

Taille du ménage	Est1−Est2	Est1−Est3	Est2−Est3	Est4−Est3
1	29 700	37 000	16 600	42 400
2	19 300	22 200	8 800	23 100
3	15 400	15 200	5 300	15 500
4	6 700	6 500	1 800	6 600
≥ 5	1 700	1 700	500	1 900
Total	15 300	18 800	8 900	23 300

Population : 19,2 – 19,8 – 19,0 – 25,8 – 16,2
Échantillon : 18,6 – 23,0 – 17,8 – 24,9 – 15,7.

La proportion, dans l'échantillon, de personnes appartenant à une famille de deux personnes est beaucoup trop élevée et, bien que nous ayons corrigé pour le biais de non-réponse, l'estimateur par expansion, et donc aussi l'estimateur d'Horvitz-Thompson modifié, ne peut corriger la non-représentativité d'un échantillon. Ceci produira nécessairement des estimations biaisées de H_2. Nous devons recourir à la poststratification pour corriger l'asymétrie d'un échantillon. Nous pouvons considérer la différence entre les valeurs prévues de ces estimateurs de H_2 comme étant proche du biais de l'estimateur d'Horvitz-Thompson modifié, et nous notons qu'un intervalle de confiance à 95 % approximatif pour cette différence est (39 800, 48 400).

Pour des raisons de robustesse, nous présentons aussi les estimations d'après le modèle logit cumulatif mentionné à la section 3.1 combiné au modèle de réponse RM1 (y, z) qui, nous le savons, est mal ajusté aux données. Ces estimations sont exprimées en centaines : 591 800, 501 000, 265 200, 267 300, 128 200 et 1 753 500 pour H_1, ..., H_5 et H, respectivement. Comparativement au tableau 5, ces valeurs semblent indiquer qu'un modèle raisonnable pour la réponse joue un rôle plus important qu'un bon modèle de population. Il est également évident que la modélisation de la non-réponse importe, comme le montre la comparaison à la poststratification et à l'expansion simple.

5.2 Comparaison avec les estimations utilisées à l'heure actuelle pour l'EDC, l'Enquête sur la qualité du Recensement de 1990 et une étude de projection

Depuis 1993, un estimateur d'Horvitz-Thompson modifié de type (4.14), dont le calcul est plus simple, est utilisé pour produire les statistiques officielles d'après les données de l'EDC (voir Belsby 1995). Nous avons indiqué à la section 2 que les poids sont les inverses des probabilités d'échantillonnage des ménages, multipliés par la probabilité de réponse estimée. Les probabilités de réponse sont estimées au moyen d'un modèle logistique semblable à RM2 (y, z) en utilisant le lieu de résidence et la taille du ménage comme variables explicatives. Pour les non-répondants dont on ne connaît pas la taille du ménage, on utilise la taille enregistrée de la famille, en remplaçant (3.5).

Donc, nous pouvons considérer les poids comme étant une approximation de l'utilisation de (3.5). Évidemment, (3.5) n'est possible que si l'on envisage un modèle de population, ce qui n'a pas été fait dans le cas de l'EDC. Le tableau 7 donne la distribution estimée des ménages d'après cet estimateur d'Horvitz-Thompson modifié pour l'EDC.

L'Enquête sur la qualité du Recensement de 1990, appelée EEP 1990, compte 8 280 répondants et repose sur pratiquement la même définition du ménage que l'EDC. Le taux de réponse était de 95 %. Les estimations de H_y sont calculées par poststratification en fonction de la taille du ménage au recensement. Cependant, aucun effort n'a été fait pour corriger le biais de non-réponse éventuel en ce qui a trait à la taille réelle du ménage. L'EEP porte sur l'ensemble de la population. Le tableau 7 donne les estimations pour le groupe de personnes de 0 à 79 ans calculées selon la même méthode de poststratification que pour l'EEP.

Le tableau 7 présente aussi les estimations basées sur l'étude de projection des ménages réalisée par Keilman et Brunborg (1995). Cette étude simule la structure des ménages pour la période allant de 1990 à 2020. Les sources des données sont 28 384 individus provenant du Recensement de la population et du logement de 1990 et de l'Enquête sur la famille et la profession de 1988. Keilman et Brunborg calculent des projections pour l'ensemble de la population en 1992. Nous avons ajusté leurs estimations au groupe des 0 à 79 ans.

Les estimations présentées au tableau 7 renforcent notre impression que les estimations fondées sur la modélisation du mécanisme de réponse donne des résultats moins biaisés que ceux obtenus en ne tenant pas compte de ce mécanisme, comme dans la simple poststratification ou la simple expansion. Il en est particulièrement ainsi pour les ménages d'une personne et pour le total. L'« estimateur officiel » courant, c'est-à-dire l'estimateur d'Horvitz-Thompson modifié, semble produire des estimations dont l'ordre de grandeur est correct et qui, en fait, s'approchent plus des résultats de l'EEP de 1990 que les estimations fondées sur un modèle. Cependant, ces résultats sont plus accidentels qu'autre chose. En tant que *méthode*, l'estimateur modifié présente certains problèmes même dans le cas d'un échantillon représentatif. Nous pouvons les étudier en estimant les probabilités de réponse. Le tableau 8 donne les résultats ainsi que les estimations fondées sur RM2 (*y, z*) et (3.1) tirés du tableau 3.

Tableau 7
Totaux estimés selon la taille du ménage pour les personnes de moins de 80 ans en Norvège au 1er janvier 1993 d'après l'estimateur d'Horvitz-Thompson modifié pour l'EDC, l'EEP de 1990 et les projections, en centaines

Taille du ménage	Horvitz-Thompson modifié pour l'EDC	%	EEP 1990	%	Projections	%
1	622 900	35	626 000	35	668 300	37
2	518 500	29	494 200	28	549 000	30
3	259 900	15	291 500	16	211 900	12
4	258 500	15	250 000	14	221 500	12
≥ 5	124 600	7	115 300	6	97 500	5
Inconnue					78 500	4
Total	1 784 400	1	1 777 000	99	1 826 700	100

Tableau 8
Probabilité estimée de réponse basée sur la méthode utilisée dans l'EDC depuis 1993, en pourcentage

	Taille du ménage				
Lieu de résidence	1	2	3	4	5 ou plus
	Méthode EDC				
Région rurale	44,53	66,24	74,55	73,54	80,07
Région urbaine	36,01	57,90	67,25	66,09	73,80
	Modèle $p_{y,x}$ dans (3.1) combiné à RM2(*y*, *z*)				
Région rurale	47,77	60,90	79,05	73,26	81,52
Région urbaine	38,92	52,04	72,44	65,62	75,46

Comparativement aux probabilités de réponse estimées d'après le modèle RM2 (y, z) avec (3.1), nous voyons que le remplacement de la taille du ménage par la taille de la famille dans le groupe de non-répondants n'est pas une approximation satisfaisante. Donc, si l'on fait une comparaison avec l'estimateur d'Horvitz-Thompson modifié de la section 5.1 fondé sur le modèle saturé (4.9), ce dernier serait le modèle privilégié. Dans le cas de l'EDC, l'approche adoptée officiellement surestime la probabilité de réponse des ménages de deux personnes, ce qui, dans un échantillon représentatif, produirait une sous-estimation de H_2. Les probabilités de réponse estimées seront le plus probablement biaisées si nous utilisons la taille de la famille à la place de la taille du ménage dans le groupe de non-répondants pour estimer les paramètres du modèle de réponse. Ce biais est un problème qui s'ajoute à celui mentionné antérieurement, à savoir que l'estimateur d'Horvitz-Thompson modifié produit des estimations semblables à celles produites par l'estimateur par expansion basées sur l'imputation et ne peut corriger pour les échantillons non représentatifs (problème qui se pose dans le cas de l'EDC depuis 1993). Toutefois, pour l'EDC de 1992, l'échantillon est asymétrique et contient une proportion trop élevée de familles de deux personnes, et l'ordre de grandeur de l'estimation de H_2. sera correcte par accident.

6. Conclusions

Nous avons étudié les problèmes de modélisation et de méthodologie que pose l'estimation du nombre total de ménages de diverses tailles en Norvège d'après les données de l'Enquête sur les dépenses de consommation (EDC) de la Norvège. Le problème principal est de savoir comment corriger le biais dû à la non-réponse non ignorable. La méthode d'estimation appliquée à l'heure actuelle pour l'EDC est l'utilisation d'un estimateur d'Horvitz-Thompson modifié qui comprend une correction pour la non-réponse par estimation des probabilités de réponse. Nous avons fondamentalement examiné deux approches fondées sur un modèle, à savoir un estimateur du maximum de vraisemblance et une poststratification basée sur l'imputation d'après la taille enregistrée de la famille. Avec un modèle de population correspondant à un modèle de groupe d'après la taille de la famille uniquement, ces deux estimateurs sont identiques. Ce modèle de groupe de famille, pour la taille du ménage et une fonction de lien logistique pour la probabilité de réponse avec la taille du ménage comme variable catégorique nominale semblent donner de bons résultats pour notre problème d'estimation.

Lors de l'analyse des données de l'EDC de 1992, nous observons un biais de non-réponse important, particulièrement dans les estimations de H_1 et H, biais qui est corrigé en partie (probablement environ 50 % pour les estimations de H_1 et H) par la poststratification pure (sans imputation). Cependant, la poststratification ne tient pas compte du biais de non-réponse éventuel en fonction de la taille du ménage. Nos modèles de réponse supposent que les taux de réponse varient selon la taille réelle du ménage plutôt que selon la taille enregistrée de la famille, et il est assez évident que ce genre de modélisation de la non-réponse a de l'importance et donne lieu à des estimations moins biaisées que la simple poststratification ou la simple expansion, particulièrement pour H_1 et H.

Les estimations d'Horvitz-Thompson modifiées utilisées pour produire les statistiques officielles d'après l'EDC correspondent aux estimations par expansion basées sur l'imputation. Donc, elles ne permettent pas de corriger l'erreur due à la non-représentativité des échantillons. L'étude décrite ici montre qu'en plus d'utiliser un modèle de réponse non ignorable, il est nécessaire de poststratifier d'après la taille de la famille, c'est-à-dire d'utiliser un modèle de population sachant la taille de la famille. Par conséquent, la poststratification, la modélisation de la réponse et l'imputation sont des éléments essentiels à une approche satisfaisante.

Dans le cas de toute estimation de totaux en sondage, il faut être conscient du fait qu'un estimateur d'Horvitz-Thompson ne peut corriger le biais dû aux échantillons asymétriques, même s'il est modifié au moyen de bonnes estimations de la réponse. La poststratification devrait toujours être envisagée, ainsi que l'imputation fondée sur un modèle de réponse, non ignorable au besoin.

Annexe A1

Les données pour les régions rurale et urbaine sont présentées séparément au tableau A1.

Annexe A2

Théorème. Supposons que Y suit le modèle (3.1), c'est-à-dire que $P(Y = y \mid x, z) = p_{y,x}$ est indépendant de z, mais que, par ailleurs, les $p_{y,x}$ sont complètement inconnues, la seule contrainte étant que $\sum_y p_{y,x} = 1$, pour toutes les valeurs de x, pour tout k. Le mécanisme de réponse est paramétrisé arbitrairement, autrement dit aucune hypothèse n'est faite au sujet de $P(R = 1 \mid Y = y, x, z)$. Alors, les estimations du maximum de vraisemblance pour $p_{y,x}$ sont données par

$$\hat{p}_{y,x} = \frac{n_{xy} + n^*_{xy}}{m_x + m_{xu}}.$$

Preuve. Soit $q_{yx,z} = P(R = 1 \mid Y = y, x, z)$. La log-vraisemblance est donnée par

Tableau A1
Tailles de la famille et du ménage pour l'Enquête sur les dépenses de consommation de la Norvège de 1992, selon la région rurale ou urbaine. L'entrée supérieure est pour le groupe urbain

		Taille du ménage								
Taille de la famille		1	2	3	4	≥ 5	Réponse totale	Non-réponse	Total	Taux de réponse
1	urbain	28	24	7	2	0	61	78	139	0,439
	rural	55	24	13	7	2	101	75	176	0,574
2	urbain	6	70	12	3	0	91	84	175	0,520
	rural	3	107	25	1	3	139	76	215	0,647
3	urbain	4	8	57	11	3	83	40	123	0,675
	rural	6	17	74	29	3	129	51	180	0,717
4	urbain	0	3	15	80	5	103	43	146	0,705
	rural	2	10	22	151	12	197	80	277	0,711
≥ 5	urbain	0	1	0	6	66	73	28	101	0,723
	rural	1	3	4	11	115	134	32	166	0,807
Total urbain		38	106	91	102	74	411	273	684	0,601
Total rural		67	161	138	199	135	700	314	1 014	0,690

$$\ell = \sum_x \sum_y n_{xy}\, p_{y,\,x} + \sum_{z=0}^{1} \sum_x \sum_y n_{xy}(z)\, q_{yx,\,z}$$

$$+ \sum_{z=0}^{1} \sum_x m_{xu}(z) \log P(R=0 \mid x, z)$$

$$= \sum_x \sum_y n_{xy}\, p_{y,\,x} + \sum_{z=0}^{1} \sum_x \sum_y n_{xy}(z)\, q_{yx,\,z}$$

$$+ \sum_{z=0}^{1} \sum_x m_{xu}(z) \log(1 - \sum_{y=1}^{5} p_{y,\,x}\, q_{yx,\,z}).$$

Nous utilisons la méthode de Lagrange et maximisons $G = \ell + \sum_{x=1}^{5} \lambda_x (\sum_{y=1}^{5} p_{y,\,x} - 1)$.

Posons que les solutions sont $\hat{p}_{y,\,x}(\lambda_x)$ et déterminons les λ_x tels que $\sum_y \hat{p}_{y,\,x}(\lambda_x) = 1$, pour tout x. Quelle que soit la façon dont les $q_{yx,\,z}$ sont paramétrisés, l'emv $\hat{p}_{y,\,x}$ doit satisfaire, en résolvant les équations $\partial G / \partial p_{y,\,x} = 0$,

$$\frac{n_{xy}}{\hat{p}_{y,\,x}} - \sum_{z=0}^{1} m_{xu}(z) \frac{\hat{q}_{yx,\,z}}{\hat{P}(R=0 \mid x, z)} + \lambda_x = 0 \qquad \text{(A1)}$$

qui est équivalent à :

$$n_{xy} = \hat{p}_{y,\,x} \sum_{z=0}^{1} \frac{m_{xu}(z)}{\hat{P}(R=0 \mid x, z)}$$

$$- \sum_{z=0}^{1} m_{xu}(z) \frac{\hat{P}(R=0, Y=y \mid x, z)}{\hat{P}(R=0 \mid x, z)} - \hat{p}_{y,\,x} \lambda_x.$$

Nous déterminons λ_x par sommation sur y :

$$m_x = \sum_{z=0}^{1} \frac{m_{xu}(z)}{\hat{P}(R=0 \mid x, z)}$$

$$- \sum_{z=0}^{1} m_{xu}(z) \frac{\hat{P}(R=0 \mid x, z)}{\hat{P}(R=0 \mid x, z)} - \lambda_x,$$

donc

$$\lambda_x = \sum_{z=0}^{1} \frac{m_{xu}(z)}{\hat{P}(R=0 \mid x, z)} - (m_x + m_{xu}).$$

Il découle de (A1) que $\hat{p}_{y,\,x}$ satisfait la relation suivante :

$$\hat{p}_{y,\,x} = \frac{n_{xy}}{\left(m_x + m_{xu} - \sum_{z=0}^{1} m_{xu}(z) \dfrac{\hat{P}(R=0 \mid Y=y, x, z)}{\hat{P}(R=0 \mid x, z)} \right)}. \qquad \text{(A2)}$$

Les valeurs imputées sont données par, tiré de (4.2),

$$n^*_{xy}(z) = m_{xu}(z)\, \hat{p}_{y,\,x} \frac{\hat{P}(R=0 \mid Y=y, x, z)}{\hat{P}(R=0 \mid x, z)}$$

et, d'après (A2),

$$\hat{p}_{y,\,x} = n_{xy} \Big/ \left(m_x + m_{xu} - \sum_{z=0}^{1} \frac{n^*_{xy}(z)}{\hat{p}_{y,\,x}} \right)$$

$$= n_{xy} \Big/ \left(m_x + m_{xu} - \frac{n^*_{xy}}{\hat{p}_{y,\,x}} \right)$$

ou bien, de façon équivalente

$$\hat{p}_{y,\,x}(m_x + m_{xu}) - n^*_{xy} = n_{xy},$$

c'est-à-dire, $\hat{p}_{y,\,x} = \dfrac{n_{xy} + n^*_{xy}}{m_x + m_{xu}}$. C.Q.F.D.

Annexe A3

Tableau A2
Échantillon complété, contenant les valeurs imputées, subdivisé en deux groupes, rural et urbain.
L'entrée supérieur est pour le groupe urbain et l'entrée inférieure, pour le groupe rural.
Fondé sur le Modèle (3.1) et RM1 (y, z)

		Taille du ménage					
Taille de la famille		1	2	3	4	≥ 5	Total
1	urbain	77,8	44,1	12,9	3,9	0,3	139
	rural	103,6	43,1	18,4	8,7	2,3	176
2	urbain	10,8	137,9	22,1	3,8	0,4	175
	rural	7,5	168,6	33,9	1,7	3,3	215
3	urbain	7,5	14,3	81,3	16,4	3,6	123
	rural	10,7	25,3	104,8	35,6	3,7	180
4	urbain	0,8	6,4	21,9	110,3	6,6	146
	rural	3,5	16,7	35,1	206,9	14,8	277
≥ 5	urbain	0,5	2,4	1,0	9,0	88,2	101
	rural	1,6	4,7	5,2	14,4	140,1	166
Total / urbain		97,4	205,1	139,2	143,4	99,1	684
rural		126,9	258,4	197,4	267,3	164,2	1 014

Tableau A3
Échantillon complété, contenant les valeurs imputées, subdivisé en deux groupes, rural et urbain.
L'entrée supérieur est pour le groupe urbain et l'entrée inférieure, pour le groupe rural.
Fondé sur le modèle (3.1) et RM2 (y, z)

		Taille du ménage					
Taille de la famille		1	2	3	4	≥ 5	Total
1	urbain	81,6	42,7	10,4	4,0	0,3	139
	rural	107,5	41,5	15,9	8,8	2,3	176
2	urbain	11,9	140,4	18,3	3,9	0,5	175
	rural	8,6	170,9	30,3	1,8	3,4	215
3	urbain	9,4	16,1	75,2	18,6	3,7	123
	rural	13,4	27,7	96,5	38,5	3,9	180
4	urbain	0,8	6,2	18,9	113,5	6,6	146
	rural	3,7	16,2	29,2	213,1	14,8	277
≥ 5	urbain	0,5	2,3	0,6	9,3	88,3	101
	rural	1,7	4,6	4,6	14,9	140,2	166
Total / urbain		104,2	207,7	123,4	149,3	99,4	684
rural		134,9	260,9	176,5	277,1	164,6	1 014

Annexe A4

Table A4
Échantillon complété, contenant les valeurs imputées, subdivisé en deux groupes, rural et urbain.
L'entrée supérieur est pour le groupe urbain et l'entrée inférieure, pour le groupe rural.
Fondé sur le modèle (4.9), c'est-à-dire imputations déterminées par (4.7) et (4.8)

		Taille du ménage					
Taille de la famille		1	2	3	4	≥ 5	Total
1	urbain	79,6	47,2	9,4	2,8	0,0	139
	rural	108,3	38,5	16,9	9,9	2,4	176
2	urbain	17,1	137,7	16,0	4,2	0,0	175
	rural	5,9	171,6	32,5	1,4	3,6	215
3	urbain	11,4	15,7	76,2	15,6	4,1	123
	rural	11,8	27,3	96,2	41,1	3,6	180
4	urbain	0,0	5,9	20,0	113,2	6,9	146
	rural	3,9	16,0	28,6	214,0	14,5	277
≥ 5	urban	0,0	2,0	0,0	8,5	90,5	101
	rural	2,0	4,8	5,2	15,6	138,4	166
Total /urbain		108,1	208,5	121,6	144,3	101,5	684
rural		131,9	258,2	179,4	282,0	162,5	1 014

Tableau A5
Nombres totaux selon la taille de la famille et du ménage pour l'échantillon complété imputé. Fondé sur le modèle (4.9)

	Taille du ménage					
Taille de la famille	1	2	3	4	≥ 5	Total
1	187,9	85,7	26,3	12,7	2,4	315
2	23,0	309,2	48,6	5,7	3,6	390
3	23,2	43,0	172,4	56,7	7,7	303
4	3,9	21,9	48,7	327,2	21,3	423
≥ 5	2,0	6,8	5,2	24,1	229,0	267
Total	240,0	466,6	301,1	426,3	264,0	1 698

Bibliographie

Baker, S.G., et Laird, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.

Belsby, L. (1995). Forbruksundersøkelsen. Vektmetoder, frafallskorrigering og intervjuer-effekt. (The consumer survey. Weight methods, nonresponse correction and interviewer effect), Notater 95/18 Statistics Norway.

Bjørnstad, J.F. (1996). On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association*, 91, 791-806.

Bjørnstad, J.F., et Skjold, F. (1992). Interval estimation in the presence of nonresponse. *The American Statistical Association 1992 Proceedings of the Section on Survey Research Methods*. 233-238.

Bjørnstad, J.F., et Walsøe, H.K. (1991). Predictive likelihood in nonresponse problems. *The American Statistical Association 1991 Proceedings of the Section on Survey Research Methods*, 152-156.

Efron, B., et Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall.

Forster, J.J., et Smith, P.W.F. (1998). Model-based inference for categorical survey data subject to nonignorable nonresponse (avec discussion). *Journal of the Royal Statistical Society B*, 60, 57-70.

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

Hall, B.H., Cummins, C. et Schnake, R. (1991). *TSP Reference Manual, Version 4.2A*, Palo Alto California: TSP International.

Holt, D., et Smith, T.M.F. (1979). Post-stratification, *Journal of the Royal Statistical Society A*, 142, 33-46.

Keilman, N., et Brunborg, H. (1995). *Household Projections for Norway, 1990-2020, Part 1: Macrosimulation*, Rapport 95/21, Statistics Norway.

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

Little, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

Little, R.J.A., et Rubin, D. (1987). *Statistical Analysis with Missing Data*, New York: John Wiley & Sons, Inc.

McCullagh, P., et Nelder, J.A. (1991). *Generalized Linear Models*, *2ième éd*. London: Chapman & Hall.

Shao, J., et Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.

Statistics Norway (1990). *Survey of Consumer Expenditure 1986-88.* Official Statistics of Norway NOS B919.

Statistics Norway (1996). *Survey of Consumer Expenditure 1992-1994*. Official Statistics of Norway NOS C317.

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


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2005
Vol. 31, N° 2, pp. 233-247
Statistique Canada, N° 12-001-XPB au catalogue

Analyse bayésienne des données catégoriques manquantes non ignorables : Une application à la densité minérale osseuse et au revenu familial

Balgobin Nandram, Lawrence H. Cox et Jai Won Choi [1]

Résumé

Le problème que nous considérons nécessite l'analyse de données catégoriques provenant d'un seul tableau à double entrée avec classification partielle (c'est-à-dire avec non-réponses partielle et totale). Nous supposons qu'il s'agit de la seule information disponible. Une méthode bayésienne nous permet de modéliser divers scénarios de données manquantes sous les hypothèses d'ignorabilité et de non-ignorabilité. Nous construisons un modèle de non-réponse non ignorable que nous obtenons par extension du modèle de non-réponse ignorable au moyen d'une loi a priori dépendante des données; l'extension au modèle de non-réponse non ignorable rend le modèle de non-réponse ignorable plus robuste. Nous utilisons un modèle Dirichlet-Multinomial, corrigé pour la non-réponse, pour estimer les probabilités de cellule et un facteur de Bayes pour vérifier l'hypothèse d'association. Nous illustrons notre méthode à l'aide de données sur la densité minérale osseuse et sur le revenu familial. Une analyse de sensibilité nous permet d'évaluer l'effet du choix de la loi a priori dépendante des données. Nous comparons les modèles de non-réponse ignorable et non ignorable au moyen d'une étude par simulation et constatons qu'il existe des différences subtiles entre ces modèles.

Mots clés : Facteur de Bayes; statistique du chi-carré; fonction d'importance; simulation de Monte Carlo à chaînes de Markov; modèle Dirichlet-Multinomial; robuste; tableau de contingence à double entrée.

1. Introduction

En pratique, il est courant d'utiliser des tableaux de contingence à double entrée pour présenter les données d'enquête. Souvent, des données manquent, si bien que la classification des individus échantillonnés n'est que partielle. Donc, les tableaux à double entrée contiennent à la fois des non-réponses partielles (cas où l'une des deux caractéristiques manque) et des non-réponses totales (cas où les deux caractéristiques manquent); voir Little et Rubin (2002, section 1.3) pour les définitions des trois mécanismes donnant lieu aux données manquantes (MCAR – manquent entièrement au hasard, MAR – manquent au hasard, MNAR – ne manquent pas au hasard). Il peut donc exister quatre types de tableau (un tableau où tous les cas sont complets et, éventuellement, trois tableaux contenant, respectivement, les cas avec classification (données) ligne uniquement, les cas avec classification colonne uniquement et les cas sans classification ligne ni colonne). Il se peut que l'on ne connaisse pas le mécanisme produisant les données manquantes. Donc, nous utilisons un modèle dans lequel la fonction de vraisemblance tient compte des différences entre les données observées et les données manquantes (c'est-à-dire données manquantes non ignorables); voir Rubin (1976), ainsi que Little et Rubin (2002) pour la relation entre l'ignorabilité ou la non-ignorabilité et les trois mécanismes de production des données manquantes. Comme la méthode bayésienne offre des avantages bien connus par rapport à l'approche non bayésienne pour ce genre de problème, nous proposons une analyse bayésienne d'un tableau de contingence général $r \times c$, constitué d'un tableau avec cas complets et de trois tableaux supplémentaires. Plus précisément, nous élaborons une méthode pour estimer les probabilités de cellule et pour tester l'association entre les deux variables catégorielles.

Nous supposons que les seules données dont dispose l'analyste sont les cas complets et les trois tableaux supplémentaires. Plus précisément, nous posons qu'il n'existe aucune donnée (provenant de covariables ou d'information a priori) sur la non-ignorabilité. Notre approche bayésienne ne tient pas compte des caractéristiques du plan de sondage (c'est-à-dire pas de poids de sondage, et pas de mise en grappe ni de stratification). Pour présenter les données d'enquête au public, il arrive qu'on supprime les données sur certaines caractéristiques par souci de commodité et pour assurer la protection des renseignements personnels. Nous sommes conscients que le modèle de non-réponse ignorable et le modèle de non-réponse non ignorable pourraient l'un et l'autre être incorrects s'ils ne tiennent pas compte de ces caractéristiques. Cependant, les paramètres du modèle de non-réponse ignorable peuvent être identifiés et estimés, et nous tirons parti de ce fait pour construire un modèle de non-réponse non ignorable qui est relié au modèle de non-réponse ignorable. En outre, dans le modèle de non-réponse ignorable, nous supposons que la non-réponse est produite selon un mécanisme MAR et que

1. Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute road, Worcester MA 01609, Courriel : balnan@wpi.edu; Lawrence H. Cox et Jai Won Choi, Office of Research and Methodology, National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782. Courriel : lgc9@cdc.gov, jwc7@cdc.gov.

les cas incomplets pourraient contenir de l'information (c'est-à-dire les deux tableaux contenant les effectifs marginaux de ligne et de colonne observés, respectivement). Si l'on ne dispose d'aucun renseignement sur le degré de non-ignorabilité, il est raisonnable de généraliser le modèle de non-réponse ignorable. C'est cette approche que nous adoptons ici.

L'article comprend cinq sections. À la section 1, nous décrivons le problème plus en détail et nous examinons la méthodologie connexe. À la section 2, nous décrivons un tableau 3×3 de la densité minérale osseuse (DMO) et du revenu familial (RF) tiré de la troisième National Health and Nutrition Examination Survey (NHANES III), que nous utilisons principalement à titre d'exemple. À la section 3, nous décrivons la méthode suivie pour estimer les probabilités de cellule et nous utilisons le facteur de Bayes pour le test d'association des deux attributs. Pour atteindre ces objectifs, nous commençons par construire un modèle de non-réponse ignorable et nous montrons comment étendre ce modèle en un modèle de non-réponse non ignorable. À la section 4, nous analysons les données de la NHANES III pour démontrer nos méthodes. Nous décrivons aussi une étude par simulation réalisée pour faire une comparaison supplémentaire des modèles de non-réponse ignorable et non ignorable, ainsi qu'une analyse de sensibilité montrant que l'inférence n'est pas trop sensible au choix d'une loi a priori importante. Enfin, à la cinquième section, nous présentons nos conclusions.

