Nº 12-001-XIF au catalogue

Techniques d'enquête

Décembre 2007

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 à www.statcan.ca.

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

Renseignements pour accéder ou commander le produit

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

Ce produit nº 12-001-XPF au catalogue est aussi disponible en version imprimée standard au prix de 30 $CAN l'exemplaire et de 58 $CAN pour un abonnement annuel.

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

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

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par

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

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

Normes de service à la clientèle

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

Statistique Canada

Division des méthodes d'enquêtes auprès des entreprises

Techniques d'enquête

Décembre 2007

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2008

Tous droits réservés. Le contenu de la présente publication électronique peut être reproduit en tout ou en partie, et par quelque moyen que ce soit, sans autre permission de Statistique Canada, sous réserve que la reproduction soit effectuée uniquement à des fins d'étude privée, de recherche, de critique, de compte rendu ou en vue d'en préparer un résumé destiné aux journaux et/ou à des fins non commerciales. Statistique Canada doit être cité comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, il est interdit de reproduire le contenu de la présente publication, ou de l'emmagasiner dans un système d'extraction, ou de le transmettre sous quelque forme ou par quelque moyen que ce soit, reproduction électronique, mécanique, photographique, pour quelque fin que ce soit, sans l'autorisation écrite préalable des Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Janvier 2008

Nº 12-001-XIF au catalogue
ISSN 1712-5685

No 12-001-XPB au catalogue
ISSN 0714-0045

Périodicité : semestriel

Ottawa

This publication is available in english upon request (catalogue no. 12-001-XIE)

Note de reconnaissance

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

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 33, numéro 2, décembre 2007

Table des matières

Dans ce numéro 109

Article Sollicité Waksberg

Carl-Erik Särndal
La méthode de calage dans la théorie et la pratique des enquêtes 113

Articles Réguliers

Seppo Laaksonen
Pondération de données d'enquête recueillies en deux phases 137

Pascal Ardilly et Pierre Lavallée
Pondération dans les échantillons rotatifs : le cas de l'enquête SILC en France 149

Jay J. Kim, Jianzhu Li et Richard Valliant
Regroupement de cellules lors de la poststratification 157

Fulvia Mecatti
Un estimateur à base de sondage unique fondé sur la multiplicité pour les sondages à bases multiples 171

David Haziza
Estimation de la variance pour un ratio en présence de données imputées 179

James Chipperfield et John Preston
Bootstrap efficace pour les enquêtes-entreprises 187

Jacob J. Oleson, Chong Z. He, Dongchu Sun et Steven L. Sheriff
Estimation bayésienne pour de petites régions en cas de différence entre les strates du plan d'échantillonnage et les domaines d'étude 195

Enrico Fabrizi, Maria Rosaria Ferrante et Silvia Pacei
Estimation pour petits domaines du revenu moyen des ménages en fonction des modèles au niveau des unités pour les données d'enquêtes par panel 209

Anne Renaud
Estimation de la couverture du recensement de la population de l'an 2000 en Suisse : méthodes et résultats 221

Marc N. Elliott et Amelia Haviland
Utilisation d'un échantillon de convenance électronique comme complément à un échantillon probabiliste 233

Remerciements 239

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.
∞

Le papier utilisé dans la présente publication répond aux exigences minimales de l'"American National Standard for Information Sciences" – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.
∞

Techniques d'enquête, décembre 2007
Vol. 33, N° 2, pp. 109-110
Statistique Canada, N° 12-001-X au catalogue

Dans ce numéro

Ce numéro de *Techniques d'enquête* s'ouvre sur le septième article de la série annuelle sur invitation en l'honneur de Joseph Waksberg. Le comité éditorial aimerait remercier les membres du comité de sélection, composé de Gordon Brackstone, président, et de Bob Groves, Sharon Lohr et Wayne Fuller, d'avoir choisi l'article de Carl-Erik Särndal qui gagné le prix Waksberg de cette année. Pour souligner la réalisation, un Atelier spécial sur le calage et l'estimation dans les enquêtes (ACEE) a été organisé le 31 octobre et le 1er novembre à Statistique Canada. Le conférencier principal, le professeur Carl-Erik Särndal, a présenté l'article pour lequel il a remporté le prix Waksberg. Durant ces deux journées, douze autres conférenciers ont présenté un article et ont rendu hommage à M. Särndal.

Dans son article intitulé *L'approche par calage dans la théorie et la pratique d'enquêtes*, Särndal se penche sur le développement et l'utilisation du calage dans les enquêtes par échantillonnage. Il y décrit en détail le concept de calage et l'oppose à la régression généralisée. Il propose ensuite une description des différentes techniques de calage, notamment la méthode de la distance minimale, les variables instrumentales et le calage par modèle. Plusieurs exemples de calage sont présentés, de même que des solutions de rechange.

Pour sa part, Laaksonen aborde la pondération dans les sondages à deux phases, où on demande aux répondants ayant participé à la première phase s'ils acceptent de participer à la deuxième. La pondération doit dès lors tenir compte des non-réponses aux deux phases ainsi que des répondants de la première phase qui refusent de participer à la seconde. À l'aide de données tirées d'une enquête finlandaise sur les activités de loisir, Laaksonen procède à une évaluation empirique des variations qui surviennent avec une méthode de pondération faisant appel à la modélisation de la propension de réponse et au calage.

Dans l'article de Ardilly et Lavallée, on traite le problème de la pondération pour l'enquête SILC (*Statistics on Income and Living Conditions*) en France. Cette enquête utilise un plan de sondage rotatif avec neuf panels. Pour obtenir des estimateurs approximativement sans biais, les auteurs ont recours à la méthode du partage des poids. On discute d'abord de pondération longitudinale et ensuite de pondération transversale.

L'article de Kim, Li et Valliant porte sur le problème des cellules avec petits chiffres ou des importantes corrections de poids lorsque la stratification a posteriori est utilisée. Les auteurs décrivent d'abord plusieurs estimateurs types, puis présentent deux estimateurs de rechange fondés sur le regroupement de cellules. Ils étudient le rendement de ces estimateurs d'après leur efficacité à contrôler le biais de couverture et la variance de conception. Ces propriétés font l'objet d'une évaluation théorique et d'une étude de simulation utilisant une population fondée sur la National Health Interview Survey de 2003.

Ensuite, Mecatti propose un simple estimateur de multiplicité dans un contexte de sondages à bases multiples. Pour commencer, elle montre que l'estimateur proposé ne présente aucun biais par rapport au plan de sondage, puis elle propose un estimateur sans biais de la variance de l'estimateur de multiplicité. À l'aide de 29 populations simulées, elle compare l'estimateur de multiplicité aux estimateurs de rechange proposés dans différentes études.

Haziza se penche sur le problème de l'estimation de la variance pour un rapport de deux totaux lorsque l'imputation hot-deck marginale aléatoire a été utilisée pour remplacer les données manquantes. Deux approches d'inférence sont étudiées : la première fait appel à un modèle d'imputation et la seconde, à un modèle de non-réponse. Les estimateurs de la variance proviennent de deux cadres de travail : l'approche inversée de Shao et Steel (1999) et l'approche traditionnelle en deux phases.

Dans leur article, Chipperfield et Preston décrivent l'estimateur bootstrap pondéré de la variance sans remplacement, qui a été appliqué par le Bureau de la statistique de l'Australie pour son système généralisé d'estimations ABSEST. Ils démontrent que l'estimateur bootstrap pondéré sans remplacement est plus efficace que l'estimateur bootstrap pondéré avec remplacement dans le cas d'échantillons stratifiés, lorsque les tailles des strates sont restreintes. En outre, ils démontrent que l'estimateur bootstrap pondéré sans remplacement nécessite moins de répétitions que l'estimateur avec remplacement pour atteindre la même erreur. Pour le système ABSEST, les estimateurs bootstrap de la variance ont été privilégiés à d'autres méthodes d'estimation de la variance en raison de leur rendement en calcul, tandis que l'estimateur bootstrap sans remplacement a été choisi pour les raisons exposées plus haut.

Pour leur part, Oleson, He et Sun décrivent une approche de modélisation bayésienne pour les situations où le plan d'échantillonnage est stratifié et où la procédure d'estimation nécessite une stratification a posteriori. La méthode est illustrée à l'aide de données tirées de l'enquête de 1998 sur la chasse aux dindons au Missouri, pour laquelle les strates étaient définies selon le lieu de résidence des chasseurs mais où les estimations étaient requises au niveau des comtés.

Fabrizi, Ferrante et Pacei se penchent sur une méthodologie qui prend de plus en plus d'importance dans les logiciels modernes d'enquête-échantillon. Ils étudient l'effet de l'emprunt d'information à un panel supplémentaire pour les estimations transversales de revenu des ménages dans de petites régions d'Italie. Les méthodes proposées abordent un problème qui pourrait s'avérer pertinent pour les statistiques européennes officielles, et possiblement dans le domaine des statistiques pour petits domaines, pour les indicateurs pouvant servir à la recherche stratégique.

Renaud présente une application intéressante d'une enquête postérieure au recensement pour évaluer le sous-dénombrement net du Recensement de 2000 en Suisse. L'objectif de cette enquête était légèrement différent de celui d'autres pays, en ce qu'il n'était pas conçu pour ajuster les chiffres du recensement au sous-dénombrement net, mais plutôt pour recueillir des renseignements visant à améliorer la qualité des recensements ultérieurs.

Dans le dernier article, Elliot et Haviland envisagent de combiner un échantillon de commodité à un échantillon probabiliste pour obtenir un estimateur présentant une erreur quadratique moyenne (EQM) plus faible. L'estimateur ainsi obtenu est un agencement linéaire des estimateurs de l'échantillon de commodité et de l'échantillon probabiliste, les poids étant fonction du biais. À partir de la contribution progressive maximale de l'échantillon de commodité, les auteurs démontrent qu'il est possible d'améliorer l'EQM uniquement dans certaines situations.

Harold Mantel, Rédacteur en chef délégué

Techniques d'enquête, décembre 2007
Vol. 33, N° 2, pp. 111-112
Statistique Canada, N° 12-001-X 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. La liste des gagnants précédents est présentée ci-bas. Leurs articles sont déjà parus dans la revue *Techniques d'enquête*.

Gagnants précédents du prix Waksberg :

Gad Nathan (2001)
Wayne A. Fuller (2002)
Tim Holt (2003)
Norman Bradburn (2004)
J.N.K. Rao (2005)
Alastair Scott (2006)
Carl-Erik Särndal (2007)

Nominations :

L'auteur de l'article Waksberg de 2009 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 Robert Groves, président du comité, par courriel à bgroves@isr.umich.edu. Les candidatures et les suggestions de sujets doivent être reçues d'ici au 29 février 2008.

Article sollicité Waksberg 2007

Auteur : Carl-Erik Särndal

Carl-Erik Särndal, professeur retraité de l'Université de Montréal, est un consultant et expert qui a été associé à plusieurs instituts nationaux de statistique, en particulier Statistique Canada et Statistique Suède, de même que Statistique Finlande, l'INSEE et Eurostat. Sa liste de publications comprend trois livres importants, dont le livre très renommé Model Assisted Survey Sampling qui a eu un grand impact. Il est aussi l'auteur d'un grand nombre d'articles scientifiques, comme seul auteur ou en collaboration avec des chercheurs de plusieurs pays. Son intérêt pour la recherche en sondages a été très diversifiée, mais s'est le plus souvent rapportée aux meilleures façons d'utiliser l'information auxiliaire en échantillonnage et estimation.