1.1 Discussion du problème

Nous ne savons pas s'il convient d'utiliser un modèle de non-réponse ignorable ou un modèle de non-réponse non ignorable, mais il vaut la peine de souligner que Cohen et Duffy (2002) font remarquer que les enquêtes sur la santé sont un bon exemple de situations où il semble plausible que la propension à répondre dépende de l'état de santé. Donc, les modèles de non-réponse non ignorable sont des candidats importants pour l'analyse des données provenant d'enquêtes sur la santé. Pour un tableau de contingence général $r\times c$ (deux variables catégorielles, l'une comptant r catégories et l'autre, c catégories) en présence de non-réponse, nos objectifs sont de montrer comment a) faire une inférence au sujet des probabilités de cellule et b) vérifier l'hypothèse d'absence d'association entre les deux caractéristiques en utilisant le facteur de Bayes. Alors que (a) découle directement de la modélisation, (b) nécessite une étape supplémentaire.

Soit $\boldsymbol{I}_i$ l'indicateur de cellule pour le i^{e} individu dans un tableau $r\times c$ pour $i=1,\ldots,n$ individus. Alors, il est bien connu que, si les $\boldsymbol{I}_i$ sont *indépendants et de même loi*, la statistique du chi-carré de Pearson suit une loi $\chi^2_{(r-1)(c-1)}$. Sinon, la distribution de cette statistique n'est pas $\chi^2_{(r-1)(c-1)}$, ce qui est vérifié lorsque des données manquent et que les répondants et les non-répondants diffèrent. Le cas échéant, des ajustements doivent être faits à la statistique du chi-carré de Pearson. Dans le cadre non bayésien, Chen et Fienberg (1974) ainsi que Wang (2001) proposent des corrections pour les tableaux à double entrée incomplets. Bien que cela ne soit pas directement pertinent ici, il mérite d'être mentionné que des ajustements semblables ont été faits pour l'échantillonnage en grappes et l'échantillonnage aléatoire stratifié (Rao et Scott 1981, 1984). Essentiellement, les travaux de Chen et Fienberg (1974) et ceux de Wang (2001) ne permettent de traiter que la non-réponse partielle; la non-réponse totale est exclue, parce que la modélisation s'inspire des modèles de non-réponse ignorable (par exemple, voir la discussion dans Kalton et Kasprzyk 1986).

La méthode bayésienne nous permet d'utiliser une procédure qui ne repose pas sur la théorie asymptotique, d'intégrer le cas de données manquantes non ignorables dans la modélisation et d'obtenir une solution de remplacement de la statistique du chi-carré de Pearson pour vérifier l'hypothèse d'absence d'association; voir Little (2003) pour une discussion des avantages bien connus de l'approche bayésienne en échantillonnage. Notre solution de rechange pour la statistique du chi-carré de Pearson est basée sur le facteur de Bayes (Kass et Raftery 1995). Ce facteur est une statistique qui compare un modèle avec association à un modèle sans association par le ratio de leurs vraisemblances marginales respectives sous les modèles de non-réponse ignorable et de non-réponse non ignorable séparément.

Little et Rubin (2002, chapitre 15) discutent du problème de la non-réponse non ignorable. Par exemple, Rubin, Stern et Vehovar (1995) (également discuté dans Little et Rubin 2002, page 345) donnent une analyse intéressante du sondage d'opinion réalisé en Slovénie en novembre/décembre 1990 durant lequel des données sur trois variables dichotomiques ont été recueillies auprès de 2 074 participants éventuels au référendum; le taux de non-réponse était de 12 %. Ils ajustent des modèles de non-réponse ignorable ainsi que non ignorable (loglinéaires avec toutes les interactions) aux données et jugent que le modèle de non-réponse ignorable est satisfaisant. Ils déclarent toutefois que, naturellement, cela ne signifie pas qu'un mécanisme MAR devrait s'appliquer automatiquement à tous les cas. Les analyses reposant sur l'hypothèse que le mécanisme est MAR ne seront vraisemblablement pas adéquates si le taux de non-réponse à une enquête est élevé, si l'information sur les covariables est limitée ou s'il existe des cas pour lesquels le mécanisme de production des données manquantes est de toute évidence non ignorable (par exemple, données censurées).

1.2 Méthodes connexes

Notre méthode diffère de celle de Rubin, Stern et Vehovar (1995). Nous partons de l'approche de Nandram et Choi (2002 a, b) selon laquelle un paramètre γ centre (peut-être considéré comme un indice) le modèle de non-réponse non ignorable sur le modèle de non-réponse ignorable. Si $\gamma = 1$, le modèle de non-réponse non ignorable coïncide avec le modèle de non-réponse ignorable et, donc, le modèle de non-réponse non ignorable « dégénère » en le modèle de non-réponse ignorable quand $\gamma = 1$; voir aussi Forster et Smith (1998). Cette approche est utile, car le modèle de non-réponse non ignorable contient le modèle de non-réponse ignorable en tant que cas particulier, exprimant de ce fait l'incertitude quant à l'ignorabilité. Draper (1995) a donné à cette approche le nom d'*extension continue de modèle* et a recommandé son utilisation de préférence à une extension discrète de modèle (c'est-à-dire mélanges finis) dans la mesure du possible. Nous appelons simplement l'extension continue de modèle un modèle *à facteur d'extension*. Nandram et Choi (2002 a, b) obtiennent le centrage en prenant $\gamma \mid \nu \sim \text{Gamma}(\nu, \nu)$ avec $E(\gamma \mid \nu) = 1$, $\text{var}(\gamma \mid \nu) = 1/\nu$.

Nandram et Choi (2002 a) analysent des données binaires sur les crimes domestiques provenant de la National Crime Survey et, dans Nandram et Choi (2002 b), des données binaires sur les visites chez le médecin provenant de la National Health Interview Survey. Alors que Nandram et Choi (2002 a) contient plus de comparaisons, Nandram et Choi (2002 b) comprend un plus grand nombre d'analyses de sensibilité. Nandram, Han et Choi (2002) décrivent deux modèles bayésiens hiérarchiques, soit un modèle de non-réponse ignorable et un modèle de non-réponse non ignorable, pour l'analyse des données de dénombrement provenant de plusieurs régions, les dénombrements étant décrits pour chaque région par une loi multinomiale. Dans tous ces travaux, la question de l'association est sens objet, car il n'existe qu'une seule variable nominale.

L'approche de Nandram et Choi (2002 a, b) est séduisante, mais elle ne peut s'appliquer directement au problème considéré ici du tableau de contingence $r \times c$. Plus précisément, Nandram et Choi (2002 a, b) n'ont eu besoin que d'un seul paramètre de centrage. Pour étendre leur méthode, nous avons besoin de rc paramètres de centrage. La distribution de chacun de ces paramètres doit être centrée sur la valeur 1 pour permettre la dégénération en le modèle de non-réponse ignorable. Des contraintes d'inégalité doivent également être incluses dans le modèle de non-réponse non ignorable. Par conséquent, bien que l'idée soit intéressante, la méthodologie nécessaire pour appliquer les travaux de Nandram et Choi (2002 a, b) dépasse de loin le cadre du présent article.

Nandram, Liu, Choi et Cox (2005) étendent les travaux de Nandram, Han et Choi (2002) dans deux directions importantes afin a) de considérer plusieurs tableaux de contingence à deux variables au lieu de tableaux à une seule variable et b) d'élaborer une méthode permettant d'étudier l'association entre les deux variables catégorielles. Nandram, Liu, Choi et Cox (2005) analysent les données sur la relation entre la densité minérale osseuse (DMO) et l'âge recueillies pour 35 comtés dans le cadre de la troisième National Health and Nutrition Examination Survey. Dans chaque comté, les données sont ventilées en deux catégories d'âge et trois catégories de DMO (autrement dit, il existe 35 tableaux de contingence 2×3). Il convient de souligner que l'âge est observé pour chaque individu, mais que les valeurs de la DMO manquent pour un grand nombre d'entre eux. Donc, pour chaque comté, il existe un tableau contenant les cas complets et un tableau contenant les totaux de ligne (c'est-à-dire les cas pour lesquels l'âge est connu, mais non la valeur de la DMO). Ici, l'objectif est d'étendre les travaux de Nandram, Liu, Choi et Cox (2005) à un tableau de contingence $r \times c$ *général*. Il s'agit d'un progrès important, puisque nous avons trois tableaux supplémentaires (un tableau avec la classification ligne uniquement, un avec la classification colonne uniquement et un troisième ne contenant ni classification ligne ni classification colonne) au lieu d'un seul avec les totaux de ligne comme dans Nandram, Liu, Choi et Cox (2005).

2. Données sur la densité minérale osseuse et le revenu familial

Nous décrivons brièvement le tableau de contingence 3×3 de la densité minérale osseuse (DMO) et du revenu familial (RF). RF est une variable discrète comportant trois niveaux : faible, moyen et élevé. Bien que DMO soit une variable continue, l'Organisation mondiale de la santé l'a classée en trois niveaux : normal, ostéopénie et ostéoporose; voir Looker, Orwoll, Johnston, Lindsay, Wahner, Dunn, Calvo et Harris (1997, 1998). DMO est utilisée pour diagnostiquer l'ostéoporose, maladie qui se manifeste chez les femmes âgées, et, dans la NHANES III, elle est évaluée chez des individus ayant au moins 20 ans (autrement dit, nous utilisons les données sur les femmes blanches uniquement, ayant plus de 20 ans et présentant des problèmes de santé chroniques).

Parmi celles ayant participé à la phase d'examen physique, des données sur le RF ainsi que sur la DMO ont été recueillies auprès d'environ 62 %, des données sur la DMO uniquement auprès de 8 %, des données sur le revenu uniquement auprès de 29 %; enfin, aucune donnée sur le revenu ni sur la DMO n'ont été recueillies auprès de 1 %. L'ensemble de données utilisé pour notre étude est présenté

au tableau 1 sous forme de tableau de contingence 3×3 de la DMO et du RF. Notre problème consiste à estimer la proportion d'individus pour chaque niveau, ou cellule, DMO-RF et de vérifier l'hypothèse d'une association entre DMO et RF. Dans la NHANES III, le taux de réponse augmente jusqu'à l'âge de 20 ans, puis se stabilise après cet âge; la race, le sexe et les poids de sondage jouent un rôle mineur (voir Nandram et Choi 2005). Donc, ici, nous supposons que les seules données disponibles sont les quatre tableaux regroupant la DMO et le RF, et nous élaborons une méthode applicable à cette situation.

Tableau 1
Classification de la densité minérale osseuse (DMO) et du revenu familial (RF) pour 2 998 femmes blanches ayant au moins 20 ans (20+)

	RF				
DMO	0	1	2	Manquant	Somme
0	621	290	284	135	1 330
1	260	131	117	69	577
2	93	30	18	27	168
Manquante	456	156	266	45	923
Somme	1 430	607	685	276	2 998

Nota : DMO : 0 (> 0,82g/cm^2; normale), 1 (> 0,64, ≤ 0,82g/cm^2; ostéopénie), 2 (≤ 0,64g/cm^2; ostéoporose); RF : 0 (< 20 000 $), 1 (≥ 20 000 $, < 45 000 $), 2 (≥ 45 000 $); la DMO est mesurée uniquement pour les femmes de 20 ans et plus.

Il est difficile d'évaluer une association entre DMO et RF quand la classification est incomplète (c'est-à-dire données manquantes) pour de nombreux individus. Comme il est discuté dans la littérature ne portant pas nécessairement sur la NHANES III, il existe plusieurs variables éventuellement confusionnelles importantes, comme l'âge, l'usage du tabac, l'apport alimentaire de calcium, l'oestrogénothérapie substitutive, l'activité physique, le niveau de scolarité, l'état de santé et la consommation d'alcool (voir Ganry, Baudoin et Fardellone 2000). Selon Farahmand, Persson, Michaelsson, Baron, Parker et Ljunghall (2000), chez les femmes ménopausées de 50 à 81 ans de six comtés de la Suède, un revenu du ménage élevé est associé à une diminution du risque de fracture de la hanche. Au moyen de l'ensemble complet de données provenant de la NHANES III, Lauderdale et Rathouz (2003) étudient la régression de la teneur minérale osseuse sur les indicateurs économiques (par exemple, le niveau de scolarité et le ratio pauvreté-revenu). Ils font une correction pour tenir compte d'autres facteurs comme l'âge, la taille et le poids. Ils concluent que la densité osseuse ne reflète pas les conditions économiques aussi fortement ou uniformément que la stature. Malheureusement, les auteurs de tous ces travaux n'abordent pas la question de la non-ignorabilité des données manquantes; ces dernières ne font l'objet d'aucune discussion. En outre, le taux de réponse aux questions sur le revenu est habituellement faible.

Nous avons examiné de plus près les données pour les cas complets. Nous avons ajusté un modèle Dirichlet-Multinomial avec association et un autre sans association. Le modèle avec association est $\boldsymbol{n} \mid \boldsymbol{p} \sim \text{Multinomiale}(n, \boldsymbol{p})$ et $\boldsymbol{p} \sim \text{Dirichlet}(1, \ldots, 1)$. Notons que, par absence d'association, nous entendons que $p_{jk} = p_j^{(1)}\, p_k^{(2)}$, $j = 1, \ldots, r$, $k = 1, \ldots, c$, où $\sum_{j=1}^{r} p_j^{(1)} = 1$ et $\sum_{k=1}^{c} p_k^{(2)} = 1$. Donc, pour le modèle sans association, $\boldsymbol{n} \mid \boldsymbol{p} \sim \text{Multinomiale}(n, \boldsymbol{p})$, $\boldsymbol{p}^{(1)} \sim \text{Dirichlet}(1, \ldots, 1)$, et indépendamment $\boldsymbol{p}^2 \sim \text{Dirichlet}(1, \ldots, 1)$, où $\boldsymbol{p}^{(1)}$ et $\boldsymbol{p}^{(2)}$ possèdent r et c composantes, respectivement. Il est facile de montrer que la vraisemblance marginale avec association (as) est $p_{\text{as}}(\boldsymbol{n}) = (rc-1)!n!/(n+rc-1)!$ et que la vraisemblance marginale sans association (nas) est

$$p_{\text{nas}}(\boldsymbol{n}) = p_{\text{as}}(\boldsymbol{n}) \frac{(r-1)!(c-1)!}{(rc-1)!} \frac{(n+rc-1)!}{(n+r-1)!(n+c-1)!} \frac{\prod_{j=1}^{r} n_{j\cdot}!\prod_{k=1}^{c} n_{\cdot k}!}{\prod_{j=1}^{r}\prod_{k=1}^{c} n_{jk}!}.$$

Revenons à nos données du tableau 1. Sous l'hypothèse d'indépendance (c'est-à-dire pas d'association), la statistique du chi-carré observée est 12,7 pour quatre degrés de liberté avec une valeur p de 0,013 et nous rejetons l'hypothèse d'absence d'association. Sur l'échelle logarithmique, les vraisemblances marginales sont $p_{\text{nas}}(\boldsymbol{n}) = -46,2$ et $p_{\text{as}}(\boldsymbol{n}) = -49,6$ menant à un logarithme du facteur de Bayes de 3,40 pour la preuve d'une absence d'association relativement à une association. Par conséquent, alors que le test du chi-carré indique fortement qu'il faut rejeter l'hypothèse d'absence d'association, le logarithme du facteur de Bayes indique fortement qu'il faut l'accepter. Donc, les données concernant l'absence d'association sont contradictoires. Voir Mirkin (2001) pour une revue des interprétations de la statistique du chi-carré en tant que mesure d'association ou d'indépendance.

À quel point le facteur de Bayes est-il sensible au choix des lois a priori? En premier lieu, notons que la densité a priori que toute personne raisonnable choisirait pour résoudre le présent problème est la loi de Dirichlet. Pour le modèle avec association, nous choisissons comme lois a priori $\boldsymbol{p} \sim \text{Dirichlet}(\boldsymbol{\gamma})$, et pour le modèle sans association, $\boldsymbol{p}^{(1)} \sim \text{Dirichlet}(\boldsymbol{\alpha})$ et indépendamment $\boldsymbol{p}^{(2)} \sim \text{Dirichlet}(\boldsymbol{\beta})$. Soit $n_{j\cdot}^{(1)} = \sum_{k=1}^{c} n_{jk}$, $j = 1, \ldots, r$ et $n_{\cdot k}^{(2)} = \sum_{j=1}^{r} n_{jk}$, $k = 1, \ldots, c$. Alors, il est facile de montrer que le facteur de Bayes (FB) pour un test d'association contre absence d'association est

$$\text{FB} = \frac{D_{rc}(\boldsymbol{n}+\boldsymbol{\gamma}) / D_r(\boldsymbol{n}_{\cdot}^{(1)}+\boldsymbol{\alpha}) D_c(\boldsymbol{n}_{\cdot}^{(2)}+\boldsymbol{\beta})}{D_{rc}(\boldsymbol{\gamma}) / D_r(\boldsymbol{\alpha}) D_c(\boldsymbol{\beta})},$$

où $D_r(.)$ représente la fonction de Dirichlet avec r composantes, *etc.*; voir la section 3.1 pour la notation. Puis, nous choisissons chacune des composantes de $\boldsymbol{\alpha}, \boldsymbol{\beta}$ et $\boldsymbol{\gamma}$ comme étant κ (par exemple, dans $p_{as}(\boldsymbol{n})$ et $p_{nas}(\boldsymbol{n})$, $\kappa = 1$). Nous pouvons étudier la sensibilité du choix des lois a priori en fonction de κ. Ici $\kappa = 1$ correspond aux lois a priori qui sont habituellement utilisées dans le modèle Dirichlet-Multinomial et $\kappa = 0{,}50$ correspond à la loi a priori de Jeffreys. Donc, nous avons choisi $\kappa = 0{,}25$, 0,5, 1,0, 1,5 et 2,3 et les facteurs de Bayes correspondants (échelle logarithmique) sont 4,7, 3,6, 3,4, 3,9, 4,7 et 6,6. Par conséquent, le facteur de Bayes est sensible au choix des lois a priori, mais pas excessivement. Naturellement, s'il existe des données a priori informatives, pour lesquelles la valeur de κ est fort grande, le problème est différent.

La statistique du chi-carré de Pearson est dominée par les cellules (3, 1) et (3, 3) pour lesquelles les carrés des résidus de Pearson sont 4,61 et 6,15, respectivement (la statistique du chi-carré observée est 12,7). Il est intéressant de noter que le facteur de Bayes a tendance à lisser cet effet. Nous avons regroupé les deux catégories, ostéopénie et ostéoporose, en une seule. Pour ce tableau de contingence 2×3, la valeur de la statistique du chi-carré est 1,7 pour deux degrés de liberté avec une valeur p de 0,42. Les vraisemblance marginales sont $p_{nas}(\boldsymbol{n}) = -28{,}2$ et $p_{as}(\boldsymbol{n}) = -32{,}0$, ce qui donne un logarithme du facteur de Bayes de –3,81. Par conséquent, les deux tests donnent à penser qu'il n'existe pas d'association pour ce tableau 2×3. Donc, si l'on s'en tient à ces données, il est difficile de croire qu'il existe une association entre la DMO et le RF. La question qui se pose maintenant est celle de savoir si cette conclusion change quand on tient compte des données incomplètes.

3. Méthodologie et modèles de non-réponse

Premièrement, nous décrivons la notation. Deuxièmement, nous décrivons le modèle de non-réponse ignorable. Troisièmement, nous construisons un modèle de non-réponse non ignorable en étendant le modèle de non-réponse ignorable. Quatrièmement, nous discutons du facteur de Bayes. Enfin, nous décrivons la façon de spécifier une loi a priori importante.

3.1 Notation

Pour un tableau de contingence $r \times c$, soit $I_{jk\ell} = 1$ si ℓ^{e} individu se situe sur la j^{e} ligne et dans la k^{e} colonne, et 0 autrement. En outre, soit $J_{s\ell} = 1$ si le ℓ^{e} individu se trouvent dans le tableau s ($s = 1$: cas complets; $s = 2$: tableau avec les totaux de ligne; $s = 3$: tableau avec les totaux de colonne; $s = 4$: tableau avec les individus non classés), et $J_{s\ell} = 0$ autrement, $s = 1, 2, 3, 4$ avec $\sum_{s=1}^{4} J_{\ell s} = 1$. Le vecteur $\boldsymbol{J}_\ell, = (J_{1\ell}, J_{2\ell}, J_{3\ell}, J_{4\ell})'$ est celui dont les composantes correspondent aux quatre tableaux.

Soit p_{jk} la probabilité qu'un individu appartienne à la cellule (j, k) du tableau $r \times c$, et soit π_{sjk} la probabilité qu'un individu appartienne au s^{e} tableau, sachant la situation de cette cellule (j, k). Pour le modèle de non-réponse ignorable, $\pi_{sjk} = \pi_s$, mais pour un modèle de non-réponse non ignorable, π_{sjk} dépend au moins d'une valeur de j ainsi que de k. Nous supposons aussi que $\boldsymbol{p}$ est le vecteur $p_{jk}, j = 1, \ldots, r, k = 1, \ldots, c$, et que $\boldsymbol{\pi}_{jk}$ est un vecteur dont les composantes sont $\{\pi_{sjk}, s = 1, \ldots, 4\}$, $j = 1, \ldots, r,\ k = 1, \ldots, c$.

Alors, nous prenons

$$\boldsymbol{I}_\ell \mid \boldsymbol{p} \overset{\text{iid}}{\sim} \text{Multinomiale}\{1, \boldsymbol{p}\}, \qquad (1)$$

où $\sum_{j=1}^{r} \sum_{k=1}^{c} p_{jk} = 1,\ p_{jk} \geq 0,\ j = 1, \ldots, r, k = 1, \ldots, c$. Pour les paramètres $\boldsymbol{p}$, nous prenons

$$\boldsymbol{p} \sim \text{Dirichlet}(1, \ldots, 1),\ p_{jk} \geq 0,\ \sum_{j=1}^{r} \sum_{k=1}^{c} p_{jk} = 1. \qquad (2)$$

Désormais, nous utiliserons la notation qu'un vecteur de dimension $k, \boldsymbol{x} \sim \text{Dirichlet}(\boldsymbol{ct})$ signifie $f(\boldsymbol{x}) = \{\prod_j^k x_j^{c_j t - 1}\} / D_k(\boldsymbol{ct}), x_j \geq 0, \sum_{j=1}^{k} x_j = 1$, où $D_k(\boldsymbol{ct})\ \{\prod_{j=1}^{k} \Gamma(c, t)\} / \Gamma(t)$ est la fonction de Dirichlet avec $c_j > 0, \sum_{j=1}^{k} c_j = 1$.

Les hypothèses (1) et (2) sont les mêmes pour les modèles de non-réponse ignorable et non ignorable, et sont les hypothèses types lorsqu'il n'y a pas de données manquantes.

Soit $y_{sjk} = \sum_{\ell=1}^{n} I_{jk\ell} J_{s\ell}$, $s = 1, 2, 3, 4$ les effectifs de cellule pour les quatre cas. Ici, les valeurs y_{1jk} sont observées et les valeurs $y_{sjk}, s = 2, 3, 4$ manquent (c'est-à-dire, variables latentes). Pour y_{1jk} nous savons que $\sum_{j=1}^{r} \sum_{k=1}^{c} y_{1jk} = n_0$, le nombre d'individus pour lesquels les données sont complètes; pour y_{2jk}, nous savons que $\sum_{k=1}^{c} y_{2jk} = u_j$, où les totaux de ligne $u_j, j = 1, \ldots, r$ sont observés; pour y_{3jk}, nous savons que $\sum_{j=1}^{r} y_{3jk} = v_k$, où les totaux de colonnes $v_k, k = 1, \ldots, c$ sont observés, et pour y_{4jk} nous savons que $\sum_{j=1}^{r} \sum_{k=1}^{c} y_{4jk} = w$. Partout, nous supposons que toute inférence est faite sachant $n_0, \boldsymbol{u}, \boldsymbol{v}$ et w, et nous supprimerons cette notation chaque fois que cela est sous-entendu. Lorsque cela est commode, nous utiliserons les notations telles que $\sum_{s,j,k} y_{sjk} \equiv \sum_{s=1}^{4} \sum_{j=1}^{r} \sum_{k=1}^{c} y_{sjk}$, $\prod_{s,j,k} \pi_{sjk} \equiv \prod_{s=1}^{4} \prod_{j=1}^{r} \prod_{k=1}^{c} \pi_{sjk}$ et $\boldsymbol{y}_{(1)} = (\boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_4), \boldsymbol{y}_{(2)} = (\boldsymbol{y}_1, \boldsymbol{y}_3, \boldsymbol{y}_4)$ *etc.*, où $\boldsymbol{y}_s = (y_{sjk},\ j = 1, \ldots, r,\ k = 1, \ldots, c)$, $s = 1, 2, 3, 4$. En outre, $\sum_{s,j,k}^{4,r,c} y_{sjk} = n$. Nous utiliserons aussi $y_{s\cdot\cdot} = \sum_{j,k} y_{sjk}, y_{\cdot jk} = \sum_s y_{sjk}$ et $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_4)$.

3.2 Modèle de non-réponse ignorable

Pour le modèle de non-réponse ignorable, nous prenons

$$\boldsymbol{J}_\ell \mid \boldsymbol{\pi} \overset{\text{iid}}{\sim} \text{Multinomiale}\{1, \boldsymbol{\pi}\}. \qquad (3)$$

Autrement dit, il n'existe aucune dépendance à l'égard de la situation de cellule d'un individu.

Alors, la fonction de vraisemblance augmentée pour $\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, n_0, \boldsymbol{u}, \boldsymbol{v}, w$ est

$$g(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, n_0, \boldsymbol{u}, \boldsymbol{v}, w) \propto \left[\prod_{s=1}^{4} \pi_s^{y_{s\cdot\cdot}}\right] \left[\prod_{s=1}^{4}\prod_{j=1}^{r}\prod_{k=1}^{c} \frac{p_{jk}^{y_{sjk}}}{y_{sjk!}}\right], \quad (4)$$

sous les contraintes $\sum_{j=1}^{r}\sum_{k=1}^{c} y_{1jk} = n_0$, $\sum_{k=1}^{c} y_{2jk} = u_j$, $j = 1, \ldots, r$, $\sum_{j=1}^{r} y_{3jk} = v_k, k = 1, \ldots, c$, et $\sum_{j=1}^{r}\sum_{k=1}^{c} y_{4jk} = w$. L'équation (4) possède trois caractéristiques intéressantes. En premier lieu, sous l'hypothèse d'ignorabilité, la fonction de vraisemblance est subdivisée en deux éléments, l'un contenant les π_s uniquement et l'autre, les p_{jk}, et les inférences au sujet de ces deux paramètres sont indépendantes. En deuxième lieu, l'inférence au sujet de π_s est fondée uniquement sur les $y_{s\cdot\cdot}$ observés (c'est-à-dire que les statistiques suffisantes pour π_1, π_2, π_3 et π_4 sont essentiellement les proportions de cas dans les premier, deuxième, troisième et quatrième tableaux, respectivement). En troisième lieu, sous le modèle de non-réponse ignorable, les u_j et les v_k contiennent de l'information au sujet de p_{jk}; w ne contient aucune information au sujet des p_{jk}. Il est facile de le démontrer; en notant T l'ensemble $\{(\boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_4) : \sum_{k=1}^{c} y_{2jk} = u_j, j = 1, \ldots, r, \sum_{j=1}^{r} y_{3jk} = v_k, k = 1, \ldots, c, \sum_{j=1}^{r}\sum_{k=1}^{c} y_{4jk} = w\}$, d'après (4)

$$\sum_{(\boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_4) \in T} \prod_{s=1}^{4}\prod_{j=1}^{r}\prod_{k=1}^{c} \frac{p_{jk}^{y_{sjk}}}{y_{sjk}!} = w! \prod_{j=1}^{r} \frac{u_j!}{\left\{\sum_{k=1}^{c} p_{jk}\right\}^{u_j}} \prod_{k=1}^{c} \frac{v_k!}{\left\{\sum_{j=1}^{r} p_{jk}\right\}^{v_k}} \prod_{j=1}^{r}\prod_{k=1}^{c} \frac{p_{jk}^{y_{1jk}}}{y_{1jk}!}.$$

Enfin, pour les paramètres $\boldsymbol{\pi}$, nous prenons

$$\boldsymbol{\pi} \sim \text{Dirichlet}(1, \ldots, 1), \pi_s \geq 0, \sum_{s=1}^{4} \pi_s = 1. \quad (5)$$

Notons qu'il s'agit d'une densité de probabilité uniforme dans un espace quadridimensionnel et qu'il n'y a pas d'hyperparamètres dans ce modèle. Donc, pour le modèle de non-réponse ignorable, si nous combinons (2) et (5), la densité a priori conjointe est

$$g_1(\boldsymbol{p}, \boldsymbol{\pi}) \propto 1, \; p_{jk} \geq 0, \sum_{j=1}^{r}\sum_{k=1}^{c} p_{jk} = 1, \pi_s \geq 0, \sum_{s=1}^{4} \pi_s = 1, \quad (6)$$

ce qui est pertinent.