Membres du comité de sélection de l'article Waskberg (2007-2008)

Robert Groves, (Président)
Wayne A. Fuller, *Iowa State University*
Daniel Kasprzyk, *Mathematica Policy Research*
Leyla Mojadjer, *Westat*

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)
Gordon Brackstone (2005 - 2006)
Sharon Lohr (2006 - 2007)

Techniques d'enquête, décembre 2007
Vol. 33, N° 2, pp. 113-135
Statistique Canada, N° 12-001-X au catalogue

La méthode de calage dans la théorie et la pratique des enquêtes

Carl-Erik Särndal [1]

Résumé

Le calage est le thème central de nombreux articles récents sur l'estimation dans le contexte de l'échantillonnage. Des expressions telles que « méthode de calage » et « estimateur par calage » sont fréquentes. Comme tiennent à le souligner les auteurs de ces articles, le calage offre un moyen systématique d'intégrer des données auxiliaires dans la procédure.

Le calage est devenu un instrument méthodologique important dans la production de statistiques à grande échelle. Plusieurs organismes statistiques nationaux ont conçu des logiciels de calcul des poids, qui sont généralement calés sur les données auxiliaires disponibles dans les registres administratifs et d'autres sources de données fiables.

Le présent article fait le point sur la méthode de calage en mettant l'accent sur les progrès accomplis depuis une dizaine d'années. Le nombre d'études sur le calage augmente rapidement et nous abordons ici certaines des questions soulevées.

L'article débute par une définition de la méthode de calage, suivie d'une revue des caractéristiques importantes de cette méthode. L'estimation par calage est comparée à l'estimation par la régression (généralisée), qui est un autre moyen, conceptuellement différent, de tenir compte de l'information auxiliaire. Vient ensuite une discussion des aspects mathématiques du calage, y compris les méthodes permettant d'éviter les poids extrêmes. Dans les premières sections sont décrites des applications simples de la méthode, c'est-à-dire l'estimation d'un total de population sous échantillonnage direct, à une seule phase. Puis est envisagée la généralisation à des paramètres et à des plans d'échantillonnage plus complexes. Un trait commun de ces plans (à au moins deux phases ou deux degrés) est que l'information auxiliaire disponible peut comporter plusieurs composantes ou couches. L'application du calage dans de tels cas d'information composite est passée en revue. Plus loin, des exemples sont donnés pour illustrer comment les résultats de l'approche du calage peuvent différer de ceux obtenus grâce aux approches établies antérieurement. Enfin sont discutées des applications du calage en présence d'erreurs non dues à l'échantillonnage, en particulier les méthodes de correction du biais de non-réponse.

Mots clés : Conhérence; estimateur par régression; inférence basée sur le plan; information auxiliaire; modèles; non-réponse; plan de sondage complexe; pondération.

1. Introduction

1.1 Définition du calage

Il serait bon, pour les besoins du présent article, de faire référence à une définition du calage. Voici celle proposée ici.

Définition. L'application de la *méthode de calage* à l'estimation pour population finie consiste à :

a) calculer des poids qui tiennent compte de l'*information auxiliaire* spécifiée et sont soumis à des contraintes précisées par une ou plusieurs *équations de calage*;

b) utiliser ces poids pour calculer des estimations linéairement pondérées des totaux et d'autres paramètres de population finie, c'est-à-dire multiplier la valeur de la variable par le poids et faire la somme sur un ensemble d'unités observées;

c) se fixer l'objectif d'obtenir des estimations presque sans biais sous le plan de sondage, à condition qu'il n'y ait pas d'erreur de non-réponse ni d'autres erreurs non dues à l'échantillonnage.

Dans la littérature, le terme « calage » fait souvent référence à la partie a) seulement. Ici, le terme sera souvent appliqué aux parties a) à c) regroupées. Les définitions antérieures, quoique moins générales, concordent essentiellement avec la présente définition. Ardilly (2006) définit le calage (ou, plus précisément, le « calage généralisé ») comme une méthode de repondération utilisée lorsqu'on a accès à plusieurs variables, qualitatives ou quantitatives, sur lesquelles on souhaite effectuer, conjointement, un ajustement.

Kott (2006) définit les poids de calage comme un ensemble de poids, pour les unités de l'échantillon, qui sont calées sur les totaux connus de population de sorte que l'estimateur résultant soit convergent dans des conditions de randomisation (convergent sous le plan) ou, de manière plus rigoureuse, que le biais par rapport au plan de sondage représente, sous des conditions faibles, un apport asymptotiquement non significatif à l'erreur quadratique moyenne de l'estimateur, propriété appelée ici « presque sans biais sous le plan de sondage ».

La quatrième édition des Lignes directrices concernant la qualité de Statistique Canada (2003) dit ce qui suit : « Le *calage aux marges* est une procédure qu'on peut appliquer pour incorporer des données auxiliaires. Cette procédure rajuste les poids d'échantillonnage au moyen de multiplicateurs appelés les *facteurs de calage*, lesquels font correspondre les estimations aux totaux connus. Les poids

1. Carl-Erik Särndal, 2115 Erinbrook Crescent, #44, Ottawa, (Ontario), K1B 4J5, Canada. Courriel : carl.sarndal@rogers.com.

obtenus sont appelés les *poids de calage* ou les *poids d'estimation finaux*. De façon générale, ces poids de calage donneront des estimations convergentes qui admettront une variance d'échantillonnage plus faible que celle obtenue au moyen de l'estimateur Horvitz Thompson. »

La partie c) de la définition appelle un commentaire. Rien n'empêche de produire des poids calés sur des données auxiliaires données sans que c) soit nécessaire. Toutefois, la plupart des travaux publiés sur le calage sont dans l'esprit de c), d'où l'intérêt de l'inclure. En présence d'erreurs non dues à l'échantillonnage, les estimations sont inévitablement entachées d'un biais, qu'elles soient produites par calage ou par toute autre méthode. Allant dans le sens de c), l'inférence fondée sur le plan de sondage est considérée comme étant la situation de référence dans le présent article. La variance d'un estimateur par rapport au plan de sondage est donc une caractéristique importante. Toutefois, l'article est axé sur les « raisons qui motivent l'estimation (ponctuelle) » et faute d'espace, l'importante question de l'estimation de la variance n'y est pas abordée.

1.2 Commentaires

La définition de la section 1.1 suscite certains commentaires et des renvois aux études publiées antérieurement.

1) *Le calage comme méthode de pondération linéaire*. Le calage est intimement lié à la pratique. La fixation des grands organismes statistiques nationaux sur les méthodes de pondération est une puissante force poussant le calage. Attribuer un poids approprié à une valeur observée d'une variable et faire la sommation des valeurs pondérées de la variable pour former des agrégats appropriés est une pratique fermement enracinée. Elle est utilisée par les organismes statistiques pour estimer divers paramètres descriptifs de population finie, dont les totaux, les moyennes et les fonctions de totaux. La pondération est facile à expliquer aux utilisateurs des données et aux autres intervenants des organismes statistiques.

La pondération des unités par l'inverse de leur probabilité d'inclusion trouve depuis longtemps ses fondements scientifiques dans des articles tels que ceux de Hansen et Hurwitz (1943) et d'Horwitz et Thompson (1952). La pondération a fini par être généralement reconnue. Plus tard, la pondération par poststratification l'a été à son tour. La pondération par calage s'inscrit dans le prolongement de ces deux idées. La pondération par calage est fonction du résultat et les poids dépendent de l'échantillon observé.

Par définition, les poids correspondant à l'inverse de la probabilité d'inclusion sont égaux ou supérieurs à l'unité. Une interprétation courante est qu'une unité observée est représentative d'elle-même et d'un certain nombre d'autres unités, non observées. En revanche, les poids calés ne sont pas nécessairement égaux ou supérieurs à l'unité, à moins que l'on ne prenne tout spécialement soin dans les calculs d'obtenir cette propriété.

Le terme « calage » est nouveau en échantillonnage - il remonte à une quinzaine d'années - mais l'application de la méthode pour produire des pondérations ne l'est pas. Il y a du vrai dans ce que disent ceux qui affirment avoir fait du calage bien avant qu'on l'ait baptisé ainsi. La méthode a gagné en portée et en attrait au cours des 15 dernières années. Une méthode de pondération voisine du calage est utilisée de longue date par les instituts de sondage privés, par exemple, dans le contexte de l'échantillonnage par quota, forme d'échantillonnage probabiliste qui dépasse le cadre du présent exposé.

La pondération des valeurs observées d'une variable était déjà un sujet important avant que le calage devienne à la mode. Certains auteurs calculaient les pondérations en posant qu'elles devaient s'écarter aussi peu que possible des poids de sondage sans biais (c'est-à-dire l'inverse des probabilités d'inclusion). D'autres obtenaient les poids en reconnaissant qu'un estimateur par la régression linéaire pouvait s'écrire sous la forme d'une somme linéairement pondérée des valeurs observées de la variable étudiée. Ils ont utilisé des expressions telles que « pondération par les poids de sondage » (*survey sample weighting*), « pondération par régression » (*regression weighting*) et « pondération de cas » (*case weighting*). Ces « premiers articles » comptent Alexander (1987), Bankier, Rathwell et Majkowski (1992), Bethlehem et Keller (1987), Chambers (1996), Fuller, Loughin et Baker (1994), Kalton et Flores-Cervantes (1998), Lemaître et Dufour (1987), Särndal (1982) et Zieschang (1990). La méthode de « pondération répétée », avancée par l'organisme statistique national des Pays-Bas (CBS) sera commentée plus loin. Le terme « calage », plus récent, communique un message plus précis et une orientation plus catégorique que le terme « pondération ».

2) *Le calage comme moyen systématique d'utiliser l'information auxiliaire*. Le calage représente un moyen systématique de tenir compte des données auxiliaires. Comme le souligne Rueda et coll. (2007), dans de nombreuses conditions ordinaires, le calage offre un moyen simple et pratique d'intégrer l'information auxiliaire dans l'estimation.

L'information auxiliaire a été utilisée pour améliorer l'exactitude des estimations par sondage bien avant que le calage gagne en popularité. De nombreux articles ont été rédigés en poursuivant cet objectif, dans le contexte de situations plus ou moins spécialisées. Aujourd'hui, le calage donne une vue systématique des utilisations de l'information auxiliaire. Par exemple, il permet de traiter efficacement les données d'enquêtes pour lesquelles l'information auxiliaire existe à divers niveaux. Dans

l'échantillonnage à deux degrés, des données peuvent exister pour les unités d'échantillonnage de premier degré (grappes) et d'autres, pour les unités d'échantillonnage de deuxième degré. Dans les enquêtes avec non réponse (c'est-à-dire essentiellement toutes les enquêtes), des données peuvent exister « au niveau de la population » (totaux de population connus) et d'autres, « au niveau de l'échantillon » (valeurs des variables auxiliaires pour toutes les unités échantillonnées, répondantes et non répondantes). Le calage à l'aide de « données composites » est examiné aux sections 8 et 9.

L'estimation par la régression, ou estimation par la régression généralisée (GREG) fait concurrence au calage en tant que moyen systématique d'intégrer l'information auxiliaire. Il importe donc de faire la distinction entre l'estimation GREG (décrite à la section 3) et l'estimation par calage (décrite à la section 4). Les deux approches sont différentes.

3) *Le calage pour obtenir la convergence*. Le calage est souvent décrit comme un moyen d'obtenir des estimations convergentes. (Ici « convergence » ne signifie pas « convergence sous la plan d'échantillonnage (randomisation) », mais « convergence avec des agrégats connus ».) Les *équations de calage* imposent la convergence sur le système de poids, de sorte que, lorsqu'ils sont appliqués aux variables auxiliaires, ils confirment (concordent avec) les agrégats connus pour ces mêmes variables auxiliaires. Le désir de rendre les statistiques publiées plus crédibles est dans bien des cas le motif cité de la recherche de convergence. Certains utilisateurs des statistiques n'aiment pas constater qu'une même grandeur de population est estimée par deux chiffres ou plus qui ne concordent pas.

Les totaux par rapport auxquels est recherchée la convergence sont parfois appelés totaux de contrôle. Des expressions comme « poids contrôlés » ou « poids calés » donnent une impression d'estimation améliorée, plus exacte. Le terme « calage » a aussi la connotation de « stabilité ».

Obtenir la convergence par calage a une incidence plus générale que la simple concordance avec des totaux de population auxiliaires connus. On pourrait, par exemple, rechercher la convergence avec des totaux *estimés* de manière appropriée, d'après les données de l'enquête courante ou d'autres enquêtes.

La convergence entre les tableaux estimés d'après des enquêtes différentes est la raison qui motive la *pondération répétée*, méthode mise au point par l'organisme statistique national des Pays-Bas (CBS) et exposée dans plusieurs articles, dont Renssen et Nieuwenbroek (1997), Nieuwenbroek, Renssen et Hofman (2000), Renssen, Kroese et Willeboordse (2001), ainsi que Knottnerus et van Duin (2006). L'objectif énoncé est de répondre aux demandes des utilisateurs de produire des données de sortie numériquement convergentes. Comme le soulignent les auteurs du dernier article mentionné, la pondération répétée peut être considérée comme une étape de calage supplémentaire en vue d'un nouvel ajustement des poids déjà calés. Les poids finaux réalisent la convergence avec les valeurs de marge données.

La convergence avec des totaux connus ou estimés peut avoir l'avantage supplémentaire d'améliorer l'exactitude (réduction de la variance et/ou du biais de non réponse). Toutefois, dans certains articles, particulièrement ceux publiés par les organismes statistiques, la convergence en vue de satisfaire les utilisateurs semble un motif plus impératif que la perspective d'une plus grande exactitude.

Si la principale motivation du calage n'est pas tant la concordance avec d'autres statistiques que la réduction de la variance et/ou du biais de non-réponse, l'expression « système de poids équilibrés » est une description plus appropriée que « système de poids convergents », parce que l'objectif est alors d'équilibrer les poids afin de refléter le résultat de l'échantillonnage, la réponse à l'enquête et l'information disponible.

4) *Calage de commodité et de transparence*. Comme le font remarquer Harms et Duchesne (2006), le recours à la méthode de calage s'est répandu dans les applications réelles, parce que les estimations résultantes sont faciles à interpréter et à justifier puisqu'elles s'appuient sur les poids de sondage et des contraintes de calage naturelles. À l'utilisateur moyen, le calage sur des totaux connus paraît transparent et naturel. Les utilisateurs qui comprennent la pondération d'échantillon apprécient le fait que le calage « ne fait que modifier légèrement » les poids de sondage, tout en respectant les valeurs de contrôle. L'absence de biais n'est perturbée que de manière négligeable. Les formes de calage les plus simples ne font appel à aucune hypothèse et ne s'appuient que sur des « contraintes naturelles ». Un autre avantage qu'apprécient les utilisateurs est que, dans de nombreuses applications, le calage produit un système de poids unique, applicable à toutes les variables étudiées, lesquelles sont habituellement nombreuses dans le cas des grandes enquêtes menées par les organismes statistiques publics.

5) *Le calage combiné à d'autres termes*. Certains auteurs utilisent le mot « calage » en combinaison avec d'autres termes afin de décrire diverses lignes de pensée. Voici des exemples de cette prolifération d'expressions : calage basé sur un modèle (Wu et Sitter 2001), calage *g* (Vanderhoeft, Waeytens et Museux 2000), calage harmonisé (Webber, Latouche et Rancourt 2000), calage de plus haut niveau (Singh, Horn et Yu 1998), calage par régression (Demnati et

Rao 2004), calage non linéaire (Plikusas 2006), calage supergénéralisé (Ardilly 2006), estimateur par calage fondé sur un modèle de réseau neuronal et estimateur par calage fondé sur un modèle polynomial local (Montanari et Ranalli 2003, 2005), estimateur du pseudo-maximum de vraisemblance empirique calé sur un modèle (Wu 2003), et ainsi de suite. En outre, le calage joue un rôle important dans les méthodes de sondage indirectes proposées dans Lavalleé (2006). Dans un esprit légèrement différent, certains auteurs proposent des concepts, que nous n'examinerons pas ici, tels que l'imputation calée (Beaumont 2005a) ou le calage du biais (Chambers, Dorfman et Wehrly 1993, Zheng et Little 2003). Les pages qui suivent ne rendent pas justice à toutes les innovations qui ont lieu dans le domaine du calage, mais l'énumération des noms des méthodes donne à elle seule une idée des diverses voies qui ont été explorées.

(6) *Le calage en tant que nouvelle ligne de pensée*. Si le calage constitue « une nouvelle approche » qui se distingue nettement de celles qui l'ont précédée, nous devons nous poser des questions telles que : le calage généralise-t-il les théories ou les approches antérieures? Le calage fournit-il de meilleures réponses, plus satisfaisantes, aux questions importantes que les approches reconnues antérieurement? Aux sections 4.5 et 7.1, nous illustrons dans quelle mesure les réponses offertes par le calage sont comparables ou différent de celles résultant des courants de pensée antérieurs.

Le praticien de l'échantillonnage se heurte à des « nuisances », telles que la non-réponse, les déficiences des bases de sondage et les erreurs de mesure. Certes, l'imputation et la repondération pour corriger la non-réponse sont d'usage répandu et se font selon une foule de méthodes. Mais il s'agit dans une certaine mesure de « problèmes distincts », qui attendent encore d'être intégrés plus pleinement dans une théorie générale, plus satisfaisante, de l'inférence dans le contexte des sondages. Nombre d'articles théoriques traitent de l'estimation dans le cas d'un sondage idéal imaginaire, inexistant dans la pratique, qui ne souffre pas de la non-réponse et d'autres erreurs non dues à l'échantillonnage. Le propos n'est pas de critiquer toutes ces études théoriques excellentes, mais idéalisantes. Mais il faut aussi explorer les fondements.

Les sections 9 et 10 indiquent que le calage peut offrir une perspective plus systématique de l'inférence dans les sondages, même en présence de diverses erreurs non dues à l'échantillonnage. De fructueux développement sont à prévoir à cet égard.

2. Conditions de base pour l'estimation fondée sur le plan dans les sondages

La présente section a pour but d'établir le contexte des sections 3 à 7. Par « conditions de base », nous entendons ici un échantillonnage probabiliste d'éléments à une seule phase et une réponse complète. Les conditions réelles de sondage ne sont pas aussi simples et parfaites, mais de nombreux articles théoriques traitent néanmoins de cette situation.

Soit un échantillon probabiliste s tiré de la population finie $U = \{1, 2, ..., k, ..., N\}$. Le plan d'échantillonnage probabiliste produit, pour l'élément k, une probabilité d'inclusion connue, $\pi_k > 0$, et un poids de sondage correspondant $d_k = 1/\pi_k$. La valeur y_k de la variable étudiée y est enregistrée pour tous les $k \in s$ (réponse complète). L'objectif est d'estimer un total de population $Y = \sum_U y_k$ en se servant de l'information auxiliaire. La variable étudiée y peut être continue ou, comme dans le cas de nombreuses enquêtes menées par des organismes publics, catégorique. Par exemple, si y est dichotomique et prend la valeur $y_k = 0$ ou $y_k = 1$ selon que la personne k est occupée ou chômeuse, le paramètre $Y = \sum_U y_k$ à estimer est le chiffre de population des chômeurs. (Si $A \subseteq U$ est un ensemble d'éléments, nous écrivons $\sum_A$ pour $\sum_{k \in A}$.) L'estimateur de base sans biais sous le plan de Y est l'estimateur d'Horwitz-Thompson $\hat{Y}_{\mathrm{HT}} = \sum_s d_k y_k$. Cependant, il est inefficace si des données auxiliaires puissantes sont disponibles à l'étape de l'estimation.

La notation générale du vecteur auxiliaire sera $\mathbf{x}_k$. Dans certains pays, pour certaines enquêtes, les sources de données auxiliaires permettent de construire des vecteurs $\mathbf{x}_k$ de grande portée. Voici néanmoins quelques exemples de vecteurs simples : 1) $\mathbf{x}_k = (1, x_k)'$, où x_k est la valeur pour l'élément k d'une variable auxiliaire continue x; 2) le vecteur de classification utilisé pour coder l'appartenance à l'un de P groupes mutuellement exclusifs et exhaustifs, $\mathbf{x}_k = \boldsymbol{\gamma}_k = (\gamma_{1k}, ..., \gamma_{pk}, ..., \gamma_{Pk})'$, de sorte que, pour $p = 1, 2, ..., P$, $\gamma_{pk} = 1$ si k appartient au groupe p, et $\gamma_{pk} = 0$ autrement; 3) la combinaison de 1) et 2), $\mathbf{x}_k = (\boldsymbol{\gamma}_k', x_k \boldsymbol{\gamma}_k')'$; 4) le vecteur $\mathbf{x}_k$ qui codifie deux classifications déployées « côte à côte », la dimension de $\mathbf{x}_k$ étant $P + Q$ - 1, où P et Q sont les nombres respectifs de catégories, et où le « -1 » évite d'obtenir une matrice singulière dans le calcul des poids calés « sur les marges »; 5) l'extension de 4) à plus de deux classifications déployées « côte à côte ». Les cas 4 et 5 sont particulièrement importants en production dans les organismes statistiques nationaux.

Dans l'approche du calage, il importe au plus haut point de spécifier exactement l'*information auxiliaire*. Dans les

conditions de base, nous devons distinguer deux cas en ce qui concerne $\mathbf{x}_k$:

i) $\mathbf{x}_k$ est une valeur du vecteur auxiliaire connue pour chaque $k \in U$ (information auxiliaire complète);

ii) $\sum_U \mathbf{x}_k$ est un total connu (importé) et $\mathbf{x}_k$ est connu (observé) pour chaque $k \in s$.

C'est souvent le contexte de l'enquête qui dicte si i) ou ii) s'applique. Le cas i), celui de l'information auxiliaire complète, se présente quand $\mathbf{x}_k$ est spécifié dans la base de sondage pour chaque $k \in U$ (et est donc connu pour chaque $k \in s$). Ce contexte est typique des enquêtes auprès des particuliers et des ménages menées en Scandinavie et dans d'autres pays nord-européens dotés de registres administratifs de haute qualité qui, appariés à la base de sondage, donnent un grand nombre de variables auxiliaires possibles. Le total de population $\sum_U \mathbf{x}_k$ s'obtient simplement par sommation des $\mathbf{x}_k$.