Enfin, en combinant la fonction de vraisemblance (4) et la densité a priori conjointe (6) par la voie du théorème de Bayes, nous obtenons la densité a posteriori conjointe des paramètres $\boldsymbol{\pi}$, $\boldsymbol{p}$ et $\boldsymbol{y}_{(1)}$

$$\pi(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1) \propto \left[\prod_{s=1}^{4} \pi_s^{y_{s\cdot\cdot}}\right] \left[\prod_{s=1}^{4}\prod_{j=1}^{r}\prod_{k=1}^{c} \frac{p_{jk}^{y_{sjk}}}{y_{sjk}!}\right]. \quad (7)$$

A posteriori, $\boldsymbol{p}$ et $\boldsymbol{\pi}$ sont indépendants. L'inférence au sujet de $\boldsymbol{\pi}$ est facile, car $\boldsymbol{\pi} \mid \boldsymbol{y}_1, \boldsymbol{y}_{(1)} \sim \text{Dirichlet}(y_{1\cdot\cdot} + 1, \ldots, y_{4\cdot\cdot} + 1)$, qui est indépendant de $\boldsymbol{y}_{(1)}$. L'inférence au sujet de $\boldsymbol{p}$ peut s'obtenir en utilisant un simple échantillonneur de Gibbs, car, si nous posons que $q_{jk}^{(1)} = p_{jk} / \sum_{k'=1}^{c} p_{jk'}$ et $q_{jk}^{(2)} = p_{jk} / \sum_{j'=1}^{r} p_{j'k}$, les probabilités conditionnelles sont

$$\boldsymbol{p} \mid \boldsymbol{y} \sim \text{Dirichlet}(y_{\cdot 11} + 1, \ldots, y_{\cdot rc} + 1),$$

$$\boldsymbol{y}_{2j} \mid \boldsymbol{p}, u_j, \boldsymbol{y}_{(2)} \overset{\text{ind}}{\sim} \text{Multinomiale}(u_j, \boldsymbol{q}_j^{(1)}), \; j = 1, \ldots, r,$$

$$\boldsymbol{y}_{3k} \mid \boldsymbol{p}, v_k, \boldsymbol{y}_{(3)} \overset{\text{ind}}{\sim} \text{Multinomiale}(v_k, \boldsymbol{q}_k^{(2)}), \; k = 1, \ldots, c,$$

$$\boldsymbol{y}_4 \mid \boldsymbol{p}, w, \boldsymbol{y}_{(4)} \sim \text{Multinomiale}(w, \boldsymbol{p}). \quad (8)$$

De toute évidence, les paramètres $\boldsymbol{p}$ et $\boldsymbol{\pi}$ sont identifiables et estimables. En outre, notons que, dans (8), $\boldsymbol{y}_4$ est une variable latente et qu'elle ne contribue pas à l'inférence au sujet de $\boldsymbol{p}$. Elle facilite plutôt le calcul en fournissant un échantillonneur de Gibbs simple. Cependant, nous soulignerons que l'information contenue dans $\boldsymbol{y}_4$, par la voie de w, est importante sous un modèle de non-réponse non ignorable.

3.3 Modèle de non-réponse non ignorable

Pour les données manquantes non ignorables, nous prenons

$$\boldsymbol{J}_\ell \mid \{I_{jk\ell} = 1, I_{j'k'\ell} = 0, \; j \neq j', k \neq k', \boldsymbol{\pi}_{jk}\} \overset{\text{iid}}{\sim} \text{Multinomiale}\{1, \boldsymbol{\pi}_{jk}\}. \quad (9)$$

L'hypothèse (9) précise que les probabilités qu'un individu appartienne à l'un des quatre tableaux dépend des deux caractéristiques (c'est-à-dire classifications ligne et colonne) de l'individu. De cette façon, nous intégrons l'hypothèse selon laquelle les données manquantes ne sont pas ignorables. Il s'agit d'une extension du modèle de Nandram, Han et Choi (2002). Nous pouvons aussi prendre $\boldsymbol{\pi}_j$ ou $\boldsymbol{\pi}_k$ au lieu de $\boldsymbol{\pi}_{jk}$; la méthodologie est la même.

Ensuite, nous avons besoin de la fonction de vraisemblance. Ici, la fonction de vraisemblance augmentée pour $\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1$ est

$$g(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, n_0, \boldsymbol{u}, \boldsymbol{v}, w) \propto \left[\prod_{s,j,k}^{4,r,c} \frac{\pi_{sjk}^{y_{sjk}}}{y_{sjk}!}\right] \left[\prod_{j,k}^{r,c} p_{jk}^{y_{\cdot jk}}\right], \quad (10)$$

sous les contraintes $\sum_{j=1}^{r}\sum_{k=1}^{c} y_{1jk} = n_0$, $\sum_{k=1}^{c} y_{2jk} = u_j$, $j = 1, \ldots, r$, $\sum_{j=1}^{r} y_{3jk} = v_k, k = 1, \ldots, c$, et $\sum_{j=1}^{r}\sum_{k=1}^{c} y_{4jk} = w$.

Soulignons que, dans (10), les paramètres p_{jk} et π_{sjk} ne sont pas identifiables. Manifestement, pour estimer p_{jk}, nous avons besoin de connaître $y_{\cdot jk}$, mais nous ne connaissons que y_{1jk}. En outre, pour estimer π_{sjk}, nous devons connaître y_{sjk}, $s = 2,3,4$. Donc, les y_{sjk}, $s = 2,3,4$ ne sont pas identifiables non plus. Appliquer aux π_{sjk} des priors appropriés très informatifs aiderait, mais il ne s'agit pas d'une solution pratique. Si l'on utilise un modèle de non-réponse ignorable (c'est-à-dire $\pi_{sjk} = \pi_s$), alors tous les paramètres peuvent être identifiés. Par conséquent, une solution raisonnable consiste à essayer de lier les $\boldsymbol{\pi}_{jk}$ sur (j,k) en se servant d'une caractéristique commune. Si les $\boldsymbol{\pi}_{jk}$ proviennent d'une distribution commune dont les paramètres sont « connus », nous pourrons les estimer. Autrement dit, nous devons essayer de « renforcer l'information par emprunt », comme dans l'estimation sur petits domaines. Cela nous permettra d'estimer $\boldsymbol{y}_{(1)}$ qui, à son tour, facilitera l'estimation des p_{jk} et π_{sjk}.

Pour les $\boldsymbol{\pi}_{jk}$, nous « centrons » le modèle de non-réponse non ignorable sur le modèle de non-réponse ignorable. Plus précisément, nous supposons que

$$\boldsymbol{\pi}_{jk} \mid \boldsymbol{\mu}, \tau \overset{\text{iid}}{\sim} \text{Dirichlet}\,(\mu_1\tau, \mu_2\tau, \mu_3\tau, \mu_4\tau),$$
$$\pi_{sjk} \geq 0, \sum_{s=1}^{4} \pi_{sjk} = 1, \qquad (11)$$

$j = 1, \ldots, r$, $k = 1, \ldots, c$. Dans (11), le paramètre τ nous renseigne sur la proximité du modèle de non-réponse non ignorable par rapport au modèle de non-réponse ignorable. Par exemple, si τ est petit, les $\boldsymbol{\pi}_{jk}$ seront fort différentes et si τ est grand, les $\boldsymbol{\pi}_{jk}$ seront fort semblables. Donc, l'inférence peut être sensible au choix de τ, et il convient de choisir ce dernier prudemment. En l'absence de toute information au sujet de la non-ignorabilité, il est naturel de choisir une densité a priori pour τ telle que le modèle de non-réponse non ignorable généralise le modèle de non-réponse ignorable. Nous obtenons cette généralisation parce que, à mesure que τ tend vers l'infini, les $\boldsymbol{\pi}_{jk}$ convergent vers la même valeur sur (j,k) (pas en ce qui concerne les composantes), c'est-à-dire le modèle de non-réponse ignorable. Les paramètres $\boldsymbol{\mu}$ et τ ne sont pas identifiables, parce que les $\boldsymbol{\pi}_{jk}$ ne le sont pas. Donc, il est impossible d'estimer $\boldsymbol{\mu}$ et τ sans aucune information; un moyen naturel de procéder consiste à essayer d'utiliser certaines données déjà observées.

Plus précisément, nous prenons a priori que μ et τ sont indépendants avec

$$p(\boldsymbol{\mu}) = 1, \mu_s \geq 0, s = 1, 2, 3, 4,$$
$$\sum_{s=1}^{4} \mu_s = 1, \ \tau \sim \text{Gamma}\,(\alpha_0, \beta_0), \tau \geq 0, \qquad (12)$$

où α_0 et β_0 doivent être spécifiés; sans aucune information au sujet de α_0 et β_0, nous devons utiliser de nouveau les données. Pour faciliter la spécification de α_0 et β_0 pour le modèle de non-réponse non ignorable, nous avons utilisé le modèle de non-réponse ignorable. Le prior sur τ ajoute une variation supplémentaire, ce qui permet un certain degré de non-ignorabilité (voir la section 3.5). Soulignons de nouveau que si τ est très grand (c'est-à-dire $\alpha_0 >> \beta_0$), ce modèle de non-réponse non ignorable dégénère en le modèle de non-réponse ignorable. Donc, une question se pose quant au degré de sensibilité de l'inférence à cette spécification. Naturellement, dans (12), nous pouvons choisir d'autres lois pour τ (par exemple, la loi log-normale), mais là n'est vraiment pas la question essentielle.

Si nous combinons (2), (11) et (12), la densité a priori conjointe de $\boldsymbol{\pi}$, $\boldsymbol{p}$, $\boldsymbol{\mu}$ et τ est

$$\pi(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\mu}, \tau) \propto \left\{ \prod_{j=1}^{r} \prod_{k=1}^{c} \frac{\prod_{s=1}^{4} \pi_{sjk}^{\mu_s\tau - 1}}{D(\boldsymbol{\mu}\tau)} \right\} \tau^{\alpha_0 - 1} e^{-\beta_0\tau}. \qquad (13)$$

Mentionnons de nouveau que (13) est une densité a priori pertinente. Enfin, en combinant la fonction de vraisemblance (10) et la densité a priori conjointe (13) au moyen du théorème de Bayes, nous obtenons la densité a posteriori conjointe des paramètres $\boldsymbol{\pi}$, $\boldsymbol{p}$, $\boldsymbol{\mu}$, τ et des variables latentes $\boldsymbol{y}_{(1)}$

$$\pi(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\mu}, \tau, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1) \propto \left[\prod_{s,j,k}^{4,r,c} \left\{ \frac{(\pi_{sjk} p_{jk})^{y_{sjk}}}{y_{sjk}!} \right\} \right] \left\{ \prod_{j,k}^{r,c} \frac{\prod_{s=1}^{4} \pi_{sjk}^{\mu_s\tau - 1}}{D(\boldsymbol{\mu}\tau)} \right\} \tau^{\alpha_0 - 1} e^{-\beta_0\tau}. \qquad (14)$$

À l'annexe A, nous montrons comment ajuster le modèle de non-réponse non ignorable pour obtenir une inférence appropriée en utilisant l'échantillonneur de Gibbs.

3.4 Facteur de Bayes : Tests d'association et de non-ignorabilité

Nous élaborons un test en vue de vérifier l'association entre DMO et RF. Ce test est une évaluation de l'hypothèse selon laquelle $p_{jk} = q_{1j}q_{2k}$, $j = 1, \ldots, r$, $k = 1, \ldots, c$, et $\sum_{j=1}^{r} q_{1j} = 1$ et $\sum_{k=1}^{c} q_{2k} = 1$. Nous utilisons le facteur de Bayes, c'est-à-dire le ratio des vraisemblances marginales sous deux scénarios (à savoir association contre absence d'association). Soulignons que nous observons $\boldsymbol{y}_1$, mais que $\boldsymbol{y}_{(1)}$ est un ensemble de variables latentes, de sorte que chaque probabilité marginale est simplement la probabilité que $\boldsymbol{y}_1$ soit la valeur observée de $\boldsymbol{Y}_1$, ce que nous notons par $p(\boldsymbol{y}_1)$.

Nous fixons

$$C = \left\{ \begin{array}{l} \boldsymbol{y}_{(1)} : \sum_{k=1}^{c} y_{2jk} = u_j, \ j = 1, \ldots, r; \\ \sum_{j=1}^{r} y_{3jk} = v_k, k = 1, \ldots, c; \sum_{j=1}^{r} \sum_{k=1}^{c} y_{4jk} = w \end{array} \right\}.$$

Alors, en posant que $d = 3!n!(rc-1)!$ et $e = 3!n!(r-1)!(c-1)!$, la vraisemblance marginale pour le modèle de non-réponse ignorable (IG) est

$$p_{\mathrm{IG}}(\boldsymbol{y}_1) = \begin{cases} d \sum_{\boldsymbol{y}_{(1)} \in C} \int\int \prod_{s,j,k}^{4,r,c} \{(\pi_s p_{jk})^{y_{sjk}} / y_{sjk}!\} d\boldsymbol{\pi}\, d\boldsymbol{p}, \\ \qquad \text{association} \\ e \sum_{\boldsymbol{y}_{(1)} \in C} \int\int\int \prod_{s,j,k}^{4,r,c} \{(\pi_s q_{1j} q_{2k})^{y_{sjk}} / y_{sjk}!\} d\boldsymbol{\pi} d\boldsymbol{q}_1 d\boldsymbol{q}_2, \\ \qquad \text{pas d'association,} \end{cases} \tag{15}$$

et en posant que $\Omega_a = (\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\mu}, \tau)$ et $\Omega_{na} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{\pi}, \boldsymbol{\mu}, \tau)$, la vraisemblance marginale pour le modèle de non-réponse non ignorable (NIG) est

$$p_{\mathrm{NIG}}(\boldsymbol{y}_1) = \begin{cases} d \sum_{\boldsymbol{y}_{(1)} \in C} \int_{\Omega_a} \prod_{s,j,k}^{4,r,c} \{(\pi_{sjk} p_{jk})^{y_{sjk}} / y_{sjk}!\} g(\boldsymbol{\pi}, \boldsymbol{\mu}, \tau)\, d\Omega_a, \\ \qquad \text{association} \\ e \sum_{\boldsymbol{y}_{(1)} \in C} \int_{\Omega_{na}} \prod_{s,j,k}^{4,r,c} \{(\pi_{sjk} q_{1j} q_{2k})^{y_{sjk}} / y_{sjk}!\} g(\boldsymbol{\pi}, \boldsymbol{\mu}, \tau) d\Omega_{na}, \\ \qquad \text{pas d'association,} \end{cases} \tag{16}$$

où

$$g(\boldsymbol{\pi}, \boldsymbol{\mu}, \tau) = \frac{\beta_0^{\alpha_0} \tau^{\alpha_0 - 1} e^{-\beta_0 \tau}}{\Gamma(\alpha_0)} \prod_{j=1}^{r} \prod_{k=1}^{c} \left\{ \frac{\prod_{s=1}^{4} \pi_{sjk}^{\mu_s \tau - 1}}{D(\boldsymbol{\mu}\tau)} \right\}. \tag{17}$$

Dans l'ensemble C, la sommation demande beaucoup de calculs, parce qu'il existe de nombreux points $\boldsymbol{y}_{(1)} \in C$ (autrement dit, nous devons calculer la somme sur la totalité de ces points). Pour éviter ce problème, nous commençons par calculer la somme sur C analytiquement, puis nous obtenons le reste par intégration par la méthode de Monte Carlo.

Pour le modèle ignorable, il est facile de démontrer que

$$p_{\mathrm{IG}}(\boldsymbol{y}_1) = \begin{cases} a = \dfrac{3!n!}{n+1} \dfrac{(rc-1)!}{(n+rc-1)!}, \\ \qquad \text{association} \\ b = \dfrac{3!n!}{n+1} \dfrac{(r-1)!(c-1)!}{(n+r-1)!(n+c-1)!} \dfrac{\prod_j y_{1j\cdot}! \prod_k y_{1\cdot k}!}{\prod_j \prod_k y_{1jk}!}, \\ \qquad \text{pas d'association,} \end{cases} \tag{18}$$

où n est le nombre total d'individus dans le tableau entier. Nous décrivons l'estimation de $p_{\mathrm{NIG}}(\boldsymbol{y}_1)$ à l'annexe B.

Cependant, nous notons qu'un test d'ignorabilité ou de non-ignorabilité est subtil, parce que nous supposons que nous ne disposons d'aucune information au sujet de l'ignorabilité ou de la non-ignorabilité. Par contre, notre modèle de non-réponse non ignorable est une généralisation de notre modèle de non-réponse ignorable. Nous pensons donc que le test d'association sous le modèle de non-réponse ignorable ou sous le modèle de non-réponse non ignorable est fiable.

Enfin, nous soulignons que le facteur de Bayes peut être sensible aux spécifications a priori, particulièrement parce qu'il n'existe pas suffisamment de données pour estimer les paramètres soumis au test; voir Sinharay et Stern (2002) pour une discussion intéressante des modèles emboîtés. Nous avons étudié la sensibilité du facteur de Bayes à la spécification de α_0 et de β_0 dans (17); consulter la section 3.5 et le tableau 6. Cet exercice est utile, car il s'agit d'un prior important dans notre modèle de non-réponse non ignorable. Cependant, la comparaison principale est un test d'absence d'association réalisé séparément sous le modèle de non-réponse ignorable et sous le modèle de non-réponse non ignorable. Le paramètre τ entre uniquement dans le modèle de non-réponse non ignorable et possède le même prior sous l'hypothèse d'association et d'absence d'association.

3.5 Spécification de α_0 et β_0

La spécification des hyperparamètres α_0 et β_0 dans $\tau \sim \mathrm{Gamma}(\alpha_0, \beta_0)$ est un point essentiel de notre méthode; voir (12). Elle est importante, parce que nous utilisons cette technique pour rendre le modèle de non-réponse ignorable robuste; nous faisons une analyse de sensibilité plus loin. Notons que $E(\tau) = \alpha_0 / \beta_0$; donc, si $\alpha_0 >> \beta_0$, le modèle de non-réponse non ignorable sera semblable au modèle de non-réponse ignorable. Supposons que nous puissions observer un échantillon aléatoire $\tau^{(1)}, \ldots, \tau^{(M)}$ tiré de la loi $\mathrm{Gamma}(\alpha_0, \beta_0)$. Alors, nous pouvons utiliser une méthode simple (par exemple, la méthode des moments) pour estimer α_0 et β_0.

Comment pouvons-nous obtenir un échantillon correspondant à $\mathrm{Gamma}(\alpha_0, \beta_0)$? L'échantillonneur de Gibbs donné en (8) pour le modèle de non-réponse ignorable produit les valeurs imputées pour les effectifs de cellule manquants. Nous avons imputé les effectifs de cellule manquants M fois, $M = 1\,000$; soit $n_{1jk}^{(h)} \equiv y_{1jk}$ et $n_{sjk}^{(h)}$, $s = 2, 3, 4$, $h = 1, \ldots, M$, les effectifs de cellule manquants. Alors, pour chaque valeur de h, nous ajustons le modèle de non-réponse non ignorable sans la spécification a priori (12),

$$(n_{111}^{(h)}, \ldots, n_{1rc}^{(h)}, \ldots, n_{411}^{(h)}, \ldots, n_{4rc}^{(h)}) \mid \boldsymbol{\pi}, \boldsymbol{p} \sim \text{Multinomiale}\{n, (\pi_{111} p_{11}, \ldots, \pi_{4rc} p_{rc})\}.$$

$$\boldsymbol{p} \sim \text{Dirichlet}(\mathbf{1}), \text{ et } \boldsymbol{\pi}_{jk} \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\alpha})$$

$$\text{où } \alpha_s = \mu_s \tau,\ s = 1, 2, 3, 4.$$

Après calcul de $\boldsymbol{p}$ et $\boldsymbol{\pi}_{jk}$, par intégration, nous obtenons la fonction de vraisemblance,

$$\prod_{j=1}^{r}\prod_{k=1}^{c}\left[\frac{\Gamma\left(\sum_{s=1}^{4}\alpha_s\right)}{\Gamma\left(\sum_{s=1}^{4}(\alpha_s + n_{sjk}^{(h)})\right)}\prod_{s=1}^{4}\frac{\Gamma(\alpha_s + n_{sjk}^{(h)})}{\Gamma(\alpha_s)}\right],$$

$$\alpha_s > 0,\ s = 1, 2, 3, 4. \qquad (19)$$

En utilisant l'algorithme de Nelder-Mead pour maximiser la fonction de vraisemblance (19) sur $\alpha_s > 0,\ s = 1, 2, 3, 4$ à la h^{e} itération, nous obtenons les estimateurs du maximum de vraisemblance $\hat{\boldsymbol{\alpha}}^{(h)}, h = 1, \ldots, M$. Maintenant, en posant que $\tau^{(h)} = \sum_{s=1}^{4}\hat{\alpha}_s^{(h)}$, nous voyons $\tau^{(h)}, h = 1, \ldots, M$ comme un échantillon aléatoire tiré de Gamma (α_0, β_0).

Enfin, par la méthode des moments, nous ajustons la loi $\text{Gamma}(\alpha_0, \beta_0)$ aux « données », $\tau^{(h)}, h = 1, \ldots, M$, pour obtenir $\alpha_0 = a^2/b$ et $\beta_0 = a/b$, où $a = M^{-1}\sum_{h=1}^{M}\tau^{(h)}$ et $b = (M-1)^{-1}\sum_{h=1}^{M}(\tau^{(h)} - a)^2$. Donc, nous avons construit une loi a priori dépendante des données pour τ. Notre procédure donne $\alpha_0 = 125,\ \beta_0 = 0,35$ (c'est-à-dire que τ a une moyenne de 357 et un écart-type de 31,9). À la section 4, nous discutons de la sensibilité à ce choix.

4. Données et analyse empirique

Nous appliquons notre méthode aux données du tableau de contingence 3×3 illustrée au tableau 1. Après avoir présenté les résultats associés aux données observées ainsi qu'une analyse de sensibilité, nous décrivons une étude par simulation en vue d'évaluer la différence entre les modèles de non-réponse ignorable et non ignorable.

4.1 Analyse des données

Consulter le tableau 2 pour une comparaison du modèle de non-réponse ignorable et du modèle de non-réponse non ignorable. Nous avons également inclus l'erreur-type numérique (ETN) qui est une mesure du degré de reproductibilité des résultats numériques; nous l'avons calculée par la méthode des moyennes par lot. Donc, cela ne nous gênerait pas que l'ETN soit petite comparativement aux estimations de Monte Carlo ou aux moyennes a posteriori. Pour les deux modèles, les ETN sont faibles, avec des valeurs relativement grandes pour le modèle de non-réponse non ignorable (proche de zéro dans les deux cas de toute façon), ce qui indique que les calculs sont reproductibles. Les moyennes a posteriori (MP) sont fort semblables pour les deux modèles. Les écarts-types a posteriori (ETP) sont plus grands pour le modèle de non-réponse non ignorable que pour l'autre, ce qui rend les intervalles de confiance à 95 % plus larges. Sous le modèle de non-réponse ignorable, pour ainsi dire tous les intervalles de confiance à 95 % sont contenus dans ceux obtenus pour le modèle de non-réponse non ignorable.

Tableau 2
Comparaison des moyennes a posteriori (MP), des écarts-types a posteriori (ETP), des erreurs types numériques (ETN) et des intervalles de confiance à 95 % (IC) pour p tiré des modèles de non-réponse ignorable et non ignorable

Cellule	$\hat{p}$	MP	ETP	ETN	IC
a) Modèle de non-réponse ignorable					
(1, 1)	0,337	0,330	0,005	0,001	(0,321, 0,339)
(1, 2)	0,157	0,142	0,003	0,001	(0,136, 0,147)
(1, 3)	0,154	0,168	0,004	0,001	(0,162, 0,175)
(2, 1)	0,141	0,142	0,004	0,001	(0,134, 0,148)
(2, 2)	0,071	0,066	0,002	0,001	(0,061, 0,070)
(2, 3)	0,063	0,071	0,003	0,001	(0,066, 0,078)
(3, 1)	0,050	0,053	0,003	0,001	(0,048, 0,059)
(3, 2)	0,016	0,016	0,001	0,000	(0,013, 0,019)
(3, 3)	0,010	0,012	0,002	0,000	(0,009, 0,015)
b) Modèle de non-réponse non ignorable					
(1, 1)	0,337	0,321	0,020	0,009	(0,278, 0,355)
(1, 2)	0,157	0,143	0,008	0,003	(0,126, 0,158)
(1, 3)	0,154	0,173	0,014	0,007	(0,140, 0,196)
(2, 1)	0,141	0,139	0,019	0,009	(0,109, 0,182)
(2, 2)	0,071	0,069	0,007	0,003	(0,056, 0,085)
(2, 3)	0,063	0,071	0,013	0,006	(0,053, 0,102)
(3, 1)	0,050	0,052	0,008	0,002	(0,040, 0,070)
(3, 2)	0,016	0,019	0,003	0,001	(0,014, 0,026)
(3, 3)	0,010	0,013	0,003	0,001	(0,009, 0,020)

Nota : Pour le modèle de non-réponse ignorable, $\pi_{sjk} = \pi_s$, $s = 1, 2, 3, 4,\ j = 1, 2, 3,\ k = 1, 2, 3$. La valeur observée de p basée sur les données complètes est $\hat{p}$.

Au tableau 3, nous comparons également l'estimation de π_s dans le modèle de non-réponse ignorable à π_{sjk} dans le modèle de non-réponse non ignorable. Pour ce dernier, nous présentons la fourchette des moyennes a posteriori (MP) pour les neuf cellules de chaque s, $s = 1, 2, 3, 4$. Elle indique l'importance de la non-ignorabilité. Les MP de π_s sont comprises dans la fourchette des π_{sjk} et, comme prévu, les ETP sont plus grands pour le modèle de non-réponse non ignorable que pour l'autre. Par exemple, sur les neuf cellules, les π_{1jk} varient de 0,388 à 0,656, et ces deux chiffres diffèrent significativement de 0,615, ce qui témoigne d'un certain degré de non-ignorabilité. Donc, il existe une différence entre les modèles de non-réponse ignorable et non ignorable.

Au tableau 4, nous présentons les logarithmes des facteurs de Bayes utilisés pour tester la qualité de l'ajustement du modèle de non-réponse ignorable et du modèle de non-réponse non ignorable. Il existe de « fortes » preuves que le modèle de non-réponse ignorable est mieux ajusté que le modèle de non-réponse non ignorable aux données étudiées (Kass et Raftery 1995). Alors que le modèle de non-réponse ignorable donne de « fortes » preuves d'absence d'association, le modèle de non-réponse non ignorable donne un résultat « positif », comme l'indique

Kass et Raftery (1995). Donc, il existe de nouveau une différence entre les modèles de non-réponse ignorable et non ignorable. Toutefois, l'ETN de 1,80 a tendance à annuler ces différences. Nous concluons qu'il existe des données convaincantes donnant à penser qu'il n'y a pas d'association entre la densité minérale osseuse (DMO) et le revenu familial (RF).

Tableau 3
Comparaison des moyennes a posteriori (MP) et des écarts-types a posteriori (ETP) pour π_{sjk} tiré des modèles de non-réponse ignorable et non ignorable

	Ignorable	Non ignorable
π_1	0,615 (0,009)	0,388 (0,078) – 0,656 (0,044)
π_2	0,077 (0,005)	0,057 (0,017) – 0,195 (0,068)
π_3	0,292 (0,008)	0,217 (0,041) – 0,349 (0,053)
π_4	0,015 (0,002)	0,013 (0,005) – 0,152 (0,055)

Nota : Les ETP figurent entre parenthèses. Pour le modèle de non-réponse ignorable, les paramètres sont π_1, π_2, π_3 et π_4, et pour le modèle de non-réponse non ignorable, les paramètres sont π_{sjk}, $s = 1, 2, 3, 4$, $j = 1, 2, 3$, $k = 1, 2, 3$. Pour chaque *s*, nous avons sélectionné parmi les neuf cellules la plus petite et la plus grande MP pour former l'intervalle.