Le cas i) offre une grande latitude pour la structuration du vecteur auxiliaire $\mathbf{x}_k$. Ainsi, si x_k est une valeur de variable continue spécifiée pour chaque $k \in U$, nous sommes appelés à considérer l'inclusion de x_k^2 et d'autres fonctions de x_k dans $\mathbf{x}_k$, parce que des totaux tels que $\sum_U x_k^2$ et $\sum_U \log x_k$ se calculent facilement. Si la relation avec la variable étudiée y est curviligne, ce serait une grave omission de ne pas tenir compte de totaux connus tels que la forme quadratique ou la forme logarithmique.

Le cas ii) est celui qui s'applique dans les enquêtes où la condition i) n'est pas satisfaite, mais où $\sum_U \mathbf{x}_k$ est importé d'une source externe jugée suffisamment fiable et où la valeur individuelle $\mathbf{x}_k$ est disponible (observée pendant la collecte des données) pour chaque $k \in s$. Alors, $\sum_U \mathbf{x}_k$ est parfois appelé « total de contrôle indépendant », pour préciser que sa source est externe à l'enquête. Il est moins souple que le cas i) : si x_k est une variable pour laquelle un total $\sum_U x_k$ est importé d'une source fiable, $\sum_U x_k^2$ peut ne pas être disponible, ce qui empêchera d'inclure x_k^2 dans $\mathbf{x}_k$.

3. Estimation par la régression généralisée dans les conditions de base

3.1 Concept d'estimation par la regression généralisée

Avant de parler du calage, nous considérerons l'*estimation par la régression généralisée* (GREG) (ou simplement l'*estimation par la régression*) pour deux bonnes raisons : 1) l'estimation GREG peut aussi être considérée comme un moyen systématique de tenir compte de l'information auxiliaire et 2) certains estimateurs GREG (mais pas tous) sont des estimateurs par calage, en ce sens qu'ils peuvent être exprimés en fonction d'une pondération linéaire calée.

Les estimateurs GREG et les estimateurs par calage ont été étudiés abondamment au cours des deux dernières décennies. Rien que la terminologie, « estimation GREG » et « estimation par calage », reflète des processus de réflexion différents. Les statisticiens spécialisés dans le domaine appartiennent à deux écoles de pensée, celle de la « régression généralisée GREG » et celle du « calage ». La distinction n'est peut-être pas aussi nette, mais nous l'utiliserons ici car elle aide à structurer l'exposé. Nous n'irons pas jusqu'à soutenir que la seconde école compte plus d'adeptes parmi les organismes statistiques nationaux et la première, dans les milieux universitaires, mais il se pourrait fort bien qu'une telle tendance existe.

Le concept de l'estimateur GREG a évolué progressivement depuis le milieu des années 1970. Särndal, Swensson et Wretman (1992) expliquent l'estimation GREG simple (linéaire), tandis que Fuller (2002) présente une revue détaillée de l'estimation par la régression. L'idée centrale est que les valeurs prévues de y, $\hat{y}_k$, peuvent être produites pour chacun des N éléments de la population par ajustement d'un *modèle auxiliaire* et utilisation des valeurs du vecteur auxiliaire $\mathbf{x}_k$ connues pour tous les $k \in U$. Ces valeurs prévues servent à élaborer un estimateur presque sans biais sous le plan du total de population $Y = \sum_U y_k$ de la forme

$$\begin{aligned} \hat{Y}_{\text{GREG}} &= \sum_U \hat{y}_k + \sum_s d_k (y_k - \hat{y}_k) \\ &= \sum_s d_k y_k + \left(\sum_U \hat{y}_k - \sum_s d_k \hat{y}_k\right). \end{aligned} \qquad (3.1)$$

Le motif évident de cette construction est la perspective d'une estimation très précise $\hat{Y}_{\text{GREG}}$ grâce à un ajustement étroit du modèle auxiliaire qui aboutit à de petits résidus $y_k - \hat{y}_k$. Cette modélisation est la pierre angulaire de l'approche GREG. Certains auteurs donnent à la forme (3.1) le nom (également justifiable) d'*estimateur par différence généralisé*.

La grande variété de modèles auxiliaires possibles engendre une vaste famille d'estimateurs GREG de la forme (3.1). Le modèle auxiliaire, qui traduit la relation imaginée entre $\mathbf{x}$ et y, peut prendre de nombreuses formes : linéaire, non linéaire, linéaire généralisée, mixte (modèle contenant des termes d'effets fixes et d'effets aléatoires) et ainsi de suite. Quel que soit le choix, le modèle « ne fait qu'assister »; bien qu'il ne puisse être qualifié absolument de « vrai », l'estimateur (3.1) est presque sans biais sous le plan dans le cas de contraintes faibles sur le modèle auxiliaire et sur le plan d'échantillonnage, de sorte que $(\hat{Y}_{\text{GREG}} - Y)/N = O_p(n^{-1/2})$ et $(\hat{Y}_{\text{GREG}} - Y)/N = (\hat{Y}_{\text{GREG,lin}} - Y)/N + O_p(n^{-1})$, où la statistique $\hat{Y}_{\text{GREG,lin}}$, qui est le résultat de la linéarisation de $\hat{Y}_{\text{GREG}}$, est sans biais pour Y.

3.2 Estimateur GREG linéaire

Par estimateur GREG linéaire, nous entendons un estimateur produit par un modèle auxiliaire linéaire à effets fixes. Les prédictions sont $\hat{y}_k = \mathbf{x}_k' \mathbf{B}_{s;dq}$ avec

$$\mathbf{B}_{s;dq} = \left(\sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\sum_s d_k q_k \mathbf{x}_k y_k\right)$$

de sorte que (3.1) devient

$$\hat{Y}_{\text{GREG}} = \left(\sum_U \mathbf{x}_k\right)' \mathbf{B}_{s;dq} + \sum_s d_k (y_k - \mathbf{x}_k' \mathbf{B}_{s;dq}). \qquad (3.2)$$

Les q_k sont des facteurs d'échelle, choisis par le statisticien. Le choix type est $q_k = 1$ pour tout k. Le choix de q_k a une certaine incidence (mais souvent limitée) sur l'exactitude de $\hat{Y}_{\text{GREG}}$; la quasi-absence de biais tient pour toute spécification de q_k. (sauf choix outranciers). Bien que le modèle soit simple, l'expression GREG linéaire (3.2) contient de nombreux estimateurs, étant donné les nombreuses options possibles pour le vecteur auxiliaire $\mathbf{x}_k$ et les facteurs d'échelle q_k. Dans des conditions générales,

$$(\hat{Y}_{\text{GREG}} - Y)/N = \left(\sum_s d_k E_k - \sum_U E_k\right)\Big/ N + O_p(n^{-1})$$

où $\sum_s d_k E_k$ est l'estimateur de Horvitz-Thompson dans les résidus $E_k = y_k - \mathbf{x}_k' \mathbf{B}_{U;q}$ avec $\mathbf{B}_{U;q} = (\sum_U q_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_U q_k \mathbf{x}_k y_k)$. D'où, les propriétés fondées sur le plan $E(\hat{Y}_{\text{GREG}}) \approx Y$ et $\text{Var}(\hat{Y}_{\text{GREG}}) \approx \text{Var}(\sum_s d_k E_k)$. Une régression *linéaire* étroitement ajustée de y sur $\mathbf{x}$ est la clé d'une variance faible pour $\hat{Y}_{\text{GREG}}$ (ce qui est loin d'affirmer qu'« une régression linéaire est la régression vraie »).

L'estimateur GREG linéaire de Särndal, Swensson et Wretman (1992) était motivé par le modèle auxiliaire linéaire ξ énonçant que $E_\xi(y_k) = \boldsymbol{\beta}' \mathbf{x}_k$ et $V_\xi(y_k) = \sigma_k^2$. L'ajustement par les moindres carrés généralisés donne l'estimateur (3.2) avec $q_k = 1/\sigma_k^2$. Dans ce contexte, une estimation éclairée quant à la variation des résidus $y_k - \boldsymbol{\beta}' \mathbf{x}_k$ détermine les q_k. Si le vecteur $\mathbf{x}_k$ est fixe, l'effort de modélisation se résume à une opinion quant à la forme des résidus. Le choix $\sigma_k^2 = \sigma^2 x_k$ donne l'estimateur par le ratio classique. Si $q_k = \boldsymbol{\mu}' \mathbf{x}_k$ pour tout $k \in U$ et un vecteur constant $\boldsymbol{\mu}$, alors (3.2) se réduit à la « forme esthétique » $(\sum_U \mathbf{x}_k)' \mathbf{B}_{s;dq}$.

Comme l'ont fait remarquer Beaumont et Alavi (2004) et d'autres, l'estimateur GREG linéaire est robuste au biais (presque sans biais, bien que le modèle auxiliaire ne soit pas tout à fait « correct »), mais peut être considérablement moins efficace (avoir une plus grande erreur quadratique moyenne) que les estimateurs de rechange sensibles au modèle qui, quoique biaisés, ont parfois une variance beaucoup plus faible. Donc, on pourrait affirmer que l'estimateur GREG linéaire n'est pas robuste à la variance; néanmoins, il s'agit d'un concept fondamental dans la théorie de l'estimation fondée sur le plan de sondage.

La spécification de $\mathbf{x}_k$ devrait inclure les variables (dont les totaux de population sont connus) qui ont déjà servi à définir le plan de sondage. Les données issues de l'élaboration du plan de sondage ne devraient pas être abandonnées à l'étape de l'estimation; au contraire, une « utilisation répétée » est recommandée. Par exemple, dans le cas de l'échantillonnage aléatoire simple stratifié (EASS), le vecteur $\mathbf{x}_k$ dans l'estimateur (3.2) devrait inclure, en même temps que les autres variables disponibles, la variable muette servant d'identificateur de strate, $\boldsymbol{\gamma}_k = (\gamma_{k1}, \gamma_{k2}, ..., \gamma_{kh}, ..., \gamma_{kH})'$, où $\gamma_{kh} = 1$ si l'élément k appartient à la strate h et $\gamma_{kh} = 0$ sinon; $h = 1, ..., H$.

Nous pouvons écrire l'estimateur GREG linéaire (3.2) sous la forme d'une somme pondérée par les poids de sondage, $\hat{Y}_{\text{GREG}} = \sum_s w_k y_k$, avec

$$w_k = d_k g_k;\ g_k = 1 + q_k \boldsymbol{\lambda}' \mathbf{x}_k;$$
$$\boldsymbol{\lambda}' = \left(\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k\right)' \left(\sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1}. \qquad (3.3)$$

Les poids w_k sont *calés* (convergents) sur le total $\mathbf{x}$ de population connu : $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. Que $\hat{Y}_{\text{GREG}}$ puisse être exprimé sous la forme d'une somme linéairement pondérée à l'aide de poids calés est un sous produit fortuit. Cette propriété ne fait pas partie du raisonnement GREG, dont l'idée centrale, formulée dans (3.1), est l'ajustement d'un modèle auxiliaire. Outre l'estimateur linéaire simple, quelques autres estimateurs GREG ont la propriété de calage, comme nous le mentionnerons plus loin.

3.3 Estimateur GREG non linéaire

Deux caractéristiques de l'estimateur GREG linéaire (3.2) en font une option de choix pour la production courante des organismes statistiques : i) le total de population auxiliaire $\sum_U \mathbf{x}_k$ est exclu, de sorte que l'estimation peut avoir lieu à condition qu'une valeur exacte de ce total puisse être calculée ou importée, et ii) lorsqu'il est écrit sous forme de la somme pondérée linéairement $\hat{Y}_{\text{GREG}} = \sum_s w_k y_k$, le système de poids (3.3) est indépendant de la variable y et peut donc être appliquée à toutes les variables y de l'enquête. Il n'est pas nécessaire de connaître $\mathbf{x}_k$ individuellement pour tout $k \in U$; connaître $\sum_U \mathbf{x}_k$ suffit. Naturellement, si nous connaissons toutes les valeurs de $\mathbf{x}_k$, nous pouvons rechercher des membres plus efficaces (mais toujours presque sans biais sous le plan) de la famille d'estimateurs GREG (3.1), ce qui permettra également d'écarter une autre critique de l'estimateur GREG linéaire, à savoir qu'un modèle linéaire n'est pas raisonnable pour certains types de données. Par exemple, pour une variable y dichotomique, un modèle auxiliaire

logistique pourrait à la fois être plus réaliste et aboutir à un estimateur GREG plus précis.

Par estimateur GREG non linéaire, il est entendu ici qu'un estimateur est produit comme en (3.1) avec l'aide d'un modèle d'un autre type que « linéaire en $\mathbf{x}_k$ avec effets fixes ». Firth et Bennett (1998), ainsi que Lehtonen et Veijanen (1998) ont été parmi les premiers à étendre le concept GREG dans cette direction; à cet égard, voir aussi Chambers, Dorfman et Wehrly (1993). Ces dernières années, plusieurs auteurs ont étudié les estimateurs GREG non linéaires assistés par modèle.

La notion d'estimation GREG non linéaire est souple; une gamme d'estimateurs deviennent possibles par la voie de modèles auxiliaires ξ du type suivant :

$$E_\xi(y_k|\mathbf{x}_k) = \mu_k \quad \text{pour} \quad k \in U \tag{3.4}$$

où la moyenne sous le modèle μ_k et la variance sous le modèle $V_\xi(y_k|\mathbf{x}_k)$ se voient attribuer chacune une formule appropriée.

Le modèle (3.4) s'applique notamment quand $\mu_k = \mu_k(\mathbf{x}_k, \boldsymbol{\theta})$ est une fonction non linéaire en $\mathbf{x}_k$ spécifiée. Si l'on estime $\boldsymbol{\theta}$ par $\hat{\boldsymbol{\theta}}$, les valeurs ajustées nécessaires pour $\hat{Y}_{\text{GREG}}$ dans (3.1) sont $\hat{y}_k = \mu(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$ pour $k \in U$. Par exemple, si le modélisateur spécifie $\log \mu_k = \alpha + \beta x_k$, les prédictions à utiliser dans (3.1) sont, après estimation des paramètres $\hat{y}_k = \exp(\hat{\alpha} + \hat{\beta} x_k)$.

Les autres applications de (3.4) incluent les modèles linéaires généralisés, tels que $g(\mu_k) = \mathbf{x}_k'\boldsymbol{\theta}$ pour une fonction lien spécifiée $g(\cdot)$ et qu'une structure appropriée est donnée à $V_\xi(y_k|\mathbf{x}_k) = v(\mu_k)$. Nous estimons $\boldsymbol{\theta}$ par $\hat{\boldsymbol{\theta}}$; les valeurs ajustées nécessaires pour l'estimateur GREG non linéaire (3.1) sont $\hat{y}_k = \hat{\mu}_k = g^{-1}(\mathbf{x}_k'\hat{\boldsymbol{\theta}})$. Par exemple, si l'on utilise un modèle auxiliaire logistique, $\mathbf{x}_k'\boldsymbol{\theta} = \text{logit}(\mu_k) = \log(\mu_k/(1-\mu_k))$, et $\hat{y}_k = \hat{\mu}_k = \exp(\mathbf{x}_k'\hat{\boldsymbol{\theta}})/(1+\exp(\mathbf{x}_k'\hat{\boldsymbol{\theta}}))$.

Lehtonen et Veijanen (1998) considèrent le cas d'une variable étudiée catégorique comportant I catégories, $i = 1, 2, ..., I$, $y_{ik} = 1$ si l'élément k appartient à la catégorie i, et $y_{ik} = 0$ sinon. Par exemple, dans une enquête sur la population active comportant $I = 3$ catégories, à savoir « occupé », « chômeur » et « inactif », l'objectif est d'estimer les chiffres de population respectifs $Y_i = \sum_U y_{ik}$, $i = 1, 2, 3$. Ces auteurs utilisent le modèle auxiliaire logistique

$$E_\xi(y_{ik}|\mathbf{x}_k) = \mu_{ik};\ \mu_{ik} = \exp(\mathbf{x}_k'\boldsymbol{\theta}_i)\Big/\left(1+\sum_{i=2}^{I}\exp(\mathbf{x}_k'\boldsymbol{\theta}_i)\right). \tag{3.5}$$

Ils obtiennent les estimations $\hat{\boldsymbol{\theta}}_i$ de $\boldsymbol{\theta}_i$ en maximisant la log-vraisemblance pondérée par les poids de sondage. Les prédictions résultantes $\hat{y}_{ik} = \hat{\mu}_{ik}$ sont utilisées pour former $\hat{Y}_{i\,\text{GREG}} = \sum_U \hat{y}_{ik} + \sum_s d_k(y_{ik} - \hat{y}_{ik})$, pour $i = 1, 2, ..., I$.

Un autre développement est l'application de l'approche GREG à l'estimation pour des domaines, comme dans Lehtonen, Särndal et Veijanen (2003, 2005), ainsi que Myrskylä (2007). Dans les deux premiers de ces articles, les auteurs utilisent des modèles mixtes pour assister l'estimateur GREG non linéaire. Soit U_a un domaine, $U_a \subset U$, dont nous souhaitons estimer le total $Y_{ia} = \sum_{U_a} y_{ik}$ pour $i = 1, 2, ..., I$. Dans l'article de 2005, les prédictions pour l'estimateur GREG non linéaire sont calculées d'après le modèle mixte logistique énonçant que, pour $k \in U_a$,

$$E_\xi(y_{ik}|\mathbf{x}_k; \mathbf{u}_{ia}) = \exp(\mathbf{x}_k'\boldsymbol{\theta}_{ia})\Big/\left(1+\sum_{i=2}^{I}\exp(\mathbf{x}_k'\boldsymbol{\theta}_{ia})\right) \tag{3.6}$$

avec $\boldsymbol{\theta}_{ia} = \boldsymbol{\beta}_i + \mathbf{u}_{ia}$, où $\mathbf{u}_{ia}$ est un vecteur d'écarts aléatoires particuliers au domaine par rapport au vecteur d'effets fixes $\boldsymbol{\beta}_i$.

Les estimateurs GREG non linéaires assistés par modèle, tels que (3.5) et (3.6), nécessitent l'ajustement individuel du modèle pour chaque variable y, car il n'existe aucun système de pondérations applicable uniformément. Cependant, la question se pose de savoir s'il existe des exemples d'estimateurs GREG non linéaires tels que les avantages pratiques de l'estimateur GREG linéaire soient préservés, autrement dit une forme linéairement pondérée à l'aide de poids calés indépendants de la variable y. La réponse est affirmative. À cet égard, deux orientations intéressantes se dégagent de la littérature récente.

Breidt et Opsomer (2000), et Montanari et Ranalli (2005) considèrent des estimateurs GREG assistés par modèle polynomial local dans le cas d'une variable auxiliaire continue unique dont les valeurs x_k sont connues pour tout $k \in U$. La méthode requiert plusieurs choix, dont 1) l'ordre q de l'expression polynomiale locale, 2) la spécification de la fonction noyau et 3) la largeur de fenêtre. L'estimateur résultant peut être exprimé en fonction des poids calés sur les totaux de population des puissances de x_k, de sorte que $\sum_s w_k x_k^j = \sum_U x_k^j$ pour $j = 0, 1, ..., q$.

Breidt, Claeskens et Opsomer (2005) élaborent un estimateur GREG à fonction spline pénalisée pour une variable x unique; le modèle auxiliaire est $m(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 x + ... + \beta_q x^q + \sum_{j=1}^{K}\beta_{q+j}(x-\kappa_j)_+^q$, où $(t)_+^q = t^q$ si $t > 0$ et 0 autrement, q est le degré de la spline et les κ_k sont des nœuds espacés de manière appropriée, par exemple, les quantiles d'échantillon, uniformément espacés, des valeurs x_k. Après avoir estimé les paramètres β, ils obtiennent les prédictions $\hat{y}_k = m(x_k; \hat{\boldsymbol{\beta}})$ nécessaires pour la formule GREG générale (3.1). Les auteurs soulignent que l'estimateur GREG résultant est calé pour la partie paramétrique du modèle, c'est-à-dire $\sum_s w_k x_k^j = \sum_U x_k^j$ pour $j = 0, 1, ..., q$, ainsi que pour les termes polynomiaux tronqués dans le modèle, à condition qu'ils ne soient pas pénalisés.

Nous pouvons résumer l'estimation GREG comme il suit. L'estimateur GREG linéaire offre des avantages pratiques pour la production de statistiques à grande échelle. Il peut être exprimé sous la forme d'une somme pondérée linéairement de valeurs y_k au moyen de poids calés sur $\sum_U \mathbf{x}_k$; les poids sont indépendants des valeurs y_k et peuvent être appliqués à toutes les variables y de l'enquête. Il suffit de connaître un total de population auxiliaire $\sum_U \mathbf{x}_k$, importé d'une source fiable. L'estimateur GREG non linéaire peut donner lieu à une variance considérablement réduite, grâce aux modèles plus perfectionnés qui peuvent être envisagés lorsqu'on dispose d'information auxiliaire complète ($\mathbf{x}_k$ connu pour tout $k \in U$). La quasi-absence de biais sous le plan est préservée. Certains estimateurs GREG non linéaires peuvent s'écrire sous forme de sommes linéairement pondérées.

Dans les exercices théoriques portant sur des populations et des relations créées artificiellement, il est possible de provoquer des situations où un estimateur GREG non linéaire présente un grand avantage par rapport à l'estimateur GREG linéaire en ce qui concerne la variance. Les expériences de ce type sont importantes pour l'illustration. Cependant, pour répondre aux exigences quotidiennes de production des organismes statistiques nationaux, les estimateurs GREG non linéaires « extravagants » ne semblent présenter qu'un intérêt assez lointain à l'heure actuelle. Les modèles auxiliaires spécifiés pour l'estimation GREG doivent satisfaire aux exigences de robustesse et de faisabilité. L'attrait d'une petite réduction de la variance d'échantillonnage est balayé par les préoccupations au sujet d'autres erreurs (non due à l'échantillonnage) et aux difficultés rencontrées dans le processus de production quotidienne.