Tableau 4
Vraisemblances marginales et facteurs de Bayes pour le test d'association entre DMO et RF sous les modèles de non-réponse ignorable et non ignorable

	Association	Pas d'association	Différence
Ignorable	–49,571	–46,173	–3,398
Non ignorable	–53,129	–50,132	–2,996
ETN	1,800	1,790	

Nota : Toutes les entrées (vraisemblances marginales et leurs différences) sont exprimées sur l'échelle logarithmique. L'intégration par la méthode de Monte Carlo comprend 50 000 itérations. Les erreurs-types numériques (ETN) sont faibles comparativement aux vraisemblances marginales.

Nous avons examiné la relation entre DMO et RF quand les niveaux d'ostéopénie et d'ostéoporose sont regroupés en un seul. Sous le modèle de non-réponse ignorable, le logarithme du facteur de Bayes est égal à –2,77 (log vraisemblance marginale : –32,82 et –29,05) et, sous le modèle de non-réponse non ignorable, il est égal à –4,52 (log vraisemblance marginale : –34,25 et –4,52). Donc, nous arrivons à la même conclusion au sujet de l'absence d'association entre DMO et RF.

Nous avons également réparti les données en deux groupes d'âge, c'est-à-dire les femmes préménopausées (ayant, au plus, 49 ans; jeunes) et les femmes ménopausées (ayant au moins 50 ans; âgées). Parmi le groupe de femmes jeunes, quatre seulement faisaient de l'ostéoporose, si bien que nous avons regroupé celles faisant de l'ostéopénie et celles faisant de l'ostéoporose. Nous avons ajusté le modèle de non-réponse ignorable ainsi que le modèle de non-réponse non ignorable à ces données et obtenu des résultats comparables. Pour le groupe de femmes âgées, en utilisant le modèle de non-réponse ignorable, les logarithmes des vraisemblances marginales correspondant à l'absence d'association et à l'existence d'une association sont –43,01 et –38,91, ce qui donne un logarithme du facteur de Bayes de 4,10 pour l'absence d'association. Par conséquent, il existe de fortes preuves d'absence d'association entre DMO et RF. Pour le groupe de femmes jeunes, si nous utilisons le modèle de non-réponse ignorable, les logarithmes des vraisemblances marginales correspondant à l'absence d'association et à l'existence d'une association sont –29,93 et –28,80, ce qui donne un logarithme du facteur de Bayes de 1,13 pour l'absence d'association. Donc, il existe des indices positifs d'une absence d'association entre la densité minérale osseuse et le revenu familial pour les deux groupes d'âge. Par conséquent, il est peu probable que l'âge joue un rôle dans l'association entre les deux variables.

4.2 Analyse de sensibilité

Nous avons étudié la sensibilité de l'inférence au sujet de p_{jk} à la loi a priori de τ. Autrement dit, nous avons pris $\tau \sim \text{Gamma}(\kappa\alpha_0, \beta_0)$, où κ est un paramètre de sensibilité auquel nous avons donné la valeur de 1 dans notre analyse (notons que $E(\tau) = \kappa\alpha_0 / \beta_0$).

Notre méthode pour la spécification de α_0 et β_0 donne les valeurs de $\alpha_0 = 125$ et $\beta_0 = 0,35$; voir la section 3.5. Donner à κ une valeur supérieure à 1 induit des variations moins importantes de la moyenne a posteriori (MP) et de l'écart-type a posteriori (ETP) des p_{jk} que lui donner une valeur inférieure à 1, parce que les valeurs plus élevées de κ provoquent des variations beaucoup plus faibles de la loi a priori de τ. Au tableau 5, nous présentons les MP et les ETP des p_{jk} pour κ = 0,25, 0,50, 1,00, 2,00 et 4,00. La valeur des MP augmente avec κ et celle des ETP diminue lorsque κ passe de 0,25 à 4,00. Donc, il existe une certaine sensibilité à la spécification de α_0 et de β_0, mais les variations sont faibles. Par exemple, les MP de p_{11} sont 0,31, 0,32 et 0,33 pour κ = 0,25, 1,00 et 4,00, et, à ces valeurs de κ, les ETP sont 0,04, 0,02 et 0,01.

Nous avons également étudié la sensibilité des facteurs de Bayes au choix de κ (voir le tableau 6). Pour commencer, les ETN diminuent avec κ, mais la variation est faible. Notons que nous avons utilisé 50 000 itérations pour l'intégration par la méthode de Monte Carlo; cette taille d'échantillon est nécessaire pour que les estimations de Monte Carlo se stabilisent. Les logarithmes des vraisemblances marginales varient peu avec κ. Comme les logarithmes des facteurs de Bayes sont faibles, certaines variations sont reflétées dans l'inférence : pour κ = 0,25, 0,50 et 4,00, il existe de « fortes » preuves de l'absence d'association, mais pour κ = 1,00 et 2,00, il existe des preuves « positives » (limites) de l'absence d'association.

Tableau 5
Sensibilité des moyennes a posteriori (MP) et ainsi que des écarts-types a posteriori (ETP) des p_{jk} au choix de κ dans le modèle de non-réponse non ignorable

κ	0,25		0,50		1,00		2,00		4,00	
Cellule	MP	ETP	MP	ETP	MP	ETP	MP	ETP	MP	ETP
(1, 1)	306,93	36,09	315,01	25,81	321,81	19,95	325,37	14,55	326,16	10,46
(1, 2)	141,12	15,52	139,86	11,91	142,66	8,44	142,63	6,68	143,42	5,01
(1, 3)	161,68	25,80	167,83	18,77	173,40	13,77	176,20	8,44	175,78	6,71
(2, 1)	143,18	34,20	142,62	24,92	138,57	18,82	137,23	13,59	137,26	9,70
(2, 2)	68,46	13,12	71,06	10,09	68,44	7,48	68,79	5,72	68,11	4,45
(2, 3)	79,78	22,83	75,97	17,86	71,11	12,56	68,09	7,84	68,34	6,38
(3, 1)	59,97	21,60	53,50	12,12	52,14	7,76	50,97	5,29	51,41	4,35
(3, 2)	21,43	7,76	20,02	4,89	18,96	23,28	18,67	2,78	17,84	2,23
(3, 3)	17,45	10,38	14,12	4,28	12,93	2,99	12,05	2,34	11,69	1,99

Nota: Toutes les entrées doivent être multipliées par 10^{-3}. Dans le modèle de non-réponse non ignorable, $\pi_{sjk} \overset{\text{iid}}{\sim}$ Gamma $(\kappa\alpha_0, \beta_0)$, où κ est le paramètre de sensibilité et $\alpha_0 = 125$ et $\beta_0 = 0{,}35$.

Dans l'ensemble, il existe un certain niveau de preuve de l'absence d'association. Donc, il est intéressant de savoir que l'on ne doit pas s'inquiéter trop du choix de (α_0, β_0).

Tableau 6
Sensibilité des vraisemblances marginales et du facteur de Bayes au choix de κ dans le modèle de non-réponse non ignorable

	Association		Pas d'association		Facteur de
κ	VM	ETN	VM	ETN	Bayes
0,25	–53,37	1,90	–49,16	1,89	–4,21
0,50	–52,58	1,83	–49,49	1,82	–3,08
1,00	–52,58	1,80	–49,76	1,79	–2,82
2,00	–52,81	1,79	–49,83	1,78	–2,98
4,00	–52,95	1,78	–49,91	1,77	–3,04

Nota : Toutes les entrées sont exprimées sur l'échelle logarithmique. Dans le modèle de non-réponse non ignorable, $\pi_{sjk} \overset{\text{iid}}{\sim}$ Gamma $(\kappa\alpha_0, \beta_0)$, où κ est le paramètre de sensibilité et $\alpha_0 = 125$ et $\beta_0 = 0{,}35$.

4.3 Étude par simulation

Nous avons exécuté une étude par simulation pour poursuivre la comparaison entre les modèles de non-réponse ignorable et non ignorable. L'objectif est de confirmer qu'il existe des différences entre les deux modèles. Dans notre situation, un test fondé sur le facteur de Bayes permettra d'indiquer si ces différences existent ou non. Lorsque l'information au sujet de la non-ignorabilité est limitée (ce qui est le cas ici), il est raisonnable d'ajuster un modèle de non-réponse ignorable, parce que les paramètres sont identifiables dans ce modèle. Donc, nous procédons à la comparaison des modèles de non-réponse ignorable et non ignorable lorsque les données sont générées à partir a) du modèle de non-réponse ignorable et b) du modèle de non-réponse non ignorable. Il s'agit d'une analyse bayésienne typique.

Nous obtenons les moyennes a posteriori des p_{jk} et des π_{sjk}, notées $\tilde{p}_{jk}$ et $\tilde{\pi}_{sjk}$, respectivement, après avoir ajusté nos modèles de non-réponse non ignorable aux données observées. Pour la non-réponse ignorable, nous prenons $\tilde{\pi}_s = \sum_{j=1}^{r}\sum_{k=1}^{c}\tilde{\pi}_{sjk}/rc$, $s = 1, 2, 3, 4$. Nous obtenons les effectifs de cellules pour le modèle de non-réponse ignorable à partir de

$$(y_{111}, \ldots, y_{1rc}, \ldots, y_{411}, \ldots, y_{4rc}) \mid \tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{p}} \sim \text{Multinomiale}\{n, (\tilde{\pi}_1\ \tilde{p}_{11}, \ldots, \tilde{\pi}_4\ \tilde{p}_{rc})\}$$

et pour le modèle de non-réponse non ignorable, par tirage à partir de

$$(y_{111}, \ldots, y_{1rc}, \ldots, y_{411}, \ldots, y_{4rc}) \mid \tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{p}} \sim \text{Multinomiale}\{n, (\tilde{\pi}_{111}\tilde{p}_{11}, \ldots, \tilde{\pi}_{4rc}\tilde{p}_{rc})\},$$

où $n = 2\,998$, le nombre total d'individus dans l'ensemble de données original (voir le tableau 1). Nous avons généré 1 000 ensembles de données pour le modèle de non-réponse ignorable ainsi que pour le modèle de non-réponse non ignorable. Puis, nous avons ajusté les modèles de non-réponse ignorable et non ignorable à chaque ensemble de données exactement de la même manière que pour les données observées du tableau 1 et nous avons calculé les moyennes a posteriori (MP) et les écarts-types a posteriori (ETP) pour les p_{jk}. Au tableau 7, nous présentons les moyennes des MP et des ETP sur les 1 000 ensembles de données. La deuxième colonne (étiquetée $\hat{p}$) contient la moyenne a posteriori de p_{jk} pour les données observées sous le modèle de non-réponse non ignorable (voir le tableau 2b).

Pour (a) au tableau 7, les MP sont très proches des $\hat{p}_{jk}$ pour le modèle de non-réponse ignorable, mais pas autant si l'on ajuste le modèle de non-réponse non ignorable. Il est évident que les ETP sont environ deux fois plus grands sous le modèle de non-réponse non ignorable que sous le modèle de non-réponse ignorable. Pour (b) au tableau 7, les MP s'approchent plus des $\hat{p}_{jk}$ pour le modèle de non-réponse non ignorable que pour le modèle de non-réponse ignorable. Cependant, dans les deux cas, les ETP sont environ deux fois plus grands pour le modèle de non-réponse non ignorable que pour le modèle de non-réponse ignorable. Par exemple, au tableau 7 pour la cellule (1, 1) comparativement à 0,322 pour $\hat{p}$, dans (a), le modèle de non-réponse

Tableau 7
Comparaison des modèles de non-réponse ignorable et non ignorable au moyen de données simulées et des moyennes a posteriori (MP) ainsi que des écarts-types a posteriori (ETP) des p_{jk}

	Simulé	Ignorable (a)				Non ignorable (b)			
	Ajusté	Ignorable		Non ignorable		Ignorable		Non ignorable	
Cellule	$\hat{p}$	MP	ETP	MP	ETP	MP	ETP	MP	ETP
(1, 1)	321,81	320,73	5,72	307,42	11,30	332,02	5,10	324,44	10,60
(1, 2)	142,66	142,96	4,24	146,44	7,34	141,81	3,30	143,44	5,43
(1, 3)	173,40	172,59	4,42	173,49	7,62	168,66	4,14	174,10	7,04
(2, 1)	138,57	138,82	4,81	135,32	9,82	143,63	4,52	139,20	9,74
(2, 2)	68,44	68,44	3,55	72,01	6,02	64,51	2,91	68,20	4,76
(2, 3)	71,11	71,41	3,65	75,00	6,30	70,85	3,76	69,63	6,58
(3, 1)	52,14	52,17	3,11	53,03	4,95	53,08	3,04	52,44	4,70
(3, 2)	18,96	19,35	2,08	21,65	2,98	15,08	1,72	17,32	2,48
(3, 3)	12,93	13,54	1,78	15,64	2,55	10,95	1,85	11,20	2,18

Nota : Les données sont simulées à partir du modèle de non-réponse ignorable en (a) ou du modèle de non-réponse non ignorable en (b), et les modèles de non-réponse ignorable et non ignorable sont tous deux ajustés. Nous avons généré 1 000 ensembles de données et nous avons ajusté le modèle de non-réponse ignorable ainsi que le modèle de non-réponse non ignorable à chaque ensemble de données simulé. Les MP et les ETP sont les moyennes sur les 1 000 ensembles de données et $\hat{p}$ est la moyenne a posteriori pour les données observées que nous avons utilisées pour générer les ensembles de données. Toutes les entrées doivent être multipliées par 10^{-3}.

ignorable (non ignorable) donne une MP de 0,321 (0,307), mais dans (b), le modèle de non-réponse ignorable (non ignorable) donne une MP de 0,332 (0,324) pour d'autres exemples. Donc, les deux modèles donnent effectivement des résultats différents lors de l'estimation de $\boldsymbol{p}$.

Nous avons également considéré l'estimation de la proportion P d'ensembles de données simulés dans lesquels le modèle de non-réponse ignorable donne de meilleurs résultats que le modèle de non-réponse non ignorable. Il est coûteux de calculer la vraisemblance marginale sous le modèle de non-réponse non ignorable. Nous soulignons de nouveau qu'il faut 50 000 itérations pour que l'estimation de Monte Carlo se stabilise; il s'agit là d'une tâche énorme pour l'étude par simulation, parce que nous devons calculer les vraisemblances marginales pour 1 000 ensembles de données. Dons, nous utilisons une méthode simple pour comparer les deux modèles et nous nous attendons à ce qu'elle donne une conclusion comparable à un calcul puissant.

Plus précisément, nous calculons $\Delta^{(h)} = n\sum_{j=1}^{r}\sum_{k=1}^{c}(\hat{p}_{jk} - PM_{jk}^{(h)})^2 / PM_{jk}^{(h)}$, où $PM_{jk}^{(h)}$ est la moyenne a posteriori de p_{jk} correspondant au h^{e} ensemble de données. Nous notons $\Delta^{(h)}$ par $\Delta_{\text{IG}}^{(h)}$ pour le modèle de non-réponse ignorable et par $\Delta_{\text{NIG}}^{(h)}$ pour le modèle de non-réponse non ignorable. Nous obtenons un estimateur de P, $\hat{P}$, en comptant le nombre d'expériences parmi les 1 000 réalisées pour lesquelles $\Delta_{\text{IG}}^{(h)} > \Delta_{\text{NIG}}^{(h)}$. Pour les données générées à partir du modèle de non-réponse ignorable, $\hat{P}$ est égal à 0,236 avec une erreur-type de 0,013. Pour les données générées à partir du modèle de non-réponse non ignorable, $\hat{P}$ est égal à 0,920 avec une erreur-type de 0,009. Donc, si nous nous attendons à ce que le modèle de non-réponse ignorable soit vérifié, il sera battu par le modèle de non-réponse non ignorable environ 24 % du temps, et si nous nous attendons à ce que le modèle de non-réponse non ignorable soit vérifié, il ne sera battu par le modèle de non-réponse ignorable qu'environ (1–0,920) 100 %, soit environ 8 % du temps. Par conséquent, il existe des différences latentes entre les deux modèles. Le modèle de non-réponse non ignorable reflète un certain degré de non-ignorabilité et rend le modèle de non-réponse ignorable plus robuste. Nous considérons qu'il s'agit d'une comparaison raisonnable entre les deux modèles.

5. Conclusion

Deux nouveautés méthodologiques importantes sont exposées dans le présent article. Plus précisément, nous avons montré a) qu'il est possible d'analyser des données multinomiales provenant de tableaux de contingence $r \times c$ en présence à la fois de non-réponse partielle et totale et que le mécanisme de non-réponse peut être non ignorable, et b) qu'en utilisant le facteur de Bayes (ratio des vraisemblances marginales des deux modèles), nous pouvons vérifier s'il existe une association entre les deux caractéristiques. Essentiellement, nous avons supposé qu'il n'existait aucune information au sujet de la non-ignorabilité, nous avons supprimé toutes les caractéristiques du plan de sondage et nous avons adopté une approche prudente.

Pour le tableau de contingence 3×3 contenant des données catégoriques sur la densité minérale osseuse (DMO) et le revenu familial (RF), nous avons montré comment estimer exactement les probabilités de cellule. Pour les cas de données complètes, le facteur de Bayes donne une « forte » preuve d'absence d'association entre les

variables DMO et RF. Pour l'ensemble des données, notre facteur de Bayes indique que la preuve d'absence d'association est « forte » sous le modèle de non-réponse ignorable et qu'elle est « positive » sous le modèle de non-réponse non ignorable. Donc, il n'y a pour ainsi dire aucune différence entre les deux scénarios, c'est-à-dire celui où seules les données provenant des cas complets sont utilisées et celui où toutes les données sont utilisées. En outre, d'après le facteur de Bayes et notre étude par simulation, bien qu'il existe des différences entre le modèle de non-réponse ignorable et le modèle de non-réponse non ignorable, elles sont faibles. Nous constatons des différences en ce qui concerne l'inférence au sujet des proportions d'individus à divers niveaux DMO-RF; les moyennes a posteriori sont semblables, mais les écarts-types a posteriori sont plus grands sous le modèle de non-réponse non ignorable que sous le modèle de non-réponse ignorable.

Notre étude par simulation confirme deux propriétés (différences subtiles) de nos modèles. En premier lieu, les estimations des probabilités de cellule d'après le modèle de non-réponse ignorable (non ignorable) s'approchent plus des valeurs réelles quand il est attendu que le modèle de non-réponse ignorable (non ignorable) sera vérifié, mais, dans l'un et l'autre cas, l'écart-type des estimations d'après le modèle de non-réponse non ignorable est environ deux fois plus grand que celui des estimations d'après le modèle de non-réponse ignorable. En deuxième lieu, si l'on s'attend à ce que le modèle de non-réponse ignorable (non ignorable) tienne, celui-ci peut donner de moins bons résultats que le modèle de non-réponse non ignorable (ignorable). Cela se produit un pourcentage significativement plus élevé de fois dans le cas où l'on s'attend à ce que le modèle de non-réponse ignorable soit vérifié que dans l'autre. Donc, il existe des différences entre ces modèles. Nous suggérons d'ajuster les deux modèles et de calculer le facteur de Bayes pour décider lequel il convient d'utiliser. Nous ne recommandons pas d'utiliser ces modèles lorsqu'il existe des covariables et (ou) des données a priori appropriées pour expliquer la non-ignorabilité.

Lors de futurs travaux, nous pourrions essayer de réduire le nombre de paramètres du modèle de non-réponse non ignorable afin de réduire davantage les effets de la non-ignorabilité. Ainsi, nous pourrions envisager de représenter les données dans deux tableaux de contingence comme il suit. Les trois tableaux supplémentaires sont regroupés en un tableau supplémentaire unique dont la j^{e} ligne contient *au moins* u_j individus et la k^{e} colonne, *au moins* v_k individus; le nombre total d'individus dans ce tableau supplémentaire est $w+\sum_{j=1}^{r} u_j + \sum_{k=1}^{c} v_k$; voir la section 3.1 pour la notation. Enfin, il convient de souligner que l'analyse complète des données provenant d'une enquête complexe nécessite une apport d'information (covariables et information a priori) au sujet de la non-ignorabilité, des poids de sondage et des effets de mise en grappes également.

Annexe A
Ajustement du modèle de non-réponse non ignorable

Nous montrons comment utiliser l'échantillonneur de Gibbs pour faire une inférence au sujet des paramètres de (14). La densité a posteriori conditionnelle de p est

$$\boldsymbol{p} \mid \boldsymbol{y} \sim \text{Dirichlet}(y_{\cdot 11}+1, \ldots, y_{\cdot rc}+1) \qquad \text{(A.1)}$$

et la densité a posteriori conditionnelle de $\boldsymbol{\pi}_{jk}$ est

$$\boldsymbol{\pi}_{jk} \mid \{\boldsymbol{\mu}, \tau, \boldsymbol{y}\} \overset{\text{ind}}{\sim} \text{Dirichlet}\begin{pmatrix} y_{1jk}+\mu_1\tau, y_{2jk} \\ +\mu_2\tau, y_{3jk}+\mu_3\tau, y_{4jk}+\mu_4\tau \end{pmatrix} \qquad \text{(A.2)}$$

avec indépendance sur $j=1, \ldots, r, k=1, \ldots, c$.

Nous avons besoin des fonctions de masse de probabilité a posteriori conditionnelles de $\boldsymbol{y}_s, s=2, 3, 4$ sachant $\boldsymbol{y}_{(s)}, \boldsymbol{p}, \boldsymbol{\pi}_{jk}, j=1, \ldots, r, \quad k=1, \ldots, c$. D'après (14), il est clair que les $\boldsymbol{y}_s, s=2, 3, 4$ sont des vecteurs aléatoires multinomiaux conditionnellement indépendants. Plus précisément,

$$\boldsymbol{y}_{2j} \mid \{\boldsymbol{y}_1, \boldsymbol{p}, \boldsymbol{\pi}_{jk}, j=1, \ldots, r, k=1, \ldots, c\} \overset{\text{ind}}{\sim} \text{Multinomiale}(u_j, \boldsymbol{q}_j^{(2)}), \; j=1, \ldots, r,$$

$$\boldsymbol{y}_{3k} \mid \{\boldsymbol{y}_1, \boldsymbol{p}, \boldsymbol{\pi}_{jk}, j=1, \ldots, r, k=1, \ldots, c\} \overset{\text{ind}}{\sim} \text{Multinomiale}(v_k, \boldsymbol{q}_k^{(3)}), k=1, \ldots, c,$$

$$\boldsymbol{y}_4 \mid \{\boldsymbol{y}_1, \boldsymbol{p}, \boldsymbol{\pi}_{jk}, j=1, \ldots, r, k=1, \ldots, c\} \sim \text{Multinomial}(w, \boldsymbol{q}^{(4)}), \qquad \text{(A.3)}$$

où $q_{jk}^{(2)} = \pi_{2jk} p_{jk} / \sum_{k'=1}^{c} \pi_{2jk'} p_{jk'}, k=1, \ldots, c, q_{jk}^{(3)} = \pi_{3jk} p_{jk} / \sum_{j'=1}^{r} \pi_{3j'k} p_{j'k}, \; j=1, \ldots, r,$ et $q_{jk}^{(4)} = \pi_{4jk} p_{jk} / \sum_{j'=1}^{r} \sum_{k'=1}^{c} \pi_{4j'k'} p_{j'k'}, j=1, \ldots, r, k=1, \ldots, c.$

Puis, nous considérons les hyperparamètres. Si nous posons que $\delta_s = \prod_{j=1}^{r} \prod_{k=1}^{c} \pi_{sjk}$, la densité a posteriori conditionnelle conjointe de $\boldsymbol{\mu}, \tau$ est

$$p(\boldsymbol{\mu}, \tau \mid \boldsymbol{\pi}_{jk}, j=1, \ldots, r, k=1, \ldots, c\} \propto \left[\left\{ \prod_{s=1}^{4} \delta_s^{\mu_s \tau} \right\} / \{D(\boldsymbol{\mu}\tau)\}^{rc} \right] \tau^{\alpha_0 - 1} e^{-\beta_0 \tau}, \qquad \text{(A.4)}$$

où $\sum_{s=1}^{4} \mu_s = 1, \mu_s \geq 0, s=1, 2, 3, 4, \tau > 0$.

Nous utilisons la méthode du quadrillage pour obtenir des échantillons à partir de la densité a posteriori conditionnelle de $p(\boldsymbol{\mu} \mid \tau, \boldsymbol{\pi}_{jk}, j=1, \ldots, r, \quad k=1, \ldots, c)$ et $p(\tau \mid \boldsymbol{\mu}, \boldsymbol{\pi}_{jk}, j=1, \ldots, r, \; k=1, \ldots, c)$. Après transformation de τ en

$\varphi/(1-\varphi)$, les paramètres résident sur $(0,1)$ avec les contraintes appropriées, ce qui rend la méthode du quadrillage commode. Nous utilisons 50 intervalles de même largeur (obtenus par expérimentation) pour tirer $\boldsymbol{\mu}$ et φ, et une valeur aléatoire pour τ est $\varphi/(1-\varphi)$.

Nous exécutons l'échantillonneur de Gibbs en tirant une valeur aléatoire de chacune des « densités » a posteriori conditionnelles, (A.1), (A.2), (A.3) et (A.4) l'une après l'autre, et en itérant la procédure complète jusqu'à la convergence. Il s'agit d'un exemple d'échantillonnage « griddy Gibbs » (Ritter et Tanner 1992).

Annexe B
Estimation de $p_{\text{NIG}}(\boldsymbol{y}_1)$ dans (16)

En notant n_m le nombre de cas incomplets (c'est-à-dire $n = n_0 + n_m$), nous pouvons aussi montrer que, pour le modèle avec association, $p_{\text{NIG}}(\boldsymbol{y}_1) = a((n+1)!)/(n_0! n_m!)A$ et pour le modèle sans association, $p_{\text{NIG}}(\boldsymbol{y}_1) = b((n+1)!/(n_0! n_m!))B$, où a et b sont donnés par (18),

$$A = \int_{\Omega_a} \left\{ \prod_{j,k} \pi_{1jk}^{y_{1jk}} \right\} \left\{ \sum_{s=2}^{4} \sum_{j,k} \pi_{sjk}\, p_{jk} \right\}^{n_m} \left\{ \frac{\prod_{j,k} p_{jk}^{y_{1jk}}}{D(y_{111}+1,\ldots,y_{1rc}+1)} \right\} \times \prod_{j,k} \left\{ \frac{\prod_s \pi_{sjk}^{\mu_s \tau - 1}}{D(\boldsymbol{\mu}\tau)} \right\} \frac{\beta_0^{\alpha_0} \tau^{\alpha_0 - 1} e^{-\beta_0 \tau}}{\Gamma(\alpha_0)} d\Omega_a ,$$

$$B = \int_{\Omega_{na}} \left\{ \prod_{j,k} \pi_{1jk}^{y_{1jk}} \right\} \left\{ \sum_{s=2}^{4} \sum_{j,k} \pi_{sjk}\, q_{1j}\, q_{2k} \right\}^{n_m} \frac{\prod_j q_{1j}^{y_{1j}}}{D(y_{11.}+1,\ldots,y_{1r.}+1)} \times \frac{\prod_k q_{2k}^{y_{1jk}}}{D(y_{1.1}+1,\ldots,y_{1.c}+1)} \prod_{j,k} \left\{ \frac{\prod_s \pi_{sjk}^{\mu_s \tau - 1}}{D(\boldsymbol{\mu}\tau)} \right\} \frac{\beta_0^{\alpha_0} \tau^{\alpha_0 - 1} e^{-\beta_0 \tau}}{\Gamma(\alpha_0)} d\Omega_{na} . \qquad \text{(B.1)}$$

Notons que $0 < A, B < 1$ donne une vérification diagnostique utile des calculs.

Nous montrons comment calculer A dans (B.1) par la méthode d'intégration de Monte Carlo; la méthode pour calculer B est semblable. Nous préférons la méthode plus simple fondée sur l'intégration par la méthode de Monte Carlo avec une fonction d'importance (Nandram et Kim, 2002) à celle fondée sur une continuation de l'échantillonneur de Gibbs (Chib et Jeliazkov 2001).