Le passage des estimateurs GREG linéaires aux estimateurs GREG non linéaires offre des possibilités et soulève des questions. Quelle est la formule la plus appropriée de l'espérance du modèle μ_k? À quel point les résultats sont-ils sensibles à la spécification de la partie variance du modèle auxiliaire? Dans quelle mesure la rapidité des calculs est-elle un problème? Des travaux de recherche plus approfondis permettront de mieux répondre à ces questions.

4. Méthode de l'estimation par calage

4.1 Calage dans les conditions de base

Dans la méthode d'estimation GREG examinée à la section précédente, une étape essentielle consiste à produire les valeurs prévues $\hat{y}_k$ par ajustement d'un modèle auxiliaire. Par contre, telle qu'elle est définie à la section 1.1, la méthode de calage ne fait directement référence à aucun modèle. Elle met plutôt l'accent sur les données sur lesquelles peut se faire le calage. Un élément clé de l'« approche du calage » est la pondération linéaire des valeurs y observées, avec alignement des poids sur des agrégats calculables. Cette différence d'ordre conceptuel donne parfois des estimateurs différents pour les approches GREG et de calage.

La méthode de calage offre un haut degré de généralisation. Elle peut s'appliquer dans des conditions diverses, dont les plans d'échantillonnage complexes, la correction de la non-réponse et les erreurs dans les bases de sondage. Néanmoins, à la présente section, nous nous attachons aux conditions de base énoncées à la section 2, c'est-à-dire celles d'échantillonnage à une phase et de réponse complète. Nous reprenons également la notation utilisée dans cette section. Les données dont nous disposons pour estimer le total de population $Y = \sum_U y_k$ sont i) les valeurs de la variable étudiée y_k observées pour $k \in s$, ii) les poids de sondage connus $d_k = 1/\pi_k$ pour $k \in U$ et iii) les valeurs du secteur auxiliaire connues $\mathbf{x}_k$ pour $k \in U$ (ou un total importé $\sum_U \mathbf{x}_k$). Ces conditions simples sont celles énoncées dans les articles publiés par Deville et Särndal (1992), ainsi que par Deville, Särndal et Sautory (1993), auxquels la méthode doit son nom et qui ont inspiré les travaux qui ont suivi. Bien que le contexte soit simple, le calage soulève plusieurs questions, dont certaines ayant trait aux calculs, que nous passons en revue à la section 5.

Notre objectif, dans les sections 4.2 et 4.3, est de déterminer les poids w_k qui satisfont l'équation de calage $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$, puis de les utiliser pour obtenir l'estimateur par calage de Y de la forme $\hat{Y}_{\mathrm{CAL}} = \sum_s w_k y_k$, que nous pouvons comparer à l'estimateur de Horvitz-Thompson sans biais en écrivant $\hat{Y}_{\mathrm{CAL}} = \hat{Y}_{\mathrm{HT}} + \sum_s (w_k - d_k) y_k$. Il s'ensuit que le biais de $\hat{Y}_{\mathrm{CAL}}$ est $E(\hat{Y}_{\mathrm{CAL}}) - Y = E(\sum_s (w_k - d_k) y_k)$. Réaliser l'objectif de quasi-absence de biais sous le plan de sondage exige que $E(\sum_s (w_k - d_k) y_k) \approx 0$, quelle que soit la variable y. Naturellement, le calage devrait viser à produire de petits écarts $w_k - d_k$.

Nous pouvons atteindre l'objectif de « calage pour obtenir la convergence sur des totaux de population auxiliaires connus » de nombreuses façons. Nous pouvons créer de nombreux ensembles de poids calés sur le total connu $\sum_U \mathbf{x}_k$. À la présente section, nous examinerons cette prolifération des méthodes sous deux angles relevés dans la littérature, à savoir la *méthode de la distance minimale* et la *méthode du vecteur instrumental*. Demnati et Rao (2004) proposent encore un autre moyen de construire une variété de poids calés.

4.2 Méthode de la distance minimale

Dans cette méthode, le but du calage est de modifier les poids initiaux $d_k = 1/\pi_k$ afin d'obtenir de nouveaux poids

w_k, jugés comme étant « proches » des poids d_k. Pour cela, considérons la fonction de distance $G_k(w, d)$ définie pour chaque $w > 0$ de façon que $G_k(w, d) \geq 0$, $G_k(d, d) = 0$, dérivable par rapport à w, strictement convexe, à dérivée continue $g_k(w, d) = \partial G_k(w, d)/\partial w$ telle que $g_k(d, d) = 0$. Habituellement, on choisit la fonction de distance de manière que $g_k(w, d) = g(w/d)/q_k$, où les q_k sont des facteurs d'échelle positifs convenablement choisis, $g(\cdot)$ est une fonction d'un argument unique, continue, strictement croissante, avec $g(1) = 0, g'(1) = 0$. Soit $F(u) = g^{-1}(u)$ la fonction inverse de $g(\cdot)$. Minimiser la distance totale $\sum_s G_k(w_k, d_k)$ sous la contrainte de l'équation de calage $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$ mène à $w_k = d_k F(q_k \mathbf{x}_k' \boldsymbol{\lambda})$, où $\boldsymbol{\lambda}$ est la solution (à supposer qu'elle existe) de

$$\sum_s d_k \mathbf{x}_k F(q_k \mathbf{x}_k' \boldsymbol{\lambda}) = \sum_U \mathbf{x}_k. \qquad (4.1)$$

Les poids ont une propriété d'optimalité, parce qu'une fonction objective dûment spécifiée est minimisée, mais il s'agit d'une « optimalité faible » en ce sens que les spécifications possibles de la fonction de distance et des facteurs d'échelle q_k sont nombreuses.

La fonction de distance $G_k(w_k, d_k) = (w_k - d_k)^2/2d_k q_k$ a suscité beaucoup d'intérêt. Elle donne $g_k(w_k, d_k) = (w_k/d_k - 1)/q_k$; $g(w/d) = w/d - 1$; $F(u) = g^{-1}(u) = 1 + u$. L'expression « cas linéaire » est donc appropriée. La tâche consiste alors à minimiser la « distance du chi-carré » $\sum_s (w_k - d_k)^2/2d_k q_k$, sachant $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. L'équation (4.1) se lit $\sum_s d_k \mathbf{x}_k (1 + q_k \mathbf{x}_k' \boldsymbol{\lambda}) = \sum_U \mathbf{x}_k$, qui se résout facilement pour obtenir $\boldsymbol{\lambda}$. L'estimateur résultant de $Y = \sum_U y_k$ est $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$ avec les poids $w_k = d_k g_k$ donnés par (3.3). Autrement dit, $\hat{Y}_{\text{CAL}} = \hat{Y}_{\text{GREG}}$ tel qu'il est donné par (3.2) et les résidus qui déterminent la variance asymptotique sont $E_k = y_k - \mathbf{x}_k' \mathbf{B}_{U;q}$ conformément à la section 3.2. Il arrive d'obtenir certains poids négatifs w_k.

L'estimateur GREG linéaire implique l'utilisation de poids calés (sur $\sum_U \mathbf{x}_k$) et le revers de cette médaille est que le cas linéaire du calage (avec la distance du chi-carré) donne l'estimateur GREG linéaire. La tendance, dans certaines publications et applications, à entremêler l'approche GREG et l'approche du calage émane de ce fait. Bon nombre d'applications fructueuses de l'utilisation de l'information auxiliaire découlent, en tout cas, de cette linéarité bilatérale. L'Enquête sur la population active menée au Canada en est un exemple et un fait récent intéressant en ce qui la concerne est le recours à des estimateurs composites où une partie de l'information provient des résultats d'enquête des mois antérieurs, comme le décrivent Fuller et Rao (2001).

L'équation de calage est satisfaite pour tout choix des facteurs d'échelle positifs q_k dans (3.2). Un choix simple est $q_k = 1$ pour tout k, mais ce n'est pas toujours celui qui est privilégié. Par exemple, s'il existe une variable auxiliaire unique, systématiquement positive et que $\mathbf{x}_k = x_k$, nombreux sont ceux qui s'attendront intuitivement à ce que $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$ mène à l'estimateur par le ratio habituel $\sum_U x_k (\sum_s d_k y_k)/(\sum_s d_k x_k)$, ce qu'il fait, mais si l'on prend $q_k = x_k^{-1}$ et non $q_k = 1$.

Une autre fonction de distance d'un intérêt considérable est $G_k(w_k, d_k) = \{w_k \log(w_k/d_k) - w_k + d_k\}/q_k$. Elle mène à $F(u) = g^{-1}(u) = \exp(u)$, qui est le « cas exponentiel ». Alors, (4.1) se lit $\sum_s d_k \mathbf{x}_k \exp(q_k \mathbf{x}_k' \boldsymbol{\lambda}) = \sum_U \mathbf{x}_k$. Des méthodes de résolution numérique doivent être appliquées pour trouver $\boldsymbol{\lambda}$, afin d'obtenir les poids $w_k = d_k \exp(q_k \mathbf{x}_k' \boldsymbol{\lambda})$. Aucun w_k négatif ne sera obtenu.

Deville et Särndal (1992) montrent qu'une gamme de fonctions de distance satisfaisant des conditions faibles produisent des estimateurs par calage asymptotiquement équivalents. Diverses fonctions de distance sont comparées dans Deville, Särndal et Sautory (1993), Singh et Mohl (1996), ainsi que Stukel, Hidiroglou et Särndal (1996). Certaines de ces fonctions garantissent que les poids se situent entre des bornes spécifiées, de façon à exclure toute valeur trop grande ou trop faible (négative). Les changements apportés à la fonction de distance n'ont souvent qu'un effet mineur sur la variance de l'estimateur par calage $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$, même si la taille d'échantillon est assez petite. Les questions relatives à l'existence d'une solution à l'équation de calage sont abordées dans Théberge (2000).

4.3 Méthode du vecteur instrumental

La méthode du vecteur instrumental est une alternative à la minimisation de la distance. Elle est considérée dans Deville (1998), Estevao et Särndal (2000, 2006) et Kott (2006). Elle permet également de produire de nombreux ensembles de poids tous calés sur la même information.

Considérons des poids de la forme $w_k = d_k F(\boldsymbol{\lambda}' \mathbf{z}_k)$, où $\mathbf{z}_k$ est un vecteur dont les valeurs sont définies pour $k \in s$ et ayant la même dimension que le vecteur auxiliaire spécifié $\mathbf{x}_k$, et où le vecteur $\boldsymbol{\lambda}$ est déterminé d'après l'équation de calage $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. La fonction $F(\cdot)$ joue le même rôle que pour la méthode de minimisation de la distance; plusieurs choix de $F(\cdot)$ sont intéressants, par exemple, $F(u) = 1 + u$ et $F(u) = \exp(u)$.