Pour A, nous choisissons la fonction d'importance

$$\pi_{im}(\Omega_a) = \frac{\prod_{j,k} p_{jk}^{y_{1jk}}}{D(y_{111}+1,\ldots,y_{1rc}+1)} \prod_{j,k} \left[\frac{\prod_s \pi_{sjk}^{\mu_s \tau - 1}}{D(\boldsymbol{\mu}\tau)} \right] \frac{\prod_{s=1}^{4} \mu_s^{\tilde{\mu}_1 \tilde{\tau} - 1}}{D(\tilde{\boldsymbol{\mu}}\tilde{\tau})} \frac{\beta_0^{\alpha_0} \tau^{\alpha_0 - 1} e^{-\beta_0 \tau}}{\Gamma(\alpha_0)}$$

où $\tilde{\mu}_s$ et $\tilde{\tau}$ sont les estimations obtenues au moyen d'une sortie de Gibbs. Nous obtenons un échantillon de $\pi_{im}(\Omega_a)$ en le tirant à partir de $\tau \sim \text{Gamma}(\alpha_0, \beta_0)$, $\boldsymbol{\mu} \overset{\text{ind}}{\sim} \text{Dirichlet}(\tilde{\boldsymbol{\mu}}, \tilde{\tau})$, $\boldsymbol{\pi}_{jk} | \boldsymbol{\mu}, \tau \sim \text{Dirichlet}(\boldsymbol{\mu}, \tau)$ et $\boldsymbol{p} | \boldsymbol{y}_1 \sim \text{Dirichlet}(y_{111}+1, \ldots, y_{1rc}+1)$.

Puis, en posant que $w_h = \sum_j^r \sum_k^c y_{1jk} \log \pi_{1jk}^{(h)} + n_m \log[\sum_{s=2}^4 \sum_j^r \sum_k^c \pi_{sjk}^{(h)} p_{jk}^{(h)}] - \sum_{s=1}^4 (\tilde{\mu}_s \tilde{\tau} - 1) + \log \mu_s^{(h)} + \log(D(\tilde{\boldsymbol{\mu}}\tilde{\tau}))$, $h = 1, \ldots, M$, un estimateur de A est $\hat{A} = M^{-1} \sum_{h=1}^M e^{\omega_h}$. L'erreur-type numérique (ETN) de $\log(\hat{A})$ peut être approximée. En posant que $\overline{\omega} = M^{-1} \sum_{h=1}^M \omega_h$ et $S^2 = (M-1)^{-1} \sum_{k=1}^M (\omega_h - \overline{\omega})^2$, nous avons $\text{Var}(\hat{A}) \approx e^{2\overline{\omega}} S^2 / M$, $\text{Var}(\log(\hat{A})) \approx (\text{Var}(\hat{A}) / e^{2\overline{\omega}}) \approx S^2 / M$, et l'ETN est égale à $S/\sqrt{M}$, approximativement. Nous débutons avec $M = 10\,000$ échantillons indépendants provenant de la fonction d'importance et augmentons M jusqu'à l'obtention de la convergence, pour $M = 50\,000$ environ.

Remerciements

La matière présentée ici fait partie des travaux réalisés au cours de l'année universitaire 2003–2004 durant laquelle Balgobin Nandram était en congé sabbatique à titre de chercheur au National Center for Health Statistics, à Hyattsville, au Maryland. Nous remercions le rédacteur adjoint et les deux examinateurs de leurs commentaires constructifs et des trois occasions que nous avons eues de réviser le manuscrit.

Bibliographie

Chen, T., et Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 30, 629-642.

Chib, S., et Jeliazkov, I. (2001). Marginal likelihood from the metropolis-hastings output. *Journal of the American Statistical Association*, 96, 270-281.

Cohen, G., et Duffy, J.C. (2002). Are nonrespondents to health surveys less healthy than respondents. *Journal of Official Statistics*, 18, 13-23.

Draper, D. (1995). Assessment and propagation of model uncertainty (avec discussion). *Journal of the Royal Statistical Society*, Series B, 57, 45-97.

Farahmand, B.Y., Persson, P.G., Michaelsson, K., Baron, J.A., Parker, M.G. et Ljunghall, S. (2000). Socioeconomic status, marital status and hip fracture risk: A population-based case control study. *Osteoporosis International*, 11, 803-808.

Forster, J.J., et Smith, P.W.F. (1998). Model-based inference for categorical survey data subject to non-ignorable nonresponse. *Journal of the Royal Statistical Society*, Series B, 60, 57-70.

Ganry, O., Baudoin, C. et Fardellone, P. (2000). Effect of alcohol intake on bone mineral density in elderly women: The EPIDOS Study. *American Journal of Epidemiology*, 151, 8, 773-780.

Kass, R., et Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.

Kalton, G., et Kasprzyk, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.

Lauderdale, D.S., et Rathouz, P.J. (2003). Does bone mineralization reflect economic conditions? An examination using a national US sample. *Economics and Human Biology*, 1, 91-104.

Little, R.J. (2003). Bayesian Approach to Sample Survey Inference. Dans *Analysis of Survey Data*, (Éds. R.L. Chambers et C.J. Skinner), New York: John Wiley & Sons, Inc., 289-306.

Little, R.J.A., et Rubin D.B. (2002). *Statistical Analysis with Missing Data.* Édition, New York: John Wiley & Sons, Inc.

Looker, A.C., Orwoll, E.S., Johnston, C.C., Lindsay, R.L., Wahner, H.W., Dunn, W., Calvo, M.S. et Harris, T.B. (1997). Prevalence of low femoral bone density in older U.S. adults from NHANES III. *Journal of Bone and Mineral Research*, 12, 1761-1768.

Looker, A.C., Wahner, H.W., Dunn, W.L., Calvo, M.S., Harris, R.R., Heyse, S.P., Johnston, C.C. et Lindsay, R. (1998). Updated data on proximal femur bone mineral levels of us adults. *Osteoporosis International*, 8, 468-489.

Mirkin, B. (2001). Eleven ways to look at the chi-squared coefficient for contingency tables. *The American Statistician*, 55, 111-120.

Nandram, B., et Choi, J.W. (2002 a). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.

Nandram, B., et Choi, J.W. (2002 b). A Bayesian analysis of a proportion under nonignorable nonresponse. *Statistics in Medicine*, 21, 1189-1212.

Nandram, B., et Choi, J.W. (2005). Modèles de régression hiérarchiques bayésiens sous non-réponse non-ignorable pour petits domaines : Une application aux données de la NHANES. *Techniques d'enquête*, 31, 79-92.

Nandram, B., Han, G. et Choi, J.W. (2002). Un modèle bayesien hiérarchique de non-réponse non-ignorable pour les données multinomiales des petites régions. *Techniques d'enquête*, 28, 157-170.

Nandram, B., et Kim, H. (2002). Marginal likelihood for a class of Bayesian generalized linear models. *Journal of Statistical Computation and Simulation*, 72, 319-340.

Nandram, B., Liu, N., Choi, J.W. et Cox, L.H. (2005). Bayesian nonresponse models for categorical data from small areas: An application to BMD and age. *Statistics in Medicine*, 24, 1047-1074.

Rao, J.N.K., et Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

Rao, J.N.K., et Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12, 46-60.

Ritter, C., et Tanner, M.A. (1992). The Gibbs stopper and the griddy Gibbs sampler. *Journal of the American Statistical Association*, 87, 861-868.

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

Rubin, D.B., Stern, H.S. et Vehovar, V. (1995). Handling "Don't know" survey responses: The case of the slovenian plebiscite. *Journal of the American Statistical Association*, 90, 822-828.

Sinharay, S., et Stern, H.S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, 56, 196-201.

Wang, H. (2001). *Two-way Contingency Tables with Marginally and Conditionally Imputed Nonrespondents*, Thèse de doctorat, *Department of Statistics*, University of Wisconsin-Madison.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2005
Vol. 31, N° 2, pp. 249-254
Statistique Canada, N° 12-001-XPB au catalogue

L'utilisation de renseignements sur le processus de collecte des données pour traiter la non-réponse totale au moyen de l'ajustement de poids

Jean-François Beaumont [1]

Résumé

On utilise couramment l'ajustement de poids pour la non-réponse afin de compenser la non-réponse totale aux enquêtes. Souvent, on postule un modèle de non-réponse et on ajuste les poids de sondage par l'inverse de probabilités de réponse estimées. Le modèle de non-réponse est habituellement conditionnel à un vecteur de variables auxiliaires fixes qui sont observées pour chaque unité de l'échantillon, comme les variables utilisées pour construire le plan de sondage. Dans le présent article, nous envisageons d'utiliser comme variables auxiliaires éventuelles les variables du processus de collecte des données. Le nombre d'essais effectués pour joindre une unité de l'échantillon en constitue un exemple. Dans notre traitement, ces variables auxiliaires sont considérées comme aléatoires, même si on conditionne sur l'échantillon sélectionné, puisqu'elles pourraient changer si le processus de collecte des données était répété pour un échantillon donné. Nous montrons que ce caractère aléatoire n'introduit ni biais ni composante additionnelle de variance dans les estimations de totaux de population lorsque le modèle de non-réponse est bien spécifié. En outre, lorsque la non-réponse dépend des variables d'intérêt, nous soutenons que l'utilisation des variables du processus de collecte des données a tendance à réduire le biais de non-réponse si ces variables fournissent des renseignements sur les variables d'intérêt qui ne sont pas déjà inclus dans le modèle de non-réponse et si elles sont associées à la non-réponse. Par conséquent, les variables du processus de collecte des données pourraient bien être très utiles pour traiter la non-réponse totale. Nous en donnons une brève illustration à partir de l'Enquête sur la population active du Canada.

Mots clés : Biais de non-réponse; modèle de non-réponse; variance due à la non-réponse; nombre d'essais; paradonnées; probabilité de réponse.

1. Introduction

Dans les enquêtes, on traite souvent la non-réponse totale en utilisant une méthode d'ajustement de poids pour la non-réponse. Le principe de base qu'on choisit souvent consiste à ajuster les poids de sondage par l'inverse de probabilités de réponse estimées (voir, par exemple, Ekholm et Laaksonen 1991). On obtient ces probabilités de réponse estimées en postulant un modèle pour le mécanisme de non-réponse inconnu, que nous appelons le modèle de non-réponse. Pour réduire dans toute la mesure du possible le biais et la variance dus à la non-réponse, il est essentiel de conditionner sur un vecteur de variables auxiliaires qui sont observées pour chaque unité de l'échantillon et qui sont de bons prédicteurs de la non-réponse et des variables d'intérêt (Little et Vartivarian 2005). On traite habituellement les variables auxiliaires comme des variables fixes, que ce soit conditionnellement ou non à l'échantillon sélectionné.

Dans le présent article, nous envisageons d'utiliser les variables du processus de collecte des données (PCD) comme variables auxiliaires éventuelles à inclure dans le modèle de non-réponse. Le nombre d'essais effectués pour joindre une unité de l'échantillon en constitue un exemple. Ce type de données est parfois appelé paradonnées (voir Couper et Lyberg 2005 pour une référence récente sur le sujet); Holt et Elliott (1991), entre autres, l'ont utilisé pour composer avec la non-réponse totale. Dans notre traitement, contrairement à Holt et Elliott (1991), les variables du PCD sont considérées comme aléatoires, même si on les conditionne sur l'échantillon sélectionné, puisqu'elles pourraient changer si le processus de collecte des données était répété pour un échantillon donné.

Les variables du PCD peuvent s'avérer particulièrement utiles dans les enquêtes transversales où les variables auxiliaires dont on dispose pour traiter la non-réponse totale se limitent souvent aux variables utilisées pour construire le plan de sondage. Sans être inutiles, ces variables du plan de sondage ne sont souvent pas de bons prédicteurs de la non-réponse et des variables d'intérêt. Dans ce cas, les renseignements additionnels tirés du processus de collecte des données peuvent être les bienvenus. Dans les enquêtes longitudinales, on trouve une foule de variables auxiliaires éventuelles pour composer avec la non-réponse de vague. Les renseignements sur le PCD peuvent donc s'avérer moins importants pour compenser la non-réponse de vague que pour compenser la non-réponse totale dans les enquêtes transversales, mais nous n'avons pas encore étudié cet aspect en profondeur. Il se pourrait qu'aux points de changement, les variables du PCD jouent un rôle important.

Dans la section 2, nous présentons la notation et notre théorie concernant l'effet de l'utilisation de variables auxiliaires aléatoires dans le modèle de non-réponse lorsqu'on

1. Jean-François Beaumont, Division des méthodes d'enquête auprès des entreprises, Statistique Canada, 11e étage, Immeuble R.-H.-Coats, Ottawa (Ontario), Canada, K1A 0T6. Courriel : jean-francois.beaumont@statcan.ca.

estime des totaux de population. Cette question du caractère aléatoire des variables auxiliaires du PCD a été soulevée et débattue au sein du Comité consultatif sur les méthodes statistiques de Statistique Canada à la suite de la présentation de l'article d'Alavi et Beaumont (2004). L'objet de la section 2 consiste donc à éclairer cette question. L'utilisation des variables du PCD pour ajuster les poids de sondage pour la non-réponse est illustrée brièvement dans la section 3, au moyen de l'Enquête sur la population active (EPA) du Canada. La dernière section, soit la section 4, présente un bref résumé de l'article.

2. Théorie

Supposons que nous voulions estimer le total de population $t_y = \sum_{k \in U} y_k$ d'une variable d'intérêt y pour une certaine population fixe U de taille N. De cette population, on sélectionne un échantillon aléatoire s de taille n selon un plan d'échantillonnage probabiliste $p(s|\mathbf{D})$, où $\mathbf{D}$ est une matrice à N lignes contenant $\mathbf{d}_k'$ dans sa k^{e} ligne et $\mathbf{d}$ est le vecteur des variables du plan de sondage. Supposons également qu'en l'absence de non-réponse, nous utiliserions l'estimateur de Horvitz-Thompson $\hat{t}_y = \sum_{k \in s} w_k y_k$, où $w_k = 1/\pi_k$ est le poids de sondage de l'unité k et $\pi_k = P(k \in s)$ est sa probabilité de sélection.

Habituellement, pour un certain nombre de raisons, la non-réponse totale survient de sorte qu'on observe la variable y uniquement pour un sous-ensemble s_r de s, c'est-à-dire les répondants. Outre s_r, on observe également un vecteur aléatoire $\mathbf{z}$ de variables du PCD pour chaque unité de l'échantillon, selon un mécanisme conjoint $\#q(\mathbf{Z}_s, s_r \mid s, \mathbf{Y}, \mathbf{D}, \mathbf{X})$. Comme nous l'avons mentionné dans l'introduction, le nombre d'essais effectués pour joindre une unité de l'échantillon constitue un exemple de variable du PCD. Le vecteur $\mathbf{z}$ des variables du PCD et l'ensemble de répondants s_r sont aléatoires après avoir conditionné sur l'échantillon sélectionné puisque ces quantités prendraient probablement des valeurs différentes si le processus de collecte des données était répété pour un échantillon donné. La quantité $\mathbf{Z}_s$ est une matrice à n lignes contenant $\mathbf{z}_k'$ dans sa k^{e} ligne, $\mathbf{Y}$ est un vecteur à N éléments contenant y_k dans son k^{e} élément et $\mathbf{X}$ est une matrice à N lignes contenant $\mathbf{x}_k'$ dans sa k^{e} ligne. Le vecteur $\mathbf{x}$ est un vecteur de variables auxiliaires fixes additionnelles. Par exemple, ces variables auxiliaires pourraient provenir d'un fichier administratif ou, dans le cas d'une enquête longitudinale, il pourrait s'agir des variables d'intérêt observées au moment de la vague précédente. Par conséquent, on ne dispose pas nécessairement du vecteur $\mathbf{x}$ pour les unités non échantillonnées. Le tableau 1 résume la disponibilité des différents types de variables pour les répondants, les non-répondants et les unités non échantillonnées.

Tableau 1
Disponibilité des variables

	y	z	x	d
Répondants : s_r	OUI	OUI	OUI	OUI
Non-répondants : $s - s_r$	NON	OUI	OUI	OUI
Unités non échantillonnées : $U - s$	NON	NON*	OUI**	OUI

* Le vecteur **z** n'est même pas défini pour les unités non échantillonnées.
** Le vecteur **x** peut ne pas toujours être disponible pour les unités non échantillonnées.

On peut factoriser le mécanisme conjoint $\#q(\mathbf{Z}_s, s_r \mid s, \mathbf{Y}, \mathbf{D}, \mathbf{X})$ en deux mécanismes aléatoires distincts : i) $\#(\mathbf{Z}_s \mid s, \mathbf{Y}, \mathbf{D}, \mathbf{X})$ et ii) $q(s_r \mid s, \mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{Z}_s)$. Le premier est appelé mécanisme du PCD et le deuxième, mécanisme de non-réponse. Cette factorisation nous permettra plus tard d'obtenir les propriétés de notre estimateur aux poids ajustés pour la non-réponse, défini dans l'équation (2.2) ci-dessous. Nous supposons que

$$q(s_r \mid s, \mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{Z}_s) = q(s_r \mid s, \mathbf{D}_s, \mathbf{X}_s, \mathbf{Z}_s), \quad (2.1)$$

où $\mathbf{D}_s$ et $\mathbf{X}_s$ sont, respectivement, les parties correspondant à l'échantillon de $\mathbf{D}$ et de $\mathbf{X}$. Cette hypothèse implique que le mécanisme de non-réponse est indépendant de $\mathbf{Y}$ (ou non confondu avec $\mathbf{Y}$) après avoir conditionné sur s, $\mathbf{D}_s$, $\mathbf{X}_s$ et $\mathbf{Z}_s$ et que les données sont manquantes au hasard. Toutefois, nous ne formulons explicitement aucune hypothèse simplificatrice au sujet du mécanisme du PCD, de sorte qu'il pourrait bien dépendre de $\mathbf{Y}$, même après avoir conditionné sur s, $\mathbf{D}$ et $\mathbf{X}$.

Pour compenser la non-réponse totale, nous considérons l'estimateur aux poids ajustés pour la non-réponse

$$\hat{t}_y^{\text{NWA}} = \sum_{k \in s_r} \frac{w_k}{p_k(\hat{\boldsymbol{\alpha}})} y_k, \quad (2.2)$$

où $p_k(\boldsymbol{\alpha}) = P(k \in s_r \mid s, \mathbf{D}_s, \mathbf{X}_s, \mathbf{Z}_s; \boldsymbol{\alpha})$ est la probabilité de réponse conditionnelle pour une unité $k \in s$ et $\hat{\boldsymbol{\alpha}}$ est un estimateur du vecteur des paramètres inconnus du modèle de non-réponse $\boldsymbol{\alpha}$. Il est à noter qu'un modèle de non-réponse est un ensemble d'hypothèses relatives au mécanisme de non-réponse inconnu $q(s_r \mid s, \mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{Z}_s)$; l'une d'elles est l'hypothèse (2.1). Nous supposons que $\hat{\boldsymbol{\alpha}}$ est défini implicitement par l'équation $\mathbf{U}_1(\hat{\boldsymbol{\alpha}}) = \mathbf{0}$, où $\mathbf{U}_1(.)$ est un vecteur de fonctions d'estimation sans biais par rapport à q pour $\boldsymbol{\alpha}$; ainsi, $\mathbf{E}_q\{\mathbf{U}_1(\boldsymbol{\alpha}) \mid s, \mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{Z}_s\} = \mathbf{0}$. Donc, $\mathbf{U}_1(.)$ est également sans biais par rapport à $p\#q$ pour $\boldsymbol{\alpha}$. Dans le reste de l'article, nous supprimons partout le conditionnement sur $\mathbf{Y}$, $\mathbf{D}$ et $\mathbf{X}$ quand on prend les espérances et les variances, puisque ces vecteurs sont toujours considérés comme fixes. Par exemple, nous écrivons $\mathbf{E}_q\{\mathbf{U}_1(\boldsymbol{\alpha}) \mid s, \mathbf{Z}_s\} = \mathbf{0}$ au lieu de $\mathbf{E}_q\{\mathbf{U}_1(\boldsymbol{\alpha}) \mid s, \mathbf{Y}, \mathbf{D}, \mathbf{X}, \mathbf{Z}_s\} = 0$. Ceci simplifie considérablement la notation.

Il est à noter que l'estimateur aux poids ajustés pour la non-réponse (2.2) est défini implicitement par l'équation

$$U_2(\hat{\boldsymbol{\alpha}}, \hat{t}_y^{\mathrm{NWA}}) = \hat{t}_y^{\mathrm{NWA}} - \sum_{k \in s_r} \frac{w_k}{p_k(\hat{\boldsymbol{\alpha}})} y_k = 0. \qquad (2.3)$$

Si le modèle de non-réponse est spécifié correctement et, surtout, si l'hypothèse (2.1) est satisfaite, la fonction d'estimation $U_2(.,.)$ est alors sans biais par rapport à $p\#q$ pour t_y; ainsi, $\mathrm{E}_{p\#q}\{U_2(\boldsymbol{\alpha}, t_y)\} = 0$. Pour rendre l'hypothèse (2.1) aussi plausible que possible, il importe que le modèle de non-réponse soit conditionnel aux variables du plan de sondage, aux variables auxiliaires et aux variables du PCD qui sont bien corrélées avec y, pourvu que ces variables soient également associées à la non-réponse. Cette recommandation devrait être utile pour contrôler l'ampleur du biais de non-réponse, qui peut être inévitable dans une enquête réelle. Elle est aussi compatible avec la recommandation formulée par Little et Vartivarian (2005). Donc, si les variables du PCD contiennent des renseignements sur y au-delà des renseignements déjà contenus dans $\mathbf{d}$ et $\mathbf{x}$, l'utilisation des variables du PCD peut alors s'avérer utile pour réduire le biais de non-réponse si ces variables sont associées à la non-réponse.

Posons maintenant $\boldsymbol{\theta} = (\boldsymbol{\alpha}', t_y)'$, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}', \hat{t}_y^{\mathrm{NWA}})'$ et $\mathbf{U}(\tilde{\boldsymbol{\theta}}) = \{\mathbf{U}_1'(\tilde{\boldsymbol{\alpha}}), U_2(\tilde{\boldsymbol{\alpha}}, \tilde{t}_y)\}'$, pour un certain vecteur $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\alpha}}', \tilde{t}_y)'$. Comme nous l'avons mentionné plus haut, $\hat{\boldsymbol{\theta}}$ est défini implicitement par l'équation $\mathbf{U}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ et la fonction d'estimation $\mathbf{U}(.)$ est sans biais par rapport à $p\#q$ pour $\boldsymbol{\theta}$ puisque $\mathbf{E}_{p\#q}\{\mathbf{U}(\boldsymbol{\theta})\} = \mathbf{0}$. En utilisant une approximation de Taylor du premier degré (voir Binder 1983), nous avons $\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta} - \{\mathbf{H}(\boldsymbol{\theta})\}^{-1}\mathbf{U}(\boldsymbol{\theta})$, où $\mathbf{H}(\tilde{\boldsymbol{\theta}}) = \mathbf{E}_{p\#q}\{\partial\mathbf{U}(\tilde{\boldsymbol{\theta}})/\partial\tilde{\boldsymbol{\theta}}'\}$. La matrice $\{\mathbf{H}(\boldsymbol{\theta})\}^{-1}$ est donc donnée par

$$\{\mathbf{H}(\boldsymbol{\theta})\}^{-1} = \begin{pmatrix} \{\mathbf{H}_{11}(\boldsymbol{\theta})\}^{-1} & \mathbf{0} \\ -\mathbf{H}_{21}(\boldsymbol{\theta})\{\mathbf{H}_{11}(\boldsymbol{\theta})\}^{-1} & 1 \end{pmatrix}, \qquad (2.4)$$

où $\mathbf{H}_{i1}(\tilde{\boldsymbol{\theta}}) = \mathbf{E}_{p\#q}(\partial\mathbf{U}_i(\tilde{\boldsymbol{\theta}}))/(\partial\tilde{\boldsymbol{\alpha}}')$, pour $i = 1, 2$. En utilisant des conditions semblables à celles de Binder (1983), $\hat{\boldsymbol{\theta}}$ est asymptotiquement normal et asymptotiquement sans biais par rapport à $p\#q$ pour $\boldsymbol{\theta}$. Par conséquent, $\hat{t}_y^{\mathrm{NWA}}$ est asymptotiquement normal et asymptotiquement sans biais par rapport à $p\#q$ pour t_y. Donc, l'utilisation des variables du PCD dans le modèle de non-réponse n'introduit aucun biais dans l'estimateur aux poids ajustés pour la non-réponse $\hat{t}_y^{\mathrm{NWA}}$ pourvu que le modèle de non-réponse (spécification de $q(s_r \mid s, \mathbf{D}_s, \mathbf{X}_s, \mathbf{Z}_s)$ et l'hypothèse 2.1) soit valable. De plus, si le vrai mécanisme de non-réponse inconnu dépend de la partie correspondant à l'échantillon de $\mathbf{Y}$, $\mathbf{Y}_s$, après avoir conditionné sur s, $\mathbf{D}_s$ et $\mathbf{X}_s$, le conditionnement sur un vecteur $\mathbf{z}$ de variables du PCD aura tendance à réduire le biais de non-réponse si le mécanisme du PCD dépend de $\mathbf{Y}_s$, après avoir conditionné sur s, $\mathbf{D}_s$ et $\mathbf{X}_s$, ce qui signifie que les variables du PCD contiennent des renseignements sur y qui ne sont pas déjà contenus dans $\mathbf{d}$ et $\mathbf{x}$.

En poursuivant notre linéarisation de Taylor et en utilisant le fait que

$$\begin{aligned}\mathbf{V}_{p\#q}\{\mathbf{U}(\boldsymbol{\theta})\} = {} & \mathbf{V}_p\mathbf{E}_{\#q}\{\mathbf{U}(\boldsymbol{\theta}) \mid s\} \\ & + \mathbf{E}_p\mathbf{V}_{\#}\mathbf{E}_q\{\mathbf{U}(\boldsymbol{\theta}) \mid s, \mathbf{Z}_s\} \\ & + \mathbf{E}_{p\#}\mathbf{V}_q\{\mathbf{U}(\boldsymbol{\theta}) \mid s, \mathbf{Z}_s\},\end{aligned}$$

la matrice de variances-covariances par rapport à $p\#q$ de $\hat{\boldsymbol{\theta}}$, $\mathbf{V}_{p\#q}(\hat{\boldsymbol{\theta}})$, est approximée par

$$\begin{aligned}\dot{\mathbf{V}}_{p\#q}(\hat{\boldsymbol{\theta}}) = {} & \{\mathbf{H}(\boldsymbol{\theta})\}^{-1}\mathbf{V}_p\mathbf{E}_{\#q}\{\mathbf{U}(\boldsymbol{\theta}) \mid s\}\{\mathbf{H}'(\boldsymbol{\theta})\}^{-1} \\ & + \{\mathbf{H}(\boldsymbol{\theta})\}^{-1}\mathbf{E}_p\mathbf{V}_{\#}\mathbf{E}_q\{\mathbf{U}(\boldsymbol{\theta}) \mid s, \mathbf{Z}_s\}\{\mathbf{H}'(\boldsymbol{\theta})\}^{-1} \\ & + \{\mathbf{H}(\boldsymbol{\theta})\}^{-1}\mathbf{E}_{p\#}\mathbf{V}_q\{\mathbf{U}(\boldsymbol{\theta}) \mid s, \mathbf{Z}_s\}\{\mathbf{H}'(\boldsymbol{\theta})\}^{-1}. \qquad (2.5)\end{aligned}$$

Le premier terme du membre droit de l'équation (2.5) est appelé la variance d'échantillonnage de $\hat{\boldsymbol{\theta}}$, le deuxième, la variance du PCD de $\hat{\boldsymbol{\theta}}$ et le troisième, la variance due à la non-réponse de $\hat{\boldsymbol{\theta}}$. La variance $\mathrm{V}_{p\#q}(\hat{t}_y^{\mathrm{NWA}})$ est approximée par la valeur de la dernière ligne et de la dernière colonne de l'équation (2.5). En utilisant l'expression (2.4) et le fait que $\mathbf{E}_q\{\mathbf{U}(\boldsymbol{\theta}) \mid s, \mathbf{Z}_s\} = (\mathbf{0}', t_y - \hat{t}_y)'$, la variance approximative (2.5) se réduit à

$$\begin{aligned}\dot{\mathbf{V}}_{p\#q}(\hat{\boldsymbol{\theta}}) = {} & \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{V}_p(\hat{t}_y) \end{pmatrix} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 \end{pmatrix} \\ & + \{\mathbf{H}(\boldsymbol{\theta})\}^{-1}\mathbf{E}_{p\#}\mathbf{V}_q\{\mathbf{U}(\boldsymbol{\theta}) \mid s, \mathbf{Z}_s\}\{\mathbf{H}'(\boldsymbol{\theta})\}^{-1}. \qquad (2.6)\end{aligned}$$

La deuxième matrice du membre droit de l'équation (2.6) correspond à la variance du PCD de $\hat{\boldsymbol{\theta}}$ et contient 0 pour tous ses éléments. Donc, l'utilisation des variables aléatoires auxiliaires (du PCD) dans le modèle de non-réponse n'introduit aucun terme additionnel de variance, contrairement à l'utilisation de variables auxiliaires fixes uniquement, lorsque le modèle de non-réponse est bien spécifié. Comme les variables du PCD ont tendance à réduire le biais de non-réponse si elles sont associées à y, il semble alors avantageux d'en profiter lorsqu'on traite la non-réponse totale au moyen d'un ajustement de poids. De plus, comme l'ont souligné Little et Vartivarian (2005), l'ajout, dans le modèle de non-réponse, de variables auxiliaires associées à y a tendance à réduire la variance due à la non-réponse. On peut donc réduire l'erreur quadratique moyenne sur les deux fronts.

On peut obtenir une expression plus détaillée du terme de la variance due à la non-réponse dans l'équation (2.6), ainsi qu'un estimateur de la variance due à l'échantillonnage et à la non-réponse, de la même manière que dans Beaumont (2005). Beaumont (2005) aborde également l'effet de

l'estimation des paramètres du modèle de non-réponse sur la variance d'un estimateur d'un total de population.

3. L'exemple de l'Enquête sur la population active du Canada

L'objet de cet exemple n'est pas de présenter en détail notre analyse des données de l'Enquête sur la population active (EPA) du Canada, mais simplement de décrire certains aspects liés au choix du modèle de non-réponse et à l'estimation des probabilités de réponse. Dans cette optique, nous formulons ensuite nos principales conclusions. On trouvera dans Alavi et Beaumont (2004) des renseignements plus détaillés sur les résultats de l'analyse des données de l'EPA et sur la mise en œuvre de la nouvelle méthode, ainsi qu'une comparaison avec la méthode antérieure.

L'EPA est une enquête mensuelle menée selon un plan d'échantillonnage stratifié à plusieurs degrés (Gambino, Singh, Dufour, Kennedy et Lindeyer 1998). Les renseignements utilisés pour construire le plan de sondage et pour prélever un échantillon de logements sont essentiellement géographiques. L'échantillon est divisé en six groupes de renouvellement représentatifs et chaque logement échantillonné reste dans l'échantillon pendant six mois consécutifs. Un groupe de renouvellement contient les logements dont les membres sont interviewés pour la première fois; un autre, ceux dont les membres sont interviewés pour la deuxième fois, et ainsi de suite. C'est pourquoi, dans cinq groupes de renouvellement sur six, on trouve d'un mois à l'autre les mêmes logements échantillonnés. On utilise l'interview assistée par ordinateur pour recueillir les données d'enquête sur chaque membre des ménages sélectionnés. Ce mode de collecte permet d'obtenir une grande quantité de renseignements sur le PCD auprès des ménages répondants et non répondants.

On a considéré un modèle de non-réponse logistique pour modéliser le mécanisme de non-réponse inconnu $q(s_r \mid s, \mathbf{D}_s, \mathbf{Z}_s)$. Avec ce modèle, la probabilité de réponse inconnue pour le ménage k est exprimée par l'équation $p_k(\boldsymbol{\alpha}) = \{1 + \exp(-\boldsymbol{\alpha}'(\mathbf{zd})_k)\}^{-1}$ et l'on suppose que les ménages échantillonnés répondent indépendamment les uns des autres. Le vecteur $\mathbf{zd}$ contient les variables du PCD $\mathbf{z}$, les variables du plan de sondage fixes $\mathbf{d}$ ainsi que les interactions entre ces deux types de variables. On ne disposait d'aucun vecteur additionnel $\mathbf{x}$ de variables auxiliaires. On a utilisé deux variables du PCD : le nombre d'essais effectués pour joindre un ménage échantillonné, qu'on a divisé en cinq catégories, et le moment du dernier essai, qu'on a également divisé en cinq catégories. Les variables du plan de sondage qu'on a utilisées étaient surtout géographiques et comprenaient également l'indicateur de groupe de renouvellement. En raison d'effets éventuels dus aux intervieweurs et aux grappes, le modèle décrit ci-dessus n'est peut-être pas entièrement réaliste. On l'a utilisé pour sa simplicité et parce qu'il semblait raisonnable et supérieur à la méthode antérieure. De plus, les probabilités de réponse estimées résultant de ce modèle ont servi uniquement à produire un "score" et n'ont pas servi directement à ajuster les poids de sondage, comme il est décrit plus loin dans la présente section.

On a estimé le vecteur inconnu $\boldsymbol{\alpha}$ par la méthode du maximum de vraisemblance en utilisant la fonction d'estimation sans biais par rapport à q

$$\mathbf{U}_1(\boldsymbol{\alpha}) = \sum_{k \in s} \{r_k - p_k(\boldsymbol{\alpha})\}(\mathbf{zd})_k, \qquad (3.1)$$

où $r_k = 1$ si $k \in s_r$ et $r_k = 0$ sinon. Il est à noter qu'on n'a pas considéré de fonction d'estimation pondérée par les poids de sondage. Cette méthode concorde avec la pratique recommandée par Little et Vartivarian (2003). On peut la justifier en précisant qu'il s'agit de modéliser le mécanisme de non-réponse uniquement pour les ménages échantillonnés $k \in s$ (et non pour la population entière) et que ce mécanisme est conditionnel à s. De plus, les variables du PCD ne sont même pas définies en dehors de l'échantillon. L'utilisation de poids de sondage n'est donc pas indiquée dans ce contexte et augmente la variance de $\hat{\boldsymbol{\alpha}}$ si le modèle de non-réponse est spécifié correctement. De plus, il n'est pas évident que l'utilisation d'une fonction d'estimation pondérée par les poids de sondage apporterait systématiquement de la robustesse dans ce cas. Toutefois, il est à noter que nous ne faisons pas abstraction des renseignements sur le plan de sondage, puisqu'ils sont inclus dans le modèle de non-réponse. On peut rapprocher cette mesure de la recommandation d'inclure les renseignements sur le plan de sondage dans les modèles d'imputation (voir, par exemple, Rubin 1996).

On a effectué pour plusieurs mois une régression logistique "stepwise" afin de déterminer les variables du plan de sondage et du PCD à inclure dans le modèle de non-réponse final. Pour tous les mois considérés, la variable « nombre d'essais » a été la première à entrer dans le modèle et, donc, la plus utile pour expliquer la non-réponse. Cette variable était également corrélée avec les principales variables d'intérêt « emploi » et « chômage ». Par exemple, les personnes qui appartiennent aux ménages ayant répondu après un grand nombre d'essais, soit ceux qui sont difficiles à joindre, avaient tendance à être plus souvent occupées (voir Alavi et Beaumont 2004). Les ménages ayant fait l'objet d'un grand nombre d'essais avaient aussi tendance à être des non-répondants. Il semble donc pertinent de donner un plus grand ajustement par poids aux ménages répondants ayant fait l'objet d'un nombre d'essais élevé, puisqu'ils ont moins tendance à répondre et qu'ils sont plus portés à présenter des caractéristiques semblables à celles des non-répondants.

Le modèle de non-réponse final choisi s'ajustait dans une mesure raisonnable aux données de l'EPA pour la plupart des mois considérés, selon le test d'adéquation de Hosmer-Lemeshow. Néanmoins, on a utilisé la méthode des "scores" de Little (1986) pour obtenir une certaine robustesse contre des défaillances non détectées du modèle. On a d'abord utilisé le modèle de non-réponse logistique décrit ci-dessus pour obtenir une probabilité de réponse estimée pour chaque ménage échantillonné, puis on a divisé l'échantillon en une cinquantaine de classes homogènes par rapport à cette probabilité de réponse estimée en utilisant l'algorithme de classification mis en œuvre dans la procédure FASTCLUS de SAS. On a pu obtenir ce nombre élevé de classes grâce à la grande taille de l'échantillon de l'EPA. On l'a choisi de manière à réduire le biais de non-réponse non seulement au niveau de la population, mais également pour les plus petits domaines. On a simplement calculé l'ajustement de poids pour la non-réponse d'un ménage répondant k au sein d'une classe donnée c en utilisant l'inverse du taux de réponse non pondéré au sein de la classe c. Un seuil pour l'ajustement de poids pour la non-réponse a été fixé à 2,5 pour contrôler la variance due à la non-réponse de l'estimateur aux poids ajustés pour la non-réponse. Il n'a fallu appliquer ce seuil que pour un nombre très petit de classes, soit celles qui présentaient les plus faibles probabilités de réponse estimées. Sans ce seuil, on aurait observé à l'occasion des ajustements de poids pour la non-réponse se situant autour de 4.

On a envisagé un autre modèle de non-réponse dans lequel on a modélisé la probabilité de réponse d'un ménage k comme le produit de la probabilité de joindre le ménage k par la probabilité de réponse de ce ménage, étant donné qu'on l'a joint. Ces deux dernières probabilités ont été modélisées séparément. Bien que ce modèle semble présenter une meilleure approximation de la réalité et qu'il ait donné des résultats légèrement supérieurs (en ce sens qu'il expliquait mieux la non-réponse), on n'a pas jugé l'amélioration suffisante pour ajouter cette complexité à la méthode d'ajustement de la non-réponse. Ce modèle pourrait cependant faire l'objet d'une étude plus approfondie.

4. Conclusion

Une contribution importante de cet article est qu'il faille considérer les renseignements sur le PCD comme aléatoires lorsqu'on les utilise dans un modèle de non-réponse. Nous avons ensuite montré que l'utilisation de ces renseignements pour traiter la non-réponse totale au moyen d'un ajustement de poids n'introduisait ni biais ni composante additionnelle de variance dans les estimations de totaux de population lorsque le modèle de non-réponse est bien spécifié. En outre, nous avons soutenu que si les renseignements sur le PCD étaient associés aux variables d'intérêt et à la non-réponse, leur utilisation avait alors tendance à réduire le biais de non-réponse lorsque le mécanisme de non-réponse dépend directement des variables d'intérêt. Enfin, au moyen de l'exemple de l'EPA, nous avons montré que ces renseignements pouvaient être utiles pour composer avec la non-réponse totale à une grande enquête.

L'estimateur de réponse complète que nous avons considéré est l'estimateur de Horvitz-Thompson. Nos conclusions seraient restées les mêmes si nous avions utilisé plutôt un estimateur par la régression généralisée. Nous avons choisi l'estimateur de Horvitz-Thompson pour sa simplicité et parce qu'il était suffisant pour démontrer l'essentiel de notre exposé.

Remerciements

Je tiens à remercier les membres du Comité consultatif sur les méthodes statistiques de Statistique Canada pour avoir soulevé des questions concernant l'application de la méthode proposée à l'Enquête sur la population active du Canada et, en particulier, J.N.K. Rao et Chris Skinner pour leurs précieuses observations à la suite de la présentation au Comité. Je tiens aussi à remercier sincèrement le rédacteur associé pour ses observations et ses suggestions. Elles se sont avérées très utiles et ont permis d'améliorer la clarté de l'article. Enfin, je suis très reconnaissant à Asma Alavi et Cynthia Bocci de Statistique Canada pour avoir mis au point les programmes informatiques ayant servi à analyser les données de l'Enquête sur la population active du Canada.

Bibliographie

Alavi, A., et Beaumont, J.-F. (2004). Nonresponse adjustment plans for the Labour Force Survey. Rapport technique présenté au Comité consultatif sur les méthodes statistiques, Statistique Canada, 2-3 mai 2004.

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society*, Series B, 67, 445-458.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Couper, M., et Lyberg, L. (2005). The use of paradata in survey research. *Bulletin of the International Statistical Institute* (à paraître).

Ekholm, A., et Laaksonen, S. (1991). Weighting via response modeling in the Finnish Household Budget Survey. *Journal of Official Statistics*, 7, 325-337.

Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. et Lindeyer, J. (1998). *Méthodologie de l'Enquête sur la population active du Canada*. Statistique Canada, Catalogue numéro 71-526.

Holt, D., et Elliott, D. (1991). Methods of weighting for unit non-response. *The Statistician*, 40, 333-342.

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

Little, R.J., et Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine*, 22, 1589-1599.

Little, R.J., et Vartivarian, S. (2005). La pondération pour la non-réponse augmente-t-elle la variance des moyennes de sondage? *Techniques d'enquête*, 31, 175-183.

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

Techniques d'enquête, décembre 2005
Vol. 31, N° 2, pp. 255-260
Statistique Canada, n° 12-001-XPB au catalogue

Structure de corrélation des unités d'échantillonnage

Alfredo Bustos [1]

Résumé

Nous explicitons dans cet article certaines propriétés distributionnelles des unités d'échantillonnage qui ne sont habituellement pas décrites dans la documentation, notamment leur structure de corrélation et le fait que celle-ci ne dépend pas d'indices de population attribués arbitrairement. Ces propriétés importent pour plusieurs méthodes d'estimation, dont l'efficacité serait améliorée si on les mentionnait explicitement.

Mots clés : Recensement; enquête; échantillonnage; unités d'échantillonnage; fonction de probabilité; moyenne; covariance.

1. Introduction

Ces dernières années, la réalisation des recensements de la population et des ménages tels que nous les connaissons est devenue plus ardue pour plusieurs raisons. Par conséquent, d'autres moyens de recueillir plus fréquemment l'information requise pour la production de statistiques aux niveaux local, provincial et national ont été proposés. De grandes enquêtes nationales continues, notamment celles appelées recensement continu, réalisées auprès d'échantillons de grande taille selon des plans d'enquête complexes, sont envisagées.

Cependant, afin de produire des résultats au niveau local comparables à ceux d'un recensement, il faut mettre au point diverses méthodes d'estimation et de validation, ainsi que, dans certains cas, d'imputation et améliorer leur efficacité. Un moyen d'accroître l'efficacité consiste à tenir compte de toute l'information pertinente disponible. Naturellement, cela englobe les propriétés stochastiques des unités d'échantillonnage.

Dans la suite de l'exposé, en partant de principes fondamentaux, nous dérivons une forme générale explicite de la fonction de probabilité d'un échantillon ordonné. Nous montrons aussi comment on peut calculer cette fonction, ainsi que les probabilités d'inclusion. Enfin, nous donnons une forme générale de la matrice des corrélations des unités d'échantillonnage qui ne dépend que des probabilités d'inclusion, de sorte qu'il soit possible d'améliorer les méthodes d'estimation linéaires et du maximum de vraisemblance.

2. Le modèle de base

Le modèle de base dont nous partons représente le tirage séquentiel aléatoire de n unités à partir d'une population U formée de N de ces unités et peut être énoncé comme suit. Soit N et n deux constantes positives telles que $n \le N$, et soit V une matrice de dimensions $N \times n$, dont les composantes sont distribuées chacune comme des variables aléatoires de Bernoulli avec, éventuellement, des paramètres différents. Alors,

$$V_{N\times n} = \begin{bmatrix} \vartheta_{11} & \vartheta_{12} & \vartheta_{13} & \cdots & \vartheta_{1n} \\ \vartheta_{21} & \vartheta_{22} & \vartheta_{23} & \cdots & \vartheta_{2n} \\ \vartheta_{31} & \vartheta_{32} & \vartheta_{33} & \cdots & \vartheta_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vartheta_{N1} & \vartheta_{N2} & \vartheta_{N3} & \cdots & \vartheta_{Nn} \end{bmatrix}. \qquad (1.1)$$

Fait aussi partie du modèle la contrainte voulant que la somme des éléments de chaque colonne de V soit égale à l'unité. Autrement dit, nous exigeons que la condition

$$\sum_{I=1}^{N} \vartheta_{Ik} = 1, \text{ for } k = 1, ..., n \qquad (1.2)$$

soit satisfaite.

Cette condition est nécessaire, parce que si le j^{e} tirage donne lieu à la sélection de l'unité de population I, alors l'élément (I, j) prend la valeur de un, tandis que tous les autres éléments de la colonne j sont nuls. Notons que cela équivaut à imposer une contrainte non stochastique au comportement de toutes les composantes de la i^{e} colonne de V, indépendamment du plan d'échantillonnage. Par conséquent, les éléments appartenant à une même colonne ne se comportent pas de façon indépendante.

Lorsque l'échantillonnage a lieu avec remise (WR pour *with replacement*), la somme des éléments de la I^{e} ligne de la matrice susmentionnée suit une loi binomiale (n, p_I), puisque la distribution de chaque colonne est indépendante de celle des autres. Par ailleurs, si l'échantillonnage se fait sans remise (WOR pour *without replacement*), le total de la

1. Victor Alfredo Bustos y de la Tijera, Instituto Nacional de Estadística, Geografia e Informática, H. de Nacozari 2301, 20270, Aguascalientes, Ags., México. Courriel : alfredo.bustos@inegi.gob.mx.

ligne I ne peut prendre que deux valeurs : un, si la I^{e} unité est tirée à un certain degré, ou zéro, autrement, ce qui nous ramène au cas de Bernoulli.

Nous pouvons former des sous-ensembles disjoints de lignes conformément à divers critères. Par exemple, si nous regroupons les lignes en fonction de leur voisinage spatial, nous pourrions parler de grappes ou d'unités primaires d'échantillonnage. Si nous fondons le groupement sur un ou plusieurs indicateurs statistiques, nous utilisons habituellement le terme de strate.

Définissons maintenant les probabilités d'inclusion comme étant

$$\begin{aligned}\pi_I^{(k)} &= P(\text{unité de population } I \text{ dans l'échantillon} \\ &\qquad \text{de taille } k) \\ &= 0 \text{ si } k = 0.\end{aligned} \tag{2}$$

Notons que $\pi_I^{(n)} = \pi_I$, habituellement nommée probabilité d'inclusion de l'unité I.

Représentons maintenant par $\underline{\vartheta}_{\circ j}$ la j^{e} colonne et par $\underline{\vartheta}_{I\circ}$ la I^{e} ligne de la matrice V. Par conséquent, en nous basant sur l'expression suivante,

$$\begin{aligned}f(\underline{\vartheta}_{\circ 1}, \underline{\vartheta}_{\circ 2}, \underline{\vartheta}_{\circ 3}, ..., \underline{\vartheta}_{\circ n}) &= f(\underline{\vartheta}_{\circ 1}) f(\underline{\vartheta}_{\circ 2} \mid \underline{\vartheta}_{\circ 1}) \\ &\quad f(\underline{\vartheta}_{\circ 3} \mid \underline{\vartheta}_{\circ 1}, \underline{\vartheta}_{\circ 2}) \ldots f(\underline{\vartheta}_{\circ n} \mid \underline{\vartheta}_{\circ 1}, ..., \underline{\vartheta}_{\circ n-1})\end{aligned} \tag{3}$$

nous pouvons écrire la fonction de probabilité conjointe des éléments de V sous la forme :

$$\begin{aligned}f(\underline{\vartheta}_{\circ 1}, \underline{\vartheta}_{\circ 2}, \underline{\vartheta}_{\circ 3}, ..., \underline{\vartheta}_{\circ n}) &= \prod_{k=1}^{n}\left[\prod_{I=1}^{N}(\pi_I^{(k)} - \pi_I^{(k-1)})^{\vartheta_{Ik}}\right] \\ &= \prod_{k=1}^{n}\left[\prod_{I=1}^{N}(p_I^{(k)})^{\vartheta_{Ik}}\right]\end{aligned} \tag{4}$$

sachant que

$$\sum_{I=1}^{N}\vartheta_{Ik} = 1, k = 1, \ldots, n \text{ et}$$

$$\sum_{k=1}^{N}\vartheta_{Ik} \le \begin{cases}1, \text{ WOR} \\ n, \text{ WR}\end{cases} I = 1, \ldots, N;$$

et ici $p_I^{(k)}$, définie comme étant $p_I^{(k)} = (\pi_I^{(k)} - \pi_I^{(k-1)})$, représente la probabilité que l'unité de population I soit incluse dans l'échantillon lors du k^{e} tirage. La fonction susmentionnée est utile pour le calcul de la probabilité de tout échantillon ordonné de taille n. Manifestement, si l'on peut ignorer l'ordre d'inclusion, on obtiendra la probabilité d'un échantillon donné en ajoutant les n! valeurs obtenues au moyen de (4).

3. Les incidences de l'échantillonnage sur les propriétés stochastiques des unités de population

Conséquemment,

$$E(\vartheta_{Ik}) = p_I^{(k)} = (\pi_I^{(k)} - \pi_I^{(k-1)}) \tag{5}$$

et, donc, nous pouvons écrire

$$E[V] = \begin{bmatrix} p_1^{(1)} & p_1^{(2)} & p_1^{(3)} & \cdots & p_1^{(n)} \\ p_2^{(1)} & p_2^{(2)} & p_2^{(3)} & \cdots & p_2^{(n)} \\ p_3^{(1)} & p_3^{(2)} & p_3^{(3)} & \cdots & p_3^{(n)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_N^{(1)} & p_N^{(2)} & p_N^{(3)} & \cdots & p_N^{(n)} \end{bmatrix}. \tag{6}$$

À partir de là, nous pouvons calculer récursivement les probabilités d'inclusion étape par étape, dans les situations d'échantillonnage sans remise, comme le montre l'équation (7) qui suit.

$$p_I^{(k)} = \begin{cases} p_I & \text{si } k = 1 \\ p_I^{(k-1)} \sum_{J \ne I}^{N} \dfrac{p_J^{(k-1)}}{1 - p_J^{(k-1)}} & \text{si } k > 1. \end{cases} \tag{7}$$

Il convient de souligner que (7) nous permet de calculer les probabilités souhaitées à deux moments distincts : en premier lieu, quand aucun tirage n'a effectivement eu lieu, ce qui explique pourquoi nous calculons la moyenne sur l'ensemble de la population et, en deuxième lieu, quand on connaît le résultat du tirage précédent, moment auquel la probabilité que la J^{e} unité de population, disons, entre dans l'échantillon est égale à 1 et toutes les autres probabilités pour ce tirage sont nulles. Par conséquent, du moins en théorie, nous pouvons calculer l'inverse des facteurs dits d'expansion ou poids pour l'échantillonnage à un degré, ou étape par étape pour l'échantillonnage à plusieurs degrés. De toute évidence,

$$\pi_I^{(n)} = \sum_{k=1}^{n} p_I^{(k)}. \tag{8}$$

Si nous définissons les probabilités d'inclusion conjointes comme étant

$$\pi_{IJ}^{(k)} = P\left(\begin{array}{c}\text{unités de population } I \text{ et} \\ J \text{ dans l'échantillon de taille } k\end{array}\right), \tag{9}$$

alors nous savons qu'elles peuvent également être calculées comme suit :

$$\pi_{IJ}^{(n)} = \sum_{j=1}^{n-1}\left(p_I^{(j)} \sum_{k>j}^{n} p_J^{(k)} + p_J^{(j)} \sum_{k>j}^{n} p_I^{(k)}\right). \tag{10}$$

Par exemple, dans le cas de l'échantillonnage aléatoire simple avec remise (EAS/WR), les expressions (7), (8) et (10) donnent lieu à (7.1), (8.1) et (10.1),

$$p_I^{(k)} = \frac{1}{N} \text{ quand } k \ge 1 \tag{7.1}$$

$$\pi_I^{(n)} = \frac{n}{N} \tag{8.1}$$

$$\pi_{IJ}^{(n)} = \sum_{j=1}^{n-1}\left(p_I^{(j)} \sum_{k>j}^{n} p_J^{(k)} + p_J^{(j)} \sum_{k>j}^{n} p_I^{(k)} \right)$$
$$= \sum_{j=1}^{n-1}\left(\frac{n-j}{N^2} + \frac{n-j}{N^2} \right) = \frac{n(n-1)}{N^2}. \qquad (10.1)$$

Dans le cas de l'EAS/WOR, nous obtenons, à la place, les expressions (7.2), (8.2) et (10.2).

$$p_I^{(k)} = \frac{1}{N} \text{ quand } k \geq 1 \qquad (7.2)$$

$$\pi_I^{(n)} = \frac{n}{N} \qquad (8.2)$$

$$\pi_{IJ}^{(n)} = \sum_{j=1}^{n-1}\left(p_I^{(j)} \sum_{k>j}^{n} p_J^{(k)} + p_J^{(j)} \sum_{k>j}^{n} p_I^{(k)} \right) \text{ où } J \neq I$$
$$= \sum_{j=1}^{n-1}\left(\frac{n-j}{N(N-1)} + \frac{n-j}{N(N-1)} \right) = \frac{n(n-1)}{N(N-1)}. \qquad (10.2)$$

Considérons maintenant les vecteurs de ligne $\underline{\vartheta}_{I\circ}$. Alors, pour la matrice des covariances entre diverses lignes, nous obtenons

$$\mathrm{Cov}(\underline{\vartheta}_{I\circ}, \underline{\vartheta}_{J\circ}) = \begin{bmatrix} -p_I^{(1)}p_J^{(1)} & 0 & \cdots & 0 \\ 0 & -p_I^{(2)}p_J^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -p_I^{(n)}p_J^{(n)} \end{bmatrix}_{n\times n} \qquad (11)$$

dans tous les cas où I est différent de J.

En cas d'échantillonnage avec remise où, par conséquent, $p_I^{(j)} = p_I \ \forall\ j = 1, ..., n,$ la matrice des covariances pour le I^{e} vecteur de ligne est donnée par

$$\mathrm{Cov}(\underline{\vartheta}_{I\circ}, \underline{\vartheta}_{I\circ}) = \begin{bmatrix} p_I q_I & 0 & 0 & \cdots & 0 \\ 0 & p_I q_I & 0 & \cdots & 0 \\ 0 & 0 & p_I q_I & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & p_I q_I \end{bmatrix}_{n\times n}. \qquad (12.1)$$

Dans le cas de l'échantillonnage sans remise, la matrice des covariances susmentionnée devient

$$\mathrm{Cov}(\underline{\vartheta}_{I\circ}, \underline{\vartheta}_{I\circ}) = \begin{bmatrix} p_I^{(1)}(1-p_I^{(1)}) & -p_I^{(1)}p_I^{(2)} & \cdots & -p_I^{(1)}p_I^{(n)} \\ -p_I^{(1)}p_I^{(2)} & p_I^{(2)}(1-p_I^{(2)}) & \cdots & -p_I^{(2)}p_I^{(n)} \\ \vdots & \vdots & \ddots & \vdots \\ -p_I^{(1)}p_I^{(n)} & -p_I^{(2)}p_I^{(n)} & \cdots & p_I^{(n)}(1-p_I^{(n)}) \end{bmatrix}_{n\times n} \cdot \qquad (12.2)$$

Soit $\underline{\vartheta}$ le vecteur de dimension N qui résulte de l'addition des colonnes de V. De toute évidence, les composantes de ce vecteur peuvent être exprimées sous forme du produit de $\underline{\vartheta}_{I\circ}$ par un vecteur dont les composantes sont toutes égales à un. Autrement dit,

$$\underline{\vartheta} = \begin{pmatrix} \vartheta_1 \\ \vartheta_2 \\ \vartheta_3 \\ \vdots \\ \vartheta_N \end{pmatrix} = \begin{pmatrix} \underline{\vartheta}_{1\circ}^T \underline{1} \\ \underline{\vartheta}_{2\circ}^T \underline{1} \\ \underline{\vartheta}_{3\circ}^T \underline{1} \\ \vdots \\ \underline{\vartheta}_{N\circ}^T \underline{1} \end{pmatrix}. \qquad (13)$$

Certaines propriétés distributionnelles de ces sommes peuvent alors être obtenues directement d'après celles des lignes ou des colonnes de la matrice V.