Si nous optons pour la fonction linéaire $F(u) = 1 + u$, nous obtenons $w_k = d_k(1 + \boldsymbol{\lambda}' \mathbf{z}_k)$. Il est facile de déterminer $\boldsymbol{\lambda}$ de façon à satisfaire l'équation de calage $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. L'estimateur par calage résultant est

$$\hat{Y}_{\text{CAL}} = \sum_s w_k y_k;\ w_k = \sum_s d_k (1 + \boldsymbol{\lambda}' \mathbf{z}_k) y_k;$$
$$\boldsymbol{\lambda}' = \left(\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k\right)' \left(\sum_s d_k \mathbf{z}_k \mathbf{x}_k'\right)^{-1}. \qquad (4.2)$$

Quel que soit le choix de $\mathbf{z}_k$, les poids $w_k = d_k(1+\boldsymbol{\lambda}'\mathbf{z}_k)$ satisfont l'équation de calage. Le choix typique est $\mathbf{z}_k = \mathbf{x}_k$. En particulier, fixer $\mathbf{z}_k = q_k\mathbf{x}_k$, pour les valeurs q_k spécifiées, donne les poids (3.3).

Même les « choix délibérément maladroits » pour $\mathbf{z}_k$ donnent des résultats étonnamment bons. Par exemple, posons que x_k est une variable auxiliaire continue unique et que $\mathbf{z}_k = c_k x_k^{p-1}$. Supposons que $p=3$ et $c_k = 1$ pour quatre éléments seulement, choisis au hasard parmi les $n=100$ éléments d'un échantillon réalisé s et que $c_k = 0$ pour les 96 autres. La quasi-absence de biais de $\hat{Y}_{\mathrm{CAL}} = \sum_s d_k(1+\boldsymbol{\lambda}'\mathbf{z}_k)y_k$ existe encore. Même dans le cas d'un vecteur $\mathbf{z}$ aussi parcimonieux, l'accroissement de la variance comparativement à de meilleurs choix de $\mathbf{z}_k$ n'est pas nécessairement excessif.

Si le plan d'échantillonnage et le vecteur $\mathbf{x}$ sont tous deux fixes, Estevao et Särndal (2004) et Kott (2004) font remarquer qu'il existe un vecteur $\mathbf{z}$ asymptotiquement optimal donné par

$$\mathbf{z}_k = \mathbf{z}_{0k} = d_k^{-1}\sum_{\ell\in s}(d_k d_\ell - d_{k\ell})\mathbf{x}_\ell$$

où $d_{k\ell}$ est l'inverse de la probabilité d'inclusion de deuxième ordre $\pi_{k\ell} = P(k\ \&\ \ell \in s)$, supposée strictement positive. L'estimateur par calage résultant $\hat{Y}_{\mathrm{CAL}} = \sum_s d_k(1+\boldsymbol{\lambda}'\mathbf{z}_{0k})y_k$ est essentiellement l'« estimateur optimal sous "randomisation" » proposé au départ par Montanari (1987) et discuté depuis par de nombreux auteurs.

Andersson et Thorburn (2005) considèrent la question sous l'angle opposé et se demandent si, dans la méthode de la distance minimale, il est possible de spécifier une fonction de distance telle que sa minimisation produise l'estimateur optimal sous « randomisation ». Ils trouvent cette distance qui - ce qui n'est pas entièrement surprenant - est reliée à la distance du chi-carré (mais n'y est pas identique).

4.4 Le calage nécessite-t-il l'énoncé explicite d'un modèle

La méthode de calage présentée aux sections 4.2 et 4.3 consiste simplement à calculer les poids qui reproduisent les totaux auxiliaires spécifiés. Elle ne requiert aucun modèle auxiliaire explicite, à moins que l'on veuille à tout prix que le choix de certaines variables à inclure dans le vecteur $\mathbf{x}_k$ représente un sérieux effort de modélisation. La justification des poids repose plutôt en majeure partie sur leur convergence avec les totaux de contrôle précisés. Les premiers travaux reflètent cette attitude, d'abord ceux de Deming (1943), puis ceux d'Alexander (1987), Zieschang (1990) et d'autres. D'où la question : N'est-il pas néanmoins important de justifier ce « calage sans modèle » à l'aide d'une formulation explicite de modèle? Il est vrai que les statisticiens ont l'habitude de réfléchir en fonction de modèles et se sentent plus ou moins obligés de toujours assortir une procédure statistique de la formulation d'un modèle. Énoncer la relation connexe entre y et $\mathbf{x}$, même si elle est aussi simple que le modèle linéaire courant, pourrait effectivement avoir une certaine valeur pédagogique pour l'explication du calage.

Mais l'énoncé d'un modèle aidera-t-il les utilisateurs et les praticiens à mieux comprendre la méthode de calage? La plupart d'entre eux la considère parfaitement claire et transparente de toute façon. Ils n'ont besoin d'aucune autre justification que celle de la convergence avec les valeurs de contrôle précisées. La recherche du « vrai modèle ayant la vraie structure de variance » se traduira-t-elle par une précision sensiblement meilleure pour la majeure partie des nombreuses estimations produites dans le cadre d'une grande enquête menée par un organisme public? Peu probablement.

La section suivante traite du calage fondé sur un modèle. Dans cette variante, proposée par Wu et Sitter (2001), la modélisation joue effectivement un rôle explicite et important. Ces auteurs dénomment l'estimateur par calage linéaire $\hat{Y}_{\mathrm{CAL}} = \sum_s w_k y_k$ où les poids w_k sont donnés par (3.3) « une application de routine sans modélisation ». Cette description est pertinente, puisque la seule exigence est d'identifier les variables x à leurs totaux de population connus.

4.5 Calage fondé sur un modèle

L'idée du calage fondé sur un modèle est avancée dans Wu et Sitter (2001) et examinée plus en profondeur dans Wu (2003) et dans Montanari et Ranalli (2003, 2005). Le facteur qui motive cette approche est que l'existence d'information auxiliaire complète permet d'utiliser plus efficacement les valeurs connues de $\mathbf{x}_k$ pour chaque $k \in U$ qui ne l'est possible dans le calage sans modèle, où un total connu $\sum_U \mathbf{x}_k$ suffit. Les poids sont contraints de converger vers le total de population calculable des prédictions $\hat{y}_k$ calculées d'après un modèle formulé convenablement. Donc, le système de poids n'est pas nécessairement convergent avec le total de population connu de chaque variable auxiliaire, à moins qu'une mesure particulière soit prise afin de retenir cette propriété. Le calage fondé sur un modèle satisfait encore les trois éléments a) à c) de la définition du calage proposée à la section 3.1; en particulier, les estimateurs sont presque sans biais sous le plan de sondage.

Considérons un modèle auxiliaire non linéaire du type (3.4). Nous estimons le paramètre inconnu $\boldsymbol{\theta}$ par $\hat{\boldsymbol{\theta}}$, ce qui donne les valeurs ajustées $\hat{y}_k = \hat{\mu}_k = \mu(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$ calculées à l'aide des $\mathbf{x}_k$ connus pour tout $k \in U$. Il s'ensuit que la taille de population N est connue et devrait

jouer un rôle important dans le calage. Si nous utilisons la distance du chi-carré minimale, nous trouvons les poids de l'estimateur par calage sous un modèle $\hat{Y}_{\text{MCAL}} = \sum_s w_k y_k$ en minimisant $\sum_s (w_k - d_k)^2/(2d_k q_k)$ pour la valeur spécifiée de q_k et $d_k = 1/\pi_k$, sous la contrainte des équations de calage

$$\sum_s w_k = N; \sum_s w_k \hat{y}_k = \sum_U \hat{y}_k. \tag{4.3}$$

Pour simplifier, prenons $q_k = 1$ pour tout k; nous calculons les poids calés, réarrangeons les termes et trouvons que l'estimateur par calage sous un modèle peut s'écrire sous la forme

$$\hat{Y}_{\text{MCAL}} = N\{\bar{y}_{s;d} + (\bar{\hat{y}}_U - \bar{\hat{y}}_{s;d})\tilde{B}_{s;d}\} \tag{4.4}$$

où $\bar{y}_{s;d} = \sum_s d_k y_k / \sum_s d_k$; $\bar{\hat{y}}_{s;d} = \sum_s d_k \hat{y}_k / \sum_s d_k$, et

$$\tilde{B}_{s;d} = \left(\sum_s d_k (\hat{y}_k - \bar{\hat{y}}_{s;d}) y_k\right) \Big/ \sum_s d_k (\hat{y}_k - \bar{\hat{y}}_{s;d})^2.$$

La régression impliquée par $\tilde{B}_{s;d}$ est celle des valeurs observées de y sur les valeurs prévues de y. L'idée de cette régression viendrait rarement à l'esprit du modélisateur qui essaye de structurer la relation entre y_k et $\mathbf{x}_k$, mais elle s'avère efficace dans l'élaboration de l'estimateur par calage. Wu et Sitter (2001) démontrent que

$$(\hat{Y}_{\text{MCAL}} - Y)/N = \left(\sum_s d_k \tilde{E}_k - \sum_U \tilde{E}_k\right) \Big/ N + O_p(n^{-1})$$

avec $\tilde{E}_k = y_k - \bar{y}_U - (\mu_k - \bar{\mu}_U)\tilde{B}_U$, où $\tilde{B}_U = (\sum_U (\mu_k - \bar{\mu}_U) y_k) / \sum_U (\mu_k - \bar{\mu}_U)^2$, et $\bar{\mu}_U = \sum_U \mu_k / N$. Le coefficient $\tilde{B}_U$ peut ne pas être proche de l'unité, même dans le cas de grands échantillons. Il exprime la régression de y_k sur sa moyenne sous le modèle auxiliaire $\mu_k = \mu(\mathbf{x}_k, \boldsymbol{\beta})$. Autrement dit, $\hat{Y}_{\text{MCAL}}$ peut être considéré comme un estimateur par la régression qui utilise l'espérance du modèle μ_k comme variable auxiliaire, ce qui laisse à $\tilde{E}_k$ le rôle de résidu déterminant la variance asymptotique de $\hat{Y}_{\text{MCAL}}$.

Comment cette variance asymptotique se compare-t-elle à la formulation GREG non linéaire (3.1) sous le même modèle auxiliaire non linéaire et les mêmes $\hat{y}_k = \hat{\mu}_k$? La formule (3.1) implique l'existence d'une pente égale à l'unité dans la régression de y_k sur $\hat{y}_k = \hat{\mu}_k$. Vu sous cet angle, $\hat{Y}_{\text{GREG}}$ est un estimateur par différence plutôt qu'un estimateur par régression et est donc moins sensible aux structures présentes dans les données. L'estimateur GREG non linéaire $\hat{Y}_{\text{GREG}}$ est généralement moins efficace que $\hat{Y}_{\text{MCAL}}$. (Il est évidemment possible de modifier $\hat{Y}_{\text{GREG}}$ afin de tenir compte également de l'information contenue dans la taille connue de population N.)

Par ailleurs, comparativement à l'estimateur par calage linéaire (sans modèle) $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$ avec les poids tels qu'en (3.3), l'estimateur par calage fondé sur un modèle $\hat{Y}_{\text{MCAL}}$ donné par (4.4) peut avoir un avantage considérable pour ce qui est de la variance, mais entraîne la perte des avantages pratiques que sont la convergence vers le total connu de population $\sum_U \mathbf{x}_k$ et un système de poids polyvalent applicable à toutes les variables y. Dans (4.4) les valeurs de y sont pondérées linéairement, mais maintenant, les poids dépendent aussi des valeurs de y. On peut donc se demander si $\hat{Y}_{\text{MCAL}}$ est un estimateur par calage authentique.