Par exemple, leurs valeurs attendues sont données par

$$E(\vartheta_I) = E(\underline{\vartheta}_{I\circ}^T \underline{1}) = E\left(\sum_{k=1}^{n} \vartheta_{Ik} \right)$$
$$= \sum_{k=1}^{n} p_I^{(k)} = \pi_I^{(1)} + \sum_{k=2}^{n} (\pi_I^{(k)} - \pi_I^{(k-1)}) = \pi_I^{(n)}. \qquad (14)$$

De (1.2), nous tirons la restriction non stochastique :

$$\underline{1}'\,\underline{\vartheta} = \vartheta_1 + \vartheta_2 + \vartheta_3 + ... + \vartheta_N = n. \qquad (15)$$

De (14) et (15) découlent directement les propositions bien connues (16) et (17),

$$E[\underline{\vartheta}'] = (\pi_1^{(n)}, \pi_2^{(n)}, \pi_3^{(n)}, \ldots, \pi_N^{(n)}) \qquad (16)$$

$$\pi_1^{(n)} + \pi_2^{(n)} + \pi_3^{(n)} + \ldots + \pi_N^{(n)} = n. \qquad (17)$$

Pour les moments de deuxième ordre, nous obtenons

$$\mathrm{Cov}(\vartheta_I, \vartheta_J) = \mathrm{Cov}(\underline{1}'\,\underline{\vartheta}_{I\circ}, \underline{1}'\,\underline{\vartheta}_{J\circ})$$
$$= \underline{1}'\mathrm{Cov}(\underline{\vartheta}_{I\circ}, \underline{\vartheta}_{J\circ})\underline{1} = -\sum_{k=1}^{n} p_I^{(k)} p_J^{(k)}$$
$$= \begin{cases} -np_I p_J & \text{WR} \\ (\pi_{IJ}^{(n)} - \pi_I^{(n)}\pi_J^{(n)}) & \text{WOR}, \end{cases} \qquad (18)$$

qui indique clairement que la covariance n'est jamais positive. À leur tour, les variances sont données par

$$\mathrm{Var}(\vartheta_I) = \mathrm{Var}(\underline{1}'\,\underline{\vartheta}_{I\circ}) = \underline{1}'\,\mathrm{Cov}(\underline{\vartheta}_{I\circ})\,\underline{1}$$
$$= \begin{cases} np_I q_I & \text{WR} \\ \pi_I^{(n)}(1-\pi_I^{(n)}) & \text{WOR}. \end{cases} \qquad (19)$$

Une autre conséquence importante de (15) concerne les moments de deuxième ordre du vecteur stochastique $\underline{\vartheta}$.

$$0 = \mathrm{Var}(n) = \mathrm{Var}(\underline{1}'\,\underline{\vartheta}) = \underline{1}'\mathrm{Cov}(\underline{\vartheta})\underline{1} = \underline{1}'C\underline{1}. \qquad (20)$$

De toute évidence, les éléments diagonaux de la matrice C, la matrice des covariances de $\underline{\vartheta}$, ne sont pas tous nuls. Par conséquent, le tirage aléatoire d'un échantillon de taille fixe introduit dans les unités de population une dépendance qui donne lieu à des covariances non nulles sous-entendant

que la matrice C est singulière. Sinon, il est impossible que l'équation (20) soit satisfaite.

En fait, il est possible de prouver que la somme des éléments de toute ligne (ou colonne) de C doit être nulle, ce qui est un énoncé plus ferme. Sachant que la covariance entre une variable aléatoire et une constante est nulle, nous obtenons

$$\begin{aligned} 0 &= \mathrm{Cov}(\vartheta_I, n) = \mathrm{Cov}(\vartheta_I, \vartheta_1 + \vartheta_2 + \vartheta_3 + \cdots + \vartheta_N) \\ &= C_{I1} + C_{I2} + \cdots + C_{IN} \\ &= \mathrm{Var}(\vartheta_I) + \sum_{J \neq I} \mathrm{Cov}(\vartheta_I, \vartheta_J). \end{aligned} \tag{21}$$

Nous avons donc prouvé que, dans le cas de l'échantillonnage sans remise, (22.1) est vérifiée.

$$0 = \pi_I^{(n)}(1 - \pi_I^{(n)}) + \sum_{J \neq I} (\pi_{IJ}^{(n)} - \pi_I^{(n)} \pi_J^{(n)}). \tag{22.1}$$

Le même énoncé peut être prouvé algébriquement en notant que

$$\begin{aligned} \sum_{J \neq I} \pi_{IJ}^{(n)} &= \pi_I^{(n)} \sum_{J \neq I} \pi_{J|I}^{(n)} \\ &= (n-1)\pi_I^{(n)}, \end{aligned}$$

ce qui est évident si l'on se rend compte que la probabilité conditionnelle concernée représente la probabilité que l'unité de population J entre dans un échantillon de taille $n-1$ pour lequel s'applique aussi l'expression (19). En outre, en utilisant de nouveau (19), notons que

$$\sum_{J \neq I} \pi_J^{(n)} = (n - \pi_I^{(n)}),$$

et, par conséquent,

$$\begin{aligned} 0 &= \pi_I^{(n)}(1 - \pi_I^{(n)}) + \sum_{J \neq I} (\pi_{IJ}^{(n)} - \pi_I^{(n)} \pi_J^{(n)}) \\ &= \pi_I^{(n)} - (\pi_I^{(n)})^2 + (n-1)\pi_I^{(n)} - \pi_I^{(n)}(n - \pi_I^{(n)}). \end{aligned}$$

Pour l'échantillonnage avec remise, (21) implique que :

$$\begin{aligned} 0 &= np_I q_I + \sum_{J \neq I} (n(n-1) p_I p_J - n^2 p_I p_J) \\ &= np_I q_I - np_I \sum_{J \neq I} p_J \end{aligned} \tag{22.2}$$

condition qui, on le voit directement, s'applique.

En tout cas, l'incidence la plus importante des résultats susmentionnés est que, indépendamment du plan d'échantillonnage, la matrice de corrélation des variables aléatoires de population $\vartheta_1, \vartheta_2, \vartheta_3, ..., \vartheta_N$ est singulière. En ce qui concerne les situations pratiques décrites dans l'introduction, la conséquence la plus importante tient principalement au fait que l'inverse de la matrice des covariances est utilisée dans de nombreuses méthodes d'ajustement et d'estimation de modèles.

4. Les deux premiers moments des unités d'échantillonnage

Après avoir établi les moments de premier et de deuxième ordre du vecteur $\underline{\vartheta}$, il nous est possible de déterminer les moments correspondants de sous-vecteurs de diverses tailles et dont les composantes sont sélectionnées aléatoirement, c'est-à-dire l'échantillon. À cette fin, définissons les variables aléatoires $\vartheta_{I_1}, \vartheta_{I_2}, \vartheta_{I_3}, \ldots, \vartheta_{I_r}$, où r représente le nombre d'unités de population différentes dans l'échantillon et dont les indices $I_k, 1 \leq k \leq r \leq n,$ peuvent prendre la valeur I avec la probabilité $\pi_I^{(n)}$. Autrement dit, sous les conditions susmentionnées, nous sommes en présence d'un jeu de variables aléatoires dont les indices sont eux-mêmes aléatoires.

4.1 Moyenne et variance pour l'échantillonnage avec remise

Dans ce cas, la fonction de probabilité de ϑ_{I_j} est donnée par

$$\begin{aligned} P(\vartheta_{I_j} = x) &= \sum_{I=1}^{N} p_I P(\vartheta_I = x) \\ &= \sum_{I=1}^{N} p_I \binom{n}{x} p_I^x (1 - p_I)^{n-x}. \end{aligned} \tag{23}$$

Les deux premiers moments peuvent aussi être obtenus par la voie d'un argument conditionnel. La moyenne de sa distribution est donnée par

$$E(\vartheta_{I_j}) = \sum_{I=1}^{N} p_I E(\vartheta_I) = \sum_{I=1}^{N} np_I \, p_I = n \sum_{I=1}^{N} p_I^2. \tag{24}$$

À son tour, sa variance est calculée à l'aide de la formule bien connue

$$V(\vartheta_{I_j}) = V_{I_j}[E(\vartheta_{I_j} \mid I_j)] + E_{I_j}[V(\vartheta_{I_j} \mid I_j)]. \tag{25}$$

Dans ce cas, nous avons

$$\begin{aligned} &E(\vartheta_{I_j} \mid I_j = I) = np_I \\ &\text{et } V(\vartheta_{I_j} \mid I_j = I) = np_I(1 - p_I). \end{aligned} \tag{26}$$

Donc,

$$\begin{aligned} V_{I_j}[E(\vartheta_{I_j} \mid I_j)] &= V_{I_j}(np_{I_j}) \\ &= n^2[E_{I_j}(p_{I_j}^2) - E_{I_j}^2(p_{I_j})], \\ E_{I_j}[V(\vartheta_{I_j} \mid I_j)] &= nE_{I_j}[p_{I_j}(1 - p_{I_j})] \\ &= n[E_{I_j}(p_{I_j}) - E_{I_j}(p_{I_j}^2)] \end{aligned} \tag{27}$$

et, par conséquent

$$V(\vartheta_{I_j})$$
$$= n[E_{I_j}(p_{I_j}) - E_{I_j}(p_{I_j}^2)] + n^2[E_{I_j}(p_{I_j}^2) - E_{I_j}^2(p_{I_j})]$$
$$= \sum_{I=1}^{N} np_I^2\left(1 + (n-1)p_I - \sum_{J=1}^{N} np_J^2\right). \quad (28)$$

Pour le cas de l'EAS, l'équation (24) qui précède donne

$$E(\vartheta_{I_j}) = \frac{1}{n}\sum_{I=1}^{N}(\pi_I^{(n)})^2 = \frac{1}{n}\sum_{I=1}^{N}\left(\frac{n}{N}\right)^2 = \left(\frac{n}{N}\right)^2\frac{N}{n} = \frac{n}{N}.$$

Tandis que (28) donne

$$V(\vartheta_{I_j}) = \sum_{I=1}^{N} n\frac{1}{N^2}\left(1 + (n-1)\frac{1}{N} - \sum_{J=1}^{N} n\frac{1}{N^2}\right) = n\frac{1}{N}\left(1 - \frac{1}{N}\right).$$

4.2 Moyenne et variance pour l'échantillonnage sans remise

Dans ce cas, la fonction de probabilité de ϑ_{I_i} est donnée par

$$P(\vartheta_{I_j} = x) = \frac{1}{n}\sum_{I=1}^{N}\pi_I^{(n)}\sum_{k=1}^{n}(p_I^{(k)})^x(1 - p_I^{(k)})^{1-x} \quad (29)$$

et, par conséquent,

$$E(\vartheta_{I_j}) = \frac{1}{n}\sum_{I=1}^{N}\pi_I^{(n)}E(\vartheta_I)$$
$$= \frac{1}{n}\sum_{I=1}^{N}\pi_I^{(n)}\sum_{j=1}^{n}(p_I^{(j)}) = \frac{1}{n}\sum_{I=1}^{N}(\pi_I^{(n)})^2. \quad (30)$$

En utilisant de nouveau (25), nous commençons par noter que

$$E(\vartheta_{I_j} \mid I_j) = \pi_{I_j}^{(n)} \text{ et } V(\vartheta_{I_j} \mid I_j) = \pi_{I_j}^{(n)}(1 - \pi_{I_j}^{(n)})$$

dont nous tirons

$$V[E(\vartheta_{I_j} \mid I_j)] = V(\pi_{I_j}^{(n)}) = E[(\pi_{I_j}^{(n)})^2] - [E(\pi_{I_j}^{(n)})]^2$$

et

$$E[V(\vartheta_{I_j} \mid I_j)] = E[\pi_{I_j}^{(n)}(1 - \pi_{I_j}^{(n)})] = E[(\pi_{I_j}^{(n)})] - [E(\pi_{I_j}^{(n)})^2].$$

Donc, la variance est donnée par

$$V(\vartheta_{I_j}) = E(\pi_{I_j}^{(n)}) - E^2(\pi_{I_j}^{(n)}) = E(\pi_{I_j}^{(n)})[1 - E(\pi_{I_j}^{(n)})]$$
$$= \left(\frac{1}{n}\sum_{I=1}^{N}(\pi_I^{(n)})^2\right)\left[1 - \left(\frac{1}{n}\sum_{I=1}^{N}(\pi_I^{(n)})^2\right)\right]. \quad (31)$$

Une fois de plus, afin d'illustrer ces résultats au moyen d'un exemple, considérons l'EAS. L'expression (30) devient

$$E(\vartheta_{I_j}) = \frac{1}{n}\sum_{I=1}^{N}(\pi_I^{(n)})^2$$
$$= \frac{1}{n}\sum_{I=1}^{N}\left(\frac{n}{N}\right)^2 = \left(\frac{n}{N}\right)^2\frac{N}{n} = \frac{n}{N}. \quad (32)$$

Tandis que (31) donne

$$V(\vartheta_{I_j}) = \left(\frac{1}{n}\sum_{I=1}^{N}\left(\frac{n}{N}\right)^2\right)\left[1 - \left(\frac{1}{n}\sum_{I=1}^{N}\left(\frac{n}{N}\right)^2\right)\right]$$
$$= \frac{n}{N}\left(1 - \frac{n}{N}\right). \quad (33)$$

4.3 La covariance entre les unités d'échantillonnage

Afin d'établir la covariance entre les diverses unités d'échantillonnage, nous recourons à une simple extension de (25),

$$\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k}) = \mathrm{Cov}_{I_j, I_k}[E(\vartheta_{I_j} \mid I_j), E(\vartheta_{I_k} \mid I_k)]$$
$$+ E_{I_j, I_j}[\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k} \mid I_j, I_k)]. \quad (34)$$

Dans ce cas, nous savons que

$$E(\vartheta_{I_j} \mid I_j = I) = \pi_I^{(n)} \quad (35)$$

et

$$E(\vartheta_{I_j}, \vartheta_{I_k} \mid I_j = I, I_k = J) = \pi_{IJ}^{(n)} \quad (36)$$

tandis qu'il est facile de voir que la covariance entre crochets dans le deuxième membre de (34) est égale à

$$\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k} | I_j = I, I_k = J) = \pi_{IJ}^{(n)} - \pi_I^{(n)}\pi_j^{(n)}. \quad (37)$$

À partir de (35) et (36), nous obtenons

$$\mathrm{Cov}_{I_jI_k}[E(\vartheta_{I_j} \mid I_j), E(\vartheta_{I_k} \mid I_k)]$$
$$= E_{I_jI_k}(\pi_{I_j}^{(n)}\pi_{I_k}^{(n)}) - E_{I_j}(\pi_{I_j}^{(n)})E_{I_k}(\pi_{I_k}^{(n)}) \quad (38)$$

tandis que, de (37), nous obtenons

$$E_{I_jI_k}[\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k} \mid I_j, I_k)]$$
$$= E_{I_jI_k}(\pi_{I_jI_k}^{(n)}) - E_{I_jI_k}(\pi_{I_j}^{(n)}\pi_{I_k}^{(n)}). \quad (39)$$

Enfin, en additionnant ces deux dernières expressions, nous obtenons la covariance souhaitée

$$\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k})$$
$$= E_{I_jI_k}(\pi_{I_jI_k}^{(n)}) - [E_{I_j}(\pi_{I_j}^{(n)})][E_{I_k}(\pi_{I_k}^{(n)})]$$
$$= \frac{1}{n(n-1)}\sum_{I=1}^{N}\sum_{\substack{J=1 \\ J\neq I}}^{N}(\pi_{IJ}^{(n)})^2 - \left(\frac{1}{n}\sum_{I=1}^{N}(\pi_I^{(n)})^2\right)^2. \quad (40)$$

Dans le cas de l'EAS/WR, (40) donne

$$\begin{aligned}\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k}) &= \frac{1}{n(n-1)} \sum_{I=1}^{N} \sum_{J=1}^{N} \left(\frac{n(n-1)}{N^2} \right)^2 \\ &\quad - \left(\frac{1}{n} \sum_{I=1}^{N} \left(\frac{n}{N} \right)^2 \right)^2 \\ &= \frac{n(n-1)}{N^2} - \frac{n^2}{N^2} \\ &= -\frac{n}{N^2}, \end{aligned} \tag{41}$$

tandis que dans le cas sans remise, on peut voir que la covariance est égale à

$$\begin{aligned}\mathrm{Cov}(\vartheta_{I_j}, \vartheta_{I_k}) &= \frac{1}{n(n-1)} \sum_{I=1}^{N} \sum_{\substack{J=1 \\ J \neq I}}^{N} \left(\frac{n(n-1)}{N(N-1)} \right)^2 \\ &\quad - \left(\frac{1}{n} \sum_{I=1}^{N} \left(\frac{n}{N} \right)^2 \right)^2 \\ &= \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} \\ &= -\left(\frac{n(N-n)}{N^2(N-1)} \right). \end{aligned} \tag{42}$$

Il convient de souligner que, dans le cas de l'EAS, que le tirage se fasse avec ou sans remise, les coefficients de corrélation sont donnés par

$$\mathrm{Corr}(\vartheta_{I_j}, \vartheta_{I_k}) = \frac{-1}{(N-1)}, \tag{43}$$

indépendamment de la taille de l'échantillon.

De surcroît, nous savons que, à mesure que la valeur de n s'approche de celle de N dans l'échantillonnage sans remise, $\pi_I^{(n)}$ et $\pi_{IJ}^{(n)}$ s'approchent l'une et l'autre de l'unité. En particulier, quand $n = N$, les valeurs des expressions (31) et (40) deviennent nulles.

5. La matrice des corrélations des unités d'échantillonnage

Dès que l'on se rend compte qu'aucune des expressions (28), (31) et (40) ne dépend d'aucun des indices arbitraires utilisés pour différencier les unités de population, il devient clair que la matrice $r \times r$ des corrélations pour le vecteur aléatoire $\underline{\theta} = (\vartheta_{I_1}, \vartheta_{I_2}, \vartheta_{I_3}, \ldots, \vartheta_{I_r})$, où $r \leq n$, peut s'écrire :

$$\mathrm{Corr}(\underline{\theta}) = R_r(\rho) = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix}. \tag{44}$$

Il convient de souligner que les éléments de $R_r(\rho)$ dans (44) dépendent uniquement des probabilités d'inclusion qui, pour toute taille d'échantillon, peuvent être calculées entièrement d'après la récursion (7) et les expressions (8) et (10). Autrement dit, elles ne dépendent d'aucun paramètre de population inconnu qu'il faut estimer ni des valeurs des variables qui doivent être mesurées sur les unités d'échantillonnage.

6. Remarques finales

En théorie, l'efficacité de toute méthode d'estimation s'accroîtrait dans une certaine mesure si l'on tenait compte explicitement de la corrélation entre les unités d'échantillonnage. Il en serait certainement ainsi pour l'estimation linéaire et, dans certains cas, pour l'estimation du maximum de vraisemblance.

Par ailleurs, il convient d'insister sur le fait que $R_n(\rho)$ peut devenir singulière à mesure que la taille de l'échantillon n s'approche de la taille de la population N; il en est ainsi pour l'EAS ($R_N(-1/(N-1))$, ainsi que pour l'échantillonnage sans remise en général. Par conséquent, numériquement, nombre de méthodes d'estimation qui s'appuient sur l'inverse ou le déterminant de R plutôt que sur la matrice des corrélation proprement dite pourraient également bénéficier du remplacement de l'hypothèse simplifiante d'indépendance entre les observations par une hypothèse plus réaliste d'observations corrélées quand la taille de l'échantillon est grande relativement à taille de la population. Les cas où cela est possible se produisent à certaines étapes dans l'échantillonnage à plusieurs degrés (par exemple nombre de ménages dans un îlot) et dans de grandes enquêtes à l'échelle du pays.

Techniques d'enquête, décembre 2005
Vol. 31, N° 2, pp. 261-266
Statistique Canada, N° 12-001-XPB au catalogue

Algorithmes et codes R pour la méthode de la pseudo-vraisemblance empirique dans les sondages

Changbao Wu [1]

Résumé

Nous présentons des algorithmes informatiques pour la méthode de la pseudo-vraisemblance empirique proposée récemment pour l'analyse des données d'enquête complexes. Plusieurs algorithmes essentiels pour le calcul des estimateurs du maximum de pseudo-vraisemblance empirique et la construction des intervalles de confiance des rapports de pseudo-vraisemblance empirique sont implantés au moyen des logiciels statistiques R et S-PLUS d'usage très répandu. Les codes principaux sont écrits sous la forme de fonctions R/S-PLUS et peuvent donc être utilisés directement dans les applications d'enquête et (ou) les études en simulation.

Mots clés : Intervalle de confiance; algorithme de bisection; vraisemblance empirique; procédure de Newton-Raphson; échantillonnage stratifié; échantillonnage avec probabilités inégales.

1. Introduction

L'un des grands défis que pose l'application de méthodes statistiques avancées et souvent complexes à des sondages réels est l'implantation informatique de la méthode. Souvent, des considérations pratiques obligent à rejeter des méthodes théoriquement valides et séduisantes, mais nécessitant une quantité incroyable de calculs.

La méthode de la vraisemblance empirique, proposée pour la première fois par Owen (1988), est l'un des principaux progrès réalisés en statistique au cours des 15 dernières années. Outre le fait qu'elle soit axée sur les données et qu'elle respecte les gammes de valeurs dans l'estimation et les tests, sa nature non paramétrique et discrète est particulièrement intéressante pour la résolution de problèmes en population finie. En effet, l'une de ses premières versions, appelée méthode des estimateurs « scale-load », a été utilisée en sondage par Hartley et Rao en 1968. L'étude plus récente de cette méthode dans le contexte des sondages a donné lieu à la publication d'une série de documents de recherche et suscité chez les statisticiens d'enquête un vif intérêt qui les a poussés à l'explorer plus en détail. Wu et Rao (2004) résument brièvement les faits récents concernant la méthode de la pseudo-vraisemblance empirique (PEL pour Pseudo Empirical Likelihood).

Des progrès ont également été réalisés en ce qui concerne l'élaboration d'algorithmes. Chen, Sitter et Wu (2002) ont proposé une procédure de Newton-Raphson modifiée pour calculer les estimateurs du maximum de pseudo-vraisemblance empirique sous échantillonnage non stratifié. Wu (2004a) a poursuivi la modification de la procédure afin de permettre le traitement des plans de sondage stratifiés.

Dans le présent article, nous présentons des algorithmes informatiques permettant de calculer les estimateurs du maximum de pseudo-vraisemblance empirique et de construire les intervalles de confiance des rapports de pseudo-vraisemblance empirique connexes pour des plans de sondage complexes sous un cadre unifié, en mettant surtout l'accent sur l'implantation de ces algorithmes au moyen des logiciels R et S-PLUS. Le progiciel R, un environnement de programmation convivial compatible avec le logiciel statistique commercial S-PLUS très répandu, intéresse de plus en plus les statisticiens. L'un des avantages de l'utilisation du progiciel R est qu'il est offert gratuitement pour la recherche et qu'il peut être téléchargé facilement à partir d'Internet. Nous espérons que le présent article comblera le fossé qui existe à l'heure actuelle entre les développements théoriques et les applications pratiques de la méthode de la pseudo-vraisemblance empirique et qu'il suscitera d'autres travaux de recherche dans ce domaine en vue de rendre l'utilisation de cette méthode entièrement pratique.

L'algorithme de calcul de l'estimateur du maximum de pseudo-vraisemblance empirique sous échantillonnage non stratifié et certaines remarques sur son implantation dans R/S-PLUS sont présentés à la section 2. L'algorithme de Wu (2004a) pour l'échantillonnage stratifié est discuté à la section 3. La construction de l'intervalle de confiance du rapport de pseudo-vraisemblance empirique, qui comprend l'établissement du profil de cette statistique, est décrite en détail à la section 4. Tous les exemples de code ou de fonction R figurent à l'annexe. Ils peuvent être téléchargés à partir de la page d'accueil personnelle de l'auteur à http://www.stats.uwaterloo.ca/~cbwu/paper.html. Ces codes et fonctions ont été testés lors de l'étude en simulation

1. Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo (Ontario), N2L 3G1, Canada. Courriel : cbwu@uwaterloo.ca.

décrite dans Wu et Rao (2004) et ont donné de très bons résultats.

2. Échantillonnage non stratifié

Considérons une population finie constituée de N unités identifiables. Associées à la i^{e} unité sont des valeurs de la variable étudiée, y_i, et un vecteur de variables auxiliaires, $\boldsymbol{x}_i$. Le vecteur de moyennes de population $\overline{\boldsymbol{X}} = N^{-1}\sum_{i=1}^{N}\boldsymbol{x}_i$ est connu. Soit $\{(y_i, \boldsymbol{x}_i), i \in s\}$ les données d'échantillon, où s est l'ensemble d'unités sélectionnées selon un plan de sondage complexe. Soit $\pi_i = P(i \in s)$ les probabilités de sélection et $d_i = 1/\pi_i$ les poids de sondage.

L'estimateur du maximum de pseudo-vraisemblance empirique de la moyenne de population $\overline{Y} = N^{-1}\sum_{i=1}^{N} y_i$ est calculé comme étant $\hat{\overline{Y}}_{\text{PEL}} = \sum_{i\in s}\hat{p}_i\, y_i$, où les poids $\hat{p}_i$ sont obtenus en maximisant la fonction de pseudo log-vraisemblance empirique

$$l_{ns}(\boldsymbol{p}) = n^* \sum_{i\in s} d_i^* \log(p_i) \qquad (2.1)$$

sous les contraintes

$$0 < p_i < 1,\ \sum_{i\in s} p_i = 1 \quad \text{et} \quad \sum_{i\in s} p_i\, \boldsymbol{x}_i = \overline{\boldsymbol{X}}. \qquad (2.2)$$

La fonction de pseudo-vraisemblance empirique originale proposée par Chen et Sitter (1999) est $l(\boldsymbol{p}) = \sum_{i\in s} d_i \log(p_i)$. La fonction de pseudo-vraisemblance empirique $l_{ns}(\boldsymbol{p})$ donnée par (2.1) a été utilisée par Wu et Rao (2004), où les $d_i^* = d_i / \sum_{i\in s} d_i$ sont les poids de sondage normalisés et n^* est la taille effective d'échantillon. L'estimateur ponctuel $\hat{\overline{Y}}_{\text{PEL}} = \sum_{i\in s}\hat{p}_i\, y_i$ reste le même pour l'une et l'autre version de la fonction de vraisemblance. Le rééchelonnement utilisé dans $l_{ns}(\boldsymbol{p})$ facilite la construction des intervalles de confiance des rapports de pseudo-vraisemblance empirique.

En utilisant comme argument un multiplicateur de Lagrange standard, nous pouvons montrer que

$$\hat{p}_i = \frac{d_i^*}{1 + \boldsymbol{\lambda}'(\boldsymbol{x}_i - \overline{\boldsymbol{X}})} \quad \text{pour} \quad i \in s, \qquad (2.3)$$

où le multiplicateur de la Lagrange évalué vectoriellement, $\boldsymbol{\lambda}$, est la solution de

$$g_1(\boldsymbol{\lambda}) = \sum_{i\in s} \frac{d_i^*(\boldsymbol{x}_i - \overline{\boldsymbol{X}})}{1 + \boldsymbol{\lambda}'(\boldsymbol{x}_i - \overline{\boldsymbol{X}})} = 0.$$

Ici, la principale tâche de calcul consiste à trouver la solution de $g_1(\boldsymbol{\lambda}) = 0$, ce qui peut se faire en utilisant la procédure de Newton-Raphson modifiée proposée par Chen et coll. (2002). La modification comprend la vérification, à chaque étape de mise à jour, que la contrainte $1 + \boldsymbol{\lambda}'(\boldsymbol{x}_i - \overline{\boldsymbol{X}}) > 0$ (*i.e.*, $p_i > 0$) est encore satisfaite. Sans perte de généralité, nous supposons que $\overline{\boldsymbol{X}} = 0$ (sinon il faut remplacer partout $\boldsymbol{x}_i$ par $\boldsymbol{x}_i - \overline{\boldsymbol{X}}$). La procédure modifiée est la suivante.

Étape 0 : Soit $\boldsymbol{\lambda}_0 = \boldsymbol{0}$. Fixer $k = 0$, $\gamma_0 = 1$ et $\varepsilon = 10^{-8}$.

Étape 1 : Calculer $\boldsymbol{\Delta}_1(\boldsymbol{\lambda}_k)$ et $\boldsymbol{\Delta}_2(\boldsymbol{\lambda}_k)$, où

$$\boldsymbol{\Delta}_1(\boldsymbol{\lambda}) = \sum_{i\in s} d_i^* \frac{\boldsymbol{x}_i}{1 + \boldsymbol{\lambda}'\boldsymbol{x}_i}$$

et

$$\boldsymbol{\Delta}_2(\boldsymbol{\lambda}) = \left\{ -\sum_{i\in s} d_i^* \frac{\boldsymbol{x}_i\, \boldsymbol{x}_i'}{(1 + \boldsymbol{\lambda}'\boldsymbol{x}_i)^2} \right\}^{-1} \boldsymbol{\Delta}_1(\boldsymbol{\lambda}).$$

Si $\|\boldsymbol{\Delta}_2(\boldsymbol{\lambda}_k)\| < \varepsilon$, arrêter l'algorithme et donner la valeur de $\boldsymbol{\lambda}_k$ dans le rapport; autrement, passer à l'étape 2.

Étape 2 : Calculer $\boldsymbol{\delta}_k = \gamma_k\, \boldsymbol{\Delta}_2(\boldsymbol{\lambda}_k)$. Si $1 + (\boldsymbol{\lambda}_k - \boldsymbol{\delta}_k)'\boldsymbol{x}_i \le 0$ pour tout i, poser que $\gamma_k = \gamma_k / 2$ et répéter l'étape 2.

Étape 3 : Poser que $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \boldsymbol{\delta}_k$, $k = k + 1$ et $\gamma_{k+1} = (k+1)^{-1/2}$. Passer à l'étape 1.

Dans l'algorithme original présenté par Chen et coll. (2002), l'étape 2 consiste aussi à vérifier une fonction objective duale connexe. Bien qu'elle soit nécessaire pour la preuve théorique de la convergence de l'algorithme, cette vérification n'est pas vraiment requise pour les applications pratiques.

La fonction R Lag2(u,ds,mu) peut être utilisée pour trouver la solution de $g_1(\boldsymbol{\lambda}) = 0$ quand le vecteur de variables auxiliaires $\boldsymbol{x}$ est de dimension m et que $m \ge 2$. Quand $\boldsymbol{x}$ est univarié, une méthode de bisection extrêmement simple et stable qui sera décrite bientôt devrait être utilisée. Soit n la taille d'échantillon. Les trois arguments requis sont la matrice de données u de dimensions $n \times m$, le vecteur de poids de sondage ds de dimension $n \times 1$ et le vecteur de moyennes de population mu de dimension $m \times 1$. La sortie de la fonction Lag2(u,ds,mu) donne la valeur de $\boldsymbol{\lambda}$ qui est la solution de $g_1(\boldsymbol{\lambda}) = 0$.

La fonction Lag2(u,ds,mu) ne fournira pas de solution si i) le vecteur moyen $\overline{\boldsymbol{X}}$ n'est pas un point intérieur de l'enveloppe convexe formée par $\{\boldsymbol{x}_i, i \in s\}$, ou que ii) la matrice $\sum_{i\in s} d_i\, \boldsymbol{x}_i\, \boldsymbol{x}_i'$ n'est pas de plein rang. Dans le cas (i), l'estimateur du maximum de pseudo-vraisemblance empirique n'existe pas. Ceci se produit avec une probabilité s'approchant de zéro à mesure que la taille d'échantillon n tend vers l'infini; dans le cas (ii), on peut envisager de supprimer certaines composantes des variables $\boldsymbol{x}$ de l'ensemble de contraintes (2.2) pour éliminer le problème de colinéarité.

Si la variable $\boldsymbol{x}$ est univariée, il en est de même du multiplicateur de Lagrange $\boldsymbol{\lambda}$ concerné. Dans ces conditions, nous devons résoudre $g_2(\lambda) = \sum_{i\in s} d_i^* x_i / (1 + \lambda x_i) = 0$ pour un scalaire λ, en supposant que $\overline{X} = 0$. Il existe

une solution unique si, et uniquement si, $\min\{x_i,\ i\in s\}< 0 < \max\{x_i,\ i\in s\}$. La solution, si elle existe, est comprise entre $L=-1/\max\{x_i,\ i\in s\}$ et $U=-1/\min\{x_i,\ i\in s\}$. En notant que $g_2(\lambda)$ est une fonction monotone décroissante pour $\lambda\in(L,\ U)$, l'algorithme le plus efficace et fiable pour résoudre $g_2(\lambda)=0$ est la méthode de bisection. La fonction Lag1(u,ds,mu) fait précisément cela, où les arguments requis sont u = $(x_1,\ ...,\ x_n)$, ds = $(d_1,\ ...,\ d_n)$ et mu = $\overline{\boldsymbol{X}}$. La sortie donne la solution de $g_2(\lambda)=0$.

La fonction Lag1(u,ds,mu) peut être utilisée conjuguée à l'approche de la pseudo-vraisemblance empirique étalonnée au moyen d'un modèle (PVEEM) de Wu et Sitter (2001) pour traiter les cas où la variable $\boldsymbol{x}$ comprend un nombre élevé de dimensions. L'approche PVEEM ne comporte qu'une seule variable de prédiction de dimension tirée d'un modèle de régression linéaire multiple et le problème connexe du multiplicateur de Lagrange est toujours unidimensionnel.

3. Échantillonnage stratifié

Soit $\{(y_{hi},\ \boldsymbol{x}_{hi}),\ i\in s_h,\ h=1,\ ...,\ H\}$ les données d'échantillon provenant d'un plan de sondage stratifié. Soit $d_{hi}^*=d_{hi}/\sum_{i\in s_h}d_{hi}$ les poids de sondage normalisés pour la strate $h,\ h=1,\ ...,\ H$. La fonction de pseudo-vraisemblance empirique sous échantillonnage stratifié définie par Wu et Rao (2004) est donnée par

$$l_{st}(\boldsymbol{p}_1,\ ...,\ \boldsymbol{p}_H)=n^*\sum_{h=1}^{H}W_h\sum_{i\in s_h}d_{hi}^*\log(p_{hi}),\quad (3.1)$$

où les $W_h=N_h/N$ sont les poids de strate et n^* est la taille totale effective d'échantillon telle que définie dans Wu et Rao (2004). La valeur de n^* n'est pas nécessaire pour l'estimation ponctuelle, mais cette constante de mise à l'échelle est requise pour la construction des intervalles de confiance. Soit $\overline{\boldsymbol{X}}$ le vecteur connu des moyennes de population pour les variables auxiliaires. L'estimateur du maximum de pseudo-vraisemblance empirique de la moyenne de population $\overline{Y}=\sum_{h=1}^{H}W_h\overline{Y}_h$ est défini comme étant $\hat{\overline{Y}}_{\mathrm{PEL}}=\sum_{h=1}^{H}W_h\sum_{i\in s_h}\hat{p}_{hi}\,y_{hi}$, où les $\hat{p}_{hi}$ maximisent $l_{st}(\boldsymbol{p}_1,\ ...,\ \boldsymbol{p}_H)$ sous l'ensemble de contraintes

$$p_{hi}>0,\ \sum_{i\in s_h}p_{hi}=1,\ h=1,\ ...,\ H$$

et

$$\sum_h W_h\sum_{i\in s_h}p_{hi}\,\boldsymbol{x}_{hi}=\overline{\boldsymbol{X}}.\quad (3.2)$$

Sous échantillonnage stratifié, la principale difficulté de calcul est due au fait que la sous-normalisation des poids (c'est-à-dire $\sum_{i\in s_h}p_{hi}=1$) a lieu au niveau de la strate, alors que les contraintes d'étalonnage (c'est-à-dire $\sum_h W_h\sum_{i\in s_h}p_{hi}\,\boldsymbol{x}_{hi}=\overline{\boldsymbol{X}}$) et la maximisation contrainte de la fonction de pseudo-vraisemblance empirique se font au niveau de la population. L'algorithme proposé par Wu (2004a) pour calculer les $\hat{p}_{hi}$ se déroule comme suit : soit l'augmentation de $\boldsymbol{x}_{hi}$ afin d'inclure les $H-1$ premières variables indicatrices de strate et l'augmentation de $\overline{\boldsymbol{X}}$ afin d'inclure $(W_1,\ ...,\ W_{H-1})$ en tant que ses $H-1$ premières composantes. Dans le cas où il n'existe aucune contrainte d'étalonnage, la variable $\boldsymbol{x}$ augmentée correspond aux $H-1$ variables indicatrices de strate uniquement et $\overline{\boldsymbol{X}}=(W_1,\ ...,\ W_{H-1})$. Il s'ensuit que l'ensemble de contraintes (3.2) est équivalent à

$$p_{hi}>0,\ \sum_{h=1}^{H}W_h\sum_{i\in s_h}p_{hi}=1$$

et

$$\sum_{h=1}^{H}W_h\sum_{i\in s_h}p_{hi}\,\boldsymbol{x}_{hi}=\overline{\boldsymbol{X}},\quad (3.3)$$

où la variable $\boldsymbol{x}$ est maintenant augmentée. Soit $\boldsymbol{u}_{hi}=\boldsymbol{x}_{hi}-\overline{\boldsymbol{X}}$. Il est facile, en utilisant comme argument un multiplicateur de Lagrange standard, de montrer que

$$\hat{p}_{hi}=\frac{d_{hi}^*}{1+\boldsymbol{\lambda}'\boldsymbol{u}_{hi}},$$

où $\boldsymbol{\lambda}$ évalué vectoriellement est la solution de

$$g_3(\boldsymbol{\lambda})=\sum_h W_h\sum_{i\in s_h}\frac{d_{hi}^*\,\boldsymbol{u}_{hi}}{1+\boldsymbol{\lambda}'\boldsymbol{u}_{hi}}=0.$$

La procédure de Newton-Raphson modifiée de la section 2 pour la résolution de $g_1(\boldsymbol{\lambda})=\mathbf{0}$ peut être utilisée pour résoudre $g_3(\boldsymbol{\lambda})=\mathbf{0}$. L'étape clé du calcul sous échantillonnage stratifié consiste à donner au fichier de données un format approprié pour pouvoir appeler directement la fonction R Lag2(u,ds,mu) utilisée pour l'échantillonnage non stratifié. Des exemples de codes R pour le faire figurent en annexe.

4. Construction des intervalles de confiance des rapports de pseudo-vraisemblance empirique

Bien que les algorithmes informatiques pour l'estimateur du maximum de vraisemblance pseudo empirique sous plans de sondage stratifié et non stratifié diffèrent quelque peu, la recherche des bornes inférieure et supérieure de l'intervalle de confiance du rapport de pseudo-vraisemblance empirique pour $\overline{Y}$ comporte le même type d'analyse de profil. Sous un plan de sondage non stratifié, l'intervalle de confiance de niveau $(1-\alpha)$ du rapport de pseudo-vraisemblance empirique de $\overline{Y}$ est construit de façon telle que

$$\{\theta\mid r_{ns}(\theta)<\chi_1^2(\alpha)\},\quad (4.1)$$

où $\chi_1^2(\alpha)$ est le quantile $1-\alpha$ d'une loi χ^2 à un degré de liberté. Le rapport des log pseudo-vraisemblances empiriques $r_{ns}(\theta)$ est donné par

$$r_{ns}(\theta) = -2\{l_{ns}(\tilde{\boldsymbol{p}}) - l_{ns}(\hat{\boldsymbol{p}})\},$$

où les $\hat{\boldsymbol{p}}$ maximisent $l_{ns}(\boldsymbol{p})$ sous l'ensemble de « contraintes standard » telles que (2.2) et les $\tilde{\boldsymbol{p}}$ maximisent $l_{ns}(\boldsymbol{p})$ sous les « contraintes standard » et une contrainte supplémentaire induite par le paramètre d'intérêt, $\bar{Y}$, c'est-à-dire

$$\sum_{i\in s} p_i\, y_i = \theta. \tag{4.2}$$

Pour calculer $\tilde{\boldsymbol{p}}$, il faut traiter (4.2) comme une composante supplémentaire de l'ensemble de « contraintes standard » pour chaque valeur fixée de θ, de sorte que le processus de maximisation soit essentiellement le même qu'auparavant.

Soit $(\hat{L}, \hat{U})$ l'intervalle donné par (4.1). La méthode de bisection que nous avons proposée pour trouver $\hat{L}$ et $\hat{U}$ est fondée sur les observations suivantes :

i) La valeur minimale de $r_{ns}(\theta)$ est atteinte à $\theta = \sum_{i\in s} \hat{p}_i\, y_i = \hat{\bar{Y}}_{\mathrm{PEL}}$. Dans ce cas, $\tilde{\boldsymbol{p}} = \hat{\boldsymbol{p}}$ et $r_{ns}(\theta) = 0$.

ii) L'intervalle $(\hat{L}, \hat{U})$ est borné par $(y_{(1)}, y_{(n)})$, où $y_{(1)} = \min\{y_i, i\in s\}$ et $y_{(n)} = \max\{y_i, i\in s\}$.

iii) Le rapport de pseudo-vraisemblance empirique $r_{ns}(\theta)$ est une fonction monotone décroissante pour $\theta \in (y_{(1)}, \hat{\bar{Y}}_{\mathrm{PEL}})$ et monotone croissante pour $\theta \in (\hat{\bar{Y}}_{\mathrm{PEL}}, y_{(n)})$.

Nous pouvons arriver à la conclusion iii) en notant que $l_{ns}(\hat{\boldsymbol{p}})$ ne fait pas intervenir θ et que $l_{ns}(\tilde{\boldsymbol{p}}) = n^* \sum_{i\in s} d_i^* \log(\tilde{p}_i)$ est typiquement une fonction concave de θ. Il est également possible de montrer cela en vérifiant directement $dr_{ns}(\theta)/d\theta$. Par exemple, dans le cas où n'intervient aucune information auxiliaire, les « contraintes standard » sont $p_i > 0$ et $\sum_{i\in s} p_i = 1$. Les $\hat{p}_i$ sont donnés par d_i^* et $\hat{\bar{Y}}_{\mathrm{PEL}} = \sum_{i\in s} d_i^*\, y_i$. Les $\tilde{p}_i$ sont calculés sous la forme

$$\tilde{p}_i = \frac{d_i^*}{1+\lambda(y_i - \theta)}, \tag{4.3}$$

où λ est la solution de

$$\sum_{i\in s} \frac{d_i^*\,(y_i - \theta)}{1+\lambda(y_i - \theta)} = 0. \tag{4.4}$$

En partant de (4.3) et (4.4), et en notant que $\sum_{i\in s} d_i^* / (1+\lambda(y_i - \theta)) = 1$, il est facile de montrer que

$$\frac{d}{d\theta} r_{ns}(\theta) = 2n^* \sum_{i\in s} \frac{d_i^* \{(d\lambda/d\theta)(y_i - \theta) - \lambda\}}{1+\lambda(y_i - \theta)} = -2n^*\lambda.$$

En réécrivant $d_i^*(y_i - \theta)$ sous la forme $d_i^*(y_i - \theta)$ $[\{1+\lambda(y_i - \theta)\} - \lambda(y_i - \theta)]$ et après certains regroupements dans (4.4), nous obtenons

$$\lambda \sum_{i\in s} \frac{d_i^*(y_i - \theta)^2}{1+\lambda(y_i - \theta)} = \sum_{i\in s} d_i^*\, y_i - \theta.$$

Il s'ensuit que $dr_{ns}(\theta)/d\theta = -2n^*\lambda < 0$ si $\theta < \sum_{i\in s} d_i^*\, y_i = \hat{\bar{Y}}_{\mathrm{PEL}}$ et $dr_{ns}(\theta)/d\theta > 0$ autrement.

Les exemples de codes pour trouver $(\hat{L}, \hat{U})$ quand aucune variable auxiliaire n'est utilisée figurent à l'annexe. Dans ces conditions, $\hat{p}_i = d_i^*$ et $\hat{\bar{Y}}_{\mathrm{PEL}} = \sum_{i\in s} d_i^*\, y_i = \hat{\bar{Y}}_H$ est l'estimateur de Hajek de $\bar{Y}$. Le processus d'établissement du profil consiste à trouver λ pour chaque valeur choisie de θ et à évaluer le rapport de pseudo-vraisemblance empirique $r_{ns}(\theta)$ en fonction de la valeur seuil de la loi χ_1^2 sous le niveau de confiance $1-\alpha$ souhaité. Si l'on utilise des données auxiliaires, il faut modifier le calcul de $r_{ns}(\theta)$ pour chaque valeur fixée de θ. L'algorithme de bisection pour trouver $\hat{L}$ et $\hat{U}$ demeure le même.

La taille effective d'échantillon n^* doit être connue pour calculer le rapport de pseudo-vraisemblance empirique $r_{ns}(\theta)$. Dans le cas des plans de sondage non stratifiés, on la calcule selon $n^* = \hat{S}_y^2 / \hat{V}(y)$, où

$$\hat{S}_y^2 = \frac{1}{N(N-1)} \sum_{i\in s} \sum_{j>i} \frac{(y_i - y_j)^2}{\pi_{ij}},$$

et

$$\hat{V}(y) = \frac{1}{N^2} \sum_{i\in s} \sum_{j>i} \frac{\pi_i\,\pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j}\right)^2,$$

où $e_i = y_i - \hat{\bar{Y}}_{\mathrm{HT}}$ et $\hat{\bar{Y}}_{\mathrm{HT}} = N^{-1}\sum_{i\in s} d_i\, y_i$. Consulter Wu et Rao (2004) pour plus de précisions. Le calcul de n^* comprend les probabilités de sélection de deuxième ordre π_{ij} qui peuvent poser un vrai défi si l'on utilise un plan de sondage PPT. Dans leur étude en simulation, Wu et Rao (2004) ont utilisé la méthode d'échantillonnage PPT de Rao-Sampford. Les fonctions R pour sélectionner un échantillon PPT selon cette méthode, ainsi que pour calculer les probabilités de sélection de deuxième ordre connexes peuvent être consultées dans Wu (2004b). Des fonctions R similaires sont également disponibles dans un progiciel R complémentaire appelé « pps » [pour *probability proportional to size*], rédigé par J. Gambino (2003), qui peut être téléchargé à la page d'accueil R à http://cran.r- project.org/ en cliquant sur l'option *packages*.

Remerciements

Cette étude a été financée par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada. L'auteur remercie un rédacteur associé de ses commentaires constructifs qui lui ont permis d'améliorer l'article.

Annexe : Codes R/S-PLUS

A1. Fonction R pour résoudre $g_1(\boldsymbol{\lambda}) = 0$.

Soit m le nombre de variables auxiliaires concernées et $m \geq 2$. Trois arguments sont requis dans la fonction Lag2(u,ds,mu) :

(1) u : la matrice de données de dimensions $n \times m$ avec $\boldsymbol{x}_i$ en tant que i^{e} ligne $i = 1, ..., n$;

(2) ds : le vecteur de poids de sondage de dimension $n \times 1$ constitué de $d_1, ..., d_n$;

(3) mu : le vecteur de moyennes de population de dimension $m \times 1$ $\overline{\boldsymbol{X}}$.

La sortie de la fonction est la solution de $g_1(\boldsymbol{\lambda}) = 0$.

```
Lag2<-function(u,ds,mu)
{
  n<-length(ds)
  u<-u-rep(1,n)%*%t(mu)
  M<-0*mu
  dif<-1
  tol<-1e-08
  while(dif>tol){
    D1<-0*mu
    DD<-D1%*%t(D1)
    for(i in 1:n){
      aa<-as.numeric(1+t(M)%*%u[i,])
      D1<-D1+ds[i]*u[i,]/aa
      DD<-DD-ds[i]*(u[i,]%*%t(u[i,]))/aa^2
          }
    D2<-solve(DD,D1,tol=1e-12)
    dif<-max(abs(D2))
    rule<-1
    while(rule>0){
      rule<-0
      if(min(1+t(M-D2)%*%t(u))<=0) rule<-rule+1
      if(rule>0) D2<-D2/2
          }
    M<-M-D2
  }
  return(M)
}
```

A2. Fonction R pour la résolution de $g_2(\lambda) = 0$.

Lorsque la variable x est univariée, la solution de $g_2(\lambda) = 0$ peut être obtenue au moyen d'une méthode de bisection simple et fiable. Les trois arguments requis pour la fonction Lag1(u,ds,mu) sont u = $(x_1, ..., x_n)$, ds = $(d_1, ..., d_n)$ et mu = $\overline{\boldsymbol{X}}$. La sortie est la solution de $g_2(\lambda) = 0$.

```
Lag1<-function(u,ds,mu)
{
  L<--1/max(u-mu)
  R<--1/min(u-mu)
  dif<-1
  tol<-1e-08
  while(dif>tol){
    M<-(L+R)/2
    glam<-sum((ds*(u-mu))/(1+M*(u-mu)))
    if(glam>0) L<-M
    if(glam<0) R<-M
    dif<-abs(glam)
  }
  return(M)
}
```

A3. Exemple de code pour l'échantillonnage stratifié.

Nous devons appeler la fonction Lag2(u,ds,mu) à partir de l'échantillonnage non stratifié. L'étape essentielle est la préparation du fichier de données afin de lui donner le format approprié. Soit

(1) n = $(n_1, ..., n_H)$ le vecteur de taille d'échantillon de strate;

(2) x la matrice de données avec $\boldsymbol{x}_{hi}$ comme vecteurs de ligne, $i = 1, ..., n_h, h = 1, ..., H$;

(3) ds = $(d^*_{11}, ..., d^*_{1n_1}, ..., d^*_{H1}, ..., d^*_{Hn_H})$, où les d^*_{hi} sont les poids de sondage initiaux normalisés pour la strate h;

(4) X le vecteur de moyennes de population connues;

(5) W = $(W_1, ..., W_H)$ le vecteur de poids de strate (c'est-à-dire $W_h = N_h / N$).

Les exemples de codes qui suivent montrent comment est trouvée la solution de $g_3(\boldsymbol{\lambda}) = \mathbf{0}$ (M de l'avant-dernière ligne du code qui suit) et comment sont calculés les $\hat{p}_{hi}$ (phi de la dernière ligne).

```
###
nst<-sum(n)
k<-length(n)-1
ntot<-rep(0,k)
  ntot[1]<-n[1]
  for(j in 2:k) ntot[j]<-ntot[j-1]+n[j]
ist<-matrix(0,nst,k)
  ist[1:n[1],1]<-1
  for(j in 2:k) ist[(ntot[j-1]+1):ntot[j],j]<-1
uhi<-cbind(ist,x)
mu<-c(W[1:k],X)
whi<-rep(W[1],n[1])
  for(j in 2:(k+1)) whi<-c(whi,rep(W[j],n[j]))
dhi<-whi*ds
M<-Lag2(uhi,dhi,mu)
phi<-as.vector(ds/(1+(uhi-rep(1,nst)%*%t(mu))%*%M))
###
```

A4. Exemple de code pour trouver l'intervalle de confiance du rapport de pseudo-vraisemblance empirique.

La recherche de la borne inférieure (LB) et de la borde supérieure (UB) de l'intervalle de confiance du rapport de vraisemblance empirique doit se faire séparément. Les codes qui suivent montrent comment se fait cette recherche dans le cas où l'on n'utilise aucune information auxiliaire. Si l'on utilise ce genre d'information, il faut modifier le calcul des rapports de pseudo-vraisemblance empirique concernés (elratio) en conséquence. Soit

(1) a = $1 - \alpha$ le niveau de confiance de l'intervalle souhaité;

(2) ys = $(y_1, ..., y_n)$ les données d'échantillon;

(3) ds = $(d^*_1, ..., d^*_n)$ les poids de sondage normalisés;

(4) YEL = $\sum_{i \in s} \hat{p}_i \, y_i$ (ici $\hat{p}_i = d^*_i$) ;

(5) nss la taille d'échantillon effective estimée n^*.

```
###
tol<-1e-08
cut<-qchisq(a,1)
###
t1<-YEL
t2<-max(ys)
dif<-t2-t1
while(dif>tol){
    tau<-(t1+t2)/2
    M<-Lag1(ys,ds,tau)
    elratio<-2*nss*sum(ds*log(1+M*(ys-tau)))
    if(elratio>cut) t2<-tau
    if(elratio<=cut) t1<-tau
    dif<-t2-t1
        }
UB<-(t1+t2)/2
###
t1<-YEL
t2<-min(ys)
dif<-t1-t2
while(dif>tol){
    tau<-(t1+t2)/2
    M<-Lag1(ys,ds,tau)
    elratio<-2*nss*sum(ds*log(1+M*(ys-tau)))
    if(elratio>cut) t2<-tau
    if(elratio<=cut) t1<-tau
    dif<-t1-t2
        }
LB<-(t1+t2)/2
###
```

Bibliographie

Chen, J., Sitter, R.R. et Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89, 230-237.

Hartley, H.O., et Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

Wu, C. (2004a). Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica*, 14, 1057-1067.

Wu, C. (2004b). R/S-PLUS Implementation of pseudo empirical likelihood methods under unequal probability sampling. Document de travail 2004-07, Department of Statistics and Actuarial Science, University of Waterloo.

Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the Americal Statistical Association*, 96, 185-193.

Wu, C., et Rao, J.N.K. (2004). Pseudo empirical likelihood ratio confidence intervals for complex surveys. Document de travail 2004-06, Department of Statistics and Actuarial Science, University of Waterloo.

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont fourni de l'aide ou ont fait la critique d'un article ou plus durant l'année 2005.

A.K Adhikary, *ISI Kolkata*
M. Battaglia, *ABT Associates*
J.-F. Beaumont, *Statistique Canada*
N. Billor, *Auburn University*
C. Boudreau, *Medical College of Wisconsin*
K. Brewer, *Australian National University*
F. Butar Butar, *Sam Houston State University*
D. Cantor, *Westat*
S.R. Chowdhury, *Westat, Inc.*
S.L. Christ, *University of North Carolina*
J. Chromy, *RTI International*
R. Courtemanche, *Institut de la statistique du Québec*
A. Dessertaine, *EDF R&D-OSIRIS - CLAMART*
P. Dick, *Statistique Canada*
P. Duchesne, *Université de Montréal*
J. Dumais, *Statistique Canada*
J. Eltinge, *Bureau of Labour Statistics*
M. Feder, *Research Triangle Institute*
R. Fisher, *U.S. Census Bureau*
O. Frank, *Stockholm University*
S.G. Heeringa, *Institute for Social Research, University of Michigan*
S. Haslett, *Massey University*
D. Heng-Yan Leung, *Singapore Management University*
K. Jae Kwang, *Yonsei University*
F. Jenkins, *Westat*
J. Jiang, *University of California at Davis*
J.K. Kim, *Yonsei University*
M. Kovačević, *Statistique Canada*
S. Laaksonen, *University of Helsinki and Statistics Finland*
P. Lahiri, *University of Maryland*
F. Lapointe, *Institut de la statistique du Québec*
M.D. Larsen, *Iowa State University*
P. Lavallée, *Statistique Canada*
H. Lee, *Westat, Inc.*
R. Lehtonen, *University of Jyvaskyla*
N.T. Longford, *SNTL*
L. Magee, *McMaster University*
T. Maiti, *Iowa State University*
D. Malec, *United States Bureau of the Census*
B. Mandall, *Ohio State University*
S. Matthews, *Statistique Canada*
D. Marker, *Westat, Inc.*
D. McCaffrey, *RAND*
C.E. M'Lan, *University of Connecticut*
J. Moore, *U.S. Bureau of the Census*
R. Munnich, *University of Tubingen*
J. Opsomer, *Iowa State University*
O. Phillips, *Statistique Canada*
M. Pratesi, *University of Pisa, Italy*
J. Reiter, *Duke Univesity*
R.H. Renssen, *Statistics Netherlands*
G. Roberts, *Statistique Canada*
I. Şchiopu-Kratina, *Statistique Canada*
C.J. Schwarz, *Simon Fraser University*
A. Scott, *University of Aucklan*
J. Sedransk, *Case Western University*
R. Sitter, *Simon Fraser University*
M. Sinclair, *U.S. Department of Labor*
A. Singh, *Statistique Canada*
T.W. Smith, *NORC*
J. Stec, *InteCap, Inc*
D.G. Steel, *University of Wollongong, Australia*
L. Stokes, *Southern Methodist University*
E. Stuart, *Mathematica Policy Research, Inc.*
A. Jr. Tersine, *United States Bureau of the Census*
R. Thomas, *Carleton University*
N. Thomas, *Pfizer, Inc.*
C. Tucker, *United States Bureau of Labor*
J. van der Brakel, *Statistics Netherlands*
S.L. Vartivarian, *Mathematica Policy Research, Inc.*
J. Wang, *Merck Research Labs, Merck & Co., Inc.*
X. Wang, *Southern Methodist University*
C. Wu, *University of Waterloo*
R. Yucel, *University of Massachusetts*
W. Yung, *Statistique Canada*
E. Zanutto, *University of Pennsylvania*
H. Zheng, *Massachusetts General Hospital and Harvard Medical School*

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2005: Francine Pilon-Renaud et Roberto Guido (Division de la diffusion), Marc Bazinet (Division du marketing) et François Beaudin (Division des langues officielles et traduction). Finalement nous désirons exprimer notre reconnaissance à Christine Cousineau, Céline Ethier, Nancy Flansberry et Denis Lemire de la Division des méthodes des enquêtes auprès des ménages, pour leur apport à la coordination, la dactylographie et la rédaction.


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. 2, 2005

Volume 21, No. 3, 2005

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 33, No. 2, June/juin 2005

Volume 33, No. 3, September/septembre 2005


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

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 $\exp(\cdot)$ et $\log(\cdot)$ *etc*.
3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
3.5 Distinguer clairement les caractères ambigus (comme w, ω ; o, O, 0; l, 1).
3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.

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.