Dans une étude empirique, Wu et Sitter (2001) comparent $\hat{Y}_{\text{MCAL}} = \sum_s w_k y_k$, calé conformément à (4.3), à l'estimateur GREG non linéaire, $\hat{Y}_{\text{GREG}} = \sum_U \hat{y}_k + \sum_s d_k (y_k - \hat{y}_k)$ donné par (3.1), pour le même modèle auxiliaire non linéaire et les mêmes $\hat{y}_k = \hat{\mu}_k$. L'étude confirme que $\hat{Y}_{\text{MCAL}}$ a des meilleures propriétés de variance que l'estimateur non linéaire $\hat{Y}_{\text{GREG}}$. Les auteurs créent une population finie U de taille $N = 2\,000$ avec les valeurs (y_k, x_k), $k = 1, ..., 2\,000$, telles que $\log(y_k) = 1 + x_k + \varepsilon_k$; les 2 000 valeurs de x_k sont des réalisations de la variable aléatoire Gamma (1,1) et ε_k est une erreur normalement distribuée. L'information auxiliaire consiste en la taille de population N et les valeurs connues x_k pour $k = 1, ..., 2\,000$. Ils tirent ensuite des échantillons aléatoires simples répétés de taille n = 100. Pour les deux estimateurs, le modèle auxiliaire est le modèle loglinéaire $E_\xi(y_k | x_k) = \mu_k$ avec $\log(\mu_k) = \alpha + \beta x_k$. Ce modèle est ajusté à chaque échantillon, en utilisant la méthode d'estimation du pseudomaximum de quasi-vraisemblance. Les valeurs ajustées $\hat{y}_k = \exp(\hat{\alpha} + \hat{\beta} x_k)$ sont utilisées pour former $\hat{Y}_{\text{MCAL}}$ ainsi que $\hat{Y}_{\text{GREG}}$. La variance de simulation est nettement plus faible pour $\hat{Y}_{\text{MCAL}}$. (Le modèle GREG linéaire (3.2), identique à l'estimateur par calage sans modèle, est également inclus dans l'étude de Wu et Sitter; fait peu étonnant, il est encore moins efficace que l'estimateur GREG non linéaire sous la forte relation non linéaire imposée dans leur expérience.

Montanari et Ranalli (2005) fournissent d'autres données, pour plusieurs populations créées artificiellement, sur la comparaison entre $\hat{Y}_{\text{MCAL}}$ et l'estimateur non linéaire $\hat{Y}_{\text{GREG}}$. Leur modèle auxiliaire, $y_k = \mu_k + \varepsilon_k$, est ajusté par régression non paramétrique (lissage polynomial local) donnant les prédictions $\hat{y}_k = \hat{\mu}_k$ pour $k \in U$. Dans le cas de ce type d'ajustement du modèle, les prédictions $\hat{y}_k = \hat{\mu}_k$ sont hautement exactes. Naturellement, l'estimateur par calage fondé sur un modèle $\hat{Y}_{\text{MCAL}}$ ne donne lieu qu'à une amélioration marginale comparativement à l'estimateur non linéaire $\hat{Y}_{\text{GREG}}$.

Nous pouvons résumer la méthode du calage de la façon suivante. L'estimateur de $Y = \sum_U y_k$ a la forme linéairement pondérée $\hat{Y} = \sum_s w_k y_k$. Dans le calage linéaire (sans modèle), l'équation de calage s'écrit $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$; un total de population auxiliaire connu $\sum_U \mathbf{x}_k$ est requis, mais non une information auxiliaire complète ($\mathbf{x}_k$

connu pour tout $k \in U$). Les mêmes poids peuvent être appliqués à toutes les variables y (pondération polyvalente); l'estimateur est identique à l'estimateur GREG linéaire (mais dérivé selon un raisonnement différent). Dans le calage fondé sur un modèle, la moyenne sous le modèle auxiliaire μ_k est non linéaire en $\mathbf{x}_k$; l'information auxiliaire complète est habituellement requise; les contraintes de calage comprennent l'équation $\sum_s w_k \hat{y}_k = \sum_U \hat{y}_k$; les poids w_k dépendent des valeurs y_k, ce qui implique la perte de la propriété de polyvalence.

5. Aspects du calcul, poids extrêmes et valeurs aberrantes

Le calcul des poids calés soulève d'importantes questions d'ordre pratique qui sont traitées dans un certain nombre d'articles. Dans la production à grande échelle de statistiques d'un organisme statistique national, tous les calculs doivent se dérouler harmonieusement, de manière routinière. Les valeurs de pondération inappropriées (ou indûment variables) doivent être évitées. De nombreux praticiens soutiennent raisonnablement que tous les poids doivent être positifs (voir même supérieurs à l'unité) et que les valeurs très élevées doivent être évitées.

Quelques-uns des poids calculés selon (4.2) peuvent s'avérer très grands ou négatifs. Huang et Fuller (1978), ainsi que Park et Fuller (2005) ont proposé des méthodes permettant d'éviter ces pondérations indésirables.

Dans la méthode de minimisation de la distance, la fonction de distance peut être formulée de manière à exclure les poids négatifs, tout en satisfaisant les équations de calage données. Le logiciel CALMAR (Deville, Särndal et Sautory 1993) permet d'utiliser plusieurs fonctions de distance de ce type. Une version étendue, CALMAR2, est décrite dans LeGuennec et Sautory (2002). D'autres organismes statistiques ont développé leur propre logiciel pour le calcul des pondérations. Le SGE de Statistique Canada, le CLAN97 de Statistique Suède, le Bascula 4.0 du Bureau central de la statistique des Pays-Bas et le *g*-CALIB-S de Statistique Belgique en sont des exemples. Chacun à leur façon, ces logiciels visent à résoudre les problèmes de calcul qui se posent. Dans chaque cas, l'utilisateur doit consulter le guide de l'utilisateur afin de savoir exactement comment sont traités les problèmes de calcul, y compris la manière d'éviter les poids indésirables.

Dans le SGE, un programme mathématique minimise la distance du chi-carré, conditionnellement aux contraintes de calage et aux bornes individuelles sur les poids, de façon que ceux-ci satisfassent $A_k \le w_k \le B_k$ pour les valeurs spécifiées A_k, B_k. Bascula 4.0 est décrit dans Nieuwenbroek et Boonstra (2002). Le logiciel *g*-CALIB-S, décrit dans Vanderhoeft, Waeytens et Museux (2001), ainsi que Vanderhoeft (2001), s'appuie sur l'inverse généralisée (la Moore-Penrose) pour le calcul des poids, si bien qu'il n'y a pas lieu de s'inquiéter d'une redondance éventuelle dans l'information auxiliaire.

Dans Bankier, Houle et Luc (1997), l'objectif est double, à savoir maintenir les poids calculés entre les bornes souhaitées et laisser tomber certaines variables x afin d'éliminer les dépendances presque linéaires. Isaki, Tsay et Fuller (2004) considèrent la programmation quadratique pour obtenir, pour les ménages ainsi que pour les personnes, des poids qui sont compris entre les bornes spécifiées.

Une intervention touchant les poids (afin d'éliminer les valeurs de pondération indésirables) amène à se demander dans quelle mesure on peut s'écarter des poids de sondage d_k sans compromettre la propriété désirable d'estimation presque sans biais sous le plan. Une idée qui a été mise à l'épreuve consiste à modifier l'ensemble de contraintes de façon que la différence entre l'estimateur pour les variables auxiliaires et pour les totaux de population connus correspondants soit comprise entre des marges de tolérance. Ainsi, Chambers (1996) minimise une fonction de perte biaisée ou fonction de perte ridge.

Des valeurs aberrantes dans les variables auxiliaires peuvent être la cause des poids extrêmes. Le calage en présence de valeurs aberrantes est examiné par Duchesne (1999). Sa méthode de « calage robuste » peut introduire dans les estimations un biais qui pourrait toutefois être plus que compensé par une réduction de la variance.

Si l'ensemble de contraintes est étendu de façon à limiter les poids à des intervalles précis, il n'est pas certain que le problème d'optimisation aura une solution. L'existence de celle-ci est considérée dans Théberge (2000), qui propose aussi des méthodes de traitement des valeurs aberrantes.

6. Estimation par calage de paramètres plus complexes

La méthode de calage peut être adaptée à l'estimation de paramètres plus complexes qu'un total de population. Nous examinons certains exemples à la présente section. Nous continuons de supposer que nous nous trouvons dans des conditions d'échantillonnage à une seule phase et de réponse complète, et nous utilisons la même notation qu'à la section 2. Un exemple est l'estimation des quantiles de population (section 6.1), un autre est l'estimation des fonctions de totaux (section 6.2). D'autres exemples rentrant dans cette catégorie, que nous ne passons pas en revue ici, sont ceux de Théberge (1999), pour l'estimation de paramètres bilinéaires et de Tracy, Singh et Arnab (2003), pour le calage par rapport aux moments de deuxième ordre.

6.1 Calage de l'estimation des quantiles

La médiane et d'autres quantiles de la population finie sont des mesures descriptives importantes, particulièrement dans le cas des enquêtes économiques. Afin d'estimer les quantiles, il faut d'abord estimer la fonction de répartition de la population finie. Avant que l'usage du calage se répande, plusieurs auteurs ont considéré l'estimation des quantiles, avec ou sans l'utilisation d'information auxiliaire. Les auteurs d'articles plus récents se sont tournés vers la méthode de calage dans le même but, dont Kovaĉević (1997), Wu et Sitter (2001), Ren (2002), Tillé (2002), Harms (2003), Harms et Duchesne (2006), et Rueda et coll. (2007). Comme l'illustrent ces articles, il existe plus d'un moyen d'appliquer cette méthode. Le caractère non lisse de la fonction de répartition de la population finie cause certaines complexités dont la résolution varie selon les auteurs.

Soit $\Delta(\cdot)$ la fonction de Heaviside, définie pour tout réel z de manière que $\Delta(z)=1$ si $z \geq 0$ et $\Delta(z)=0$ si $z<0$. La fonction de répartition inconnue de la variable étudiée y est

$$F_y(t)=\frac{1}{N}\sum_U \Delta(t-y_k). \qquad (6.1)$$

Le quantile α de la population finie est défini comme étant $Q_{y\alpha}=\inf\{t|F_y(t)\geq\alpha\}$. La variable auxiliaire x_j, qui prend les valeurs x_{jk}, possède la fonction de répartition $F_{x_j}(t)=(1/N)\sum_U\Delta(t-x_{jk})$ avec le quantile α dénoté $Q_{x_j\alpha}$, $j=1,2,...,J$. Un estimateur naturel de $F_y(t)$ basé sur les poids de sondage $d_k=1/\pi_k$ est

$$\hat{F}_y(t)=\frac{1}{\sum_s d_k}\sum_s d_k\Delta(t-y_k).$$

Un estimateur par calage de $F_y(t)$ prend la forme

$$\hat{F}_{y\mathrm{CAL}}(t)=\frac{1}{\sum_s w_k}\sum_s w_k\Delta(t-y_k) \qquad (6.2)$$

où les poids w_k sont calés comme il convient sur une information auxiliaire spécifiée. Puis, nous tirons de $\hat{F}_{y\mathrm{CAL}}(t)$ l'estimateur du quantile α donné par $\hat{Q}_{y\alpha}=\inf\{t|\hat{F}_{y\mathrm{CAL}}(t)\geq\alpha\}$. Une formule analogue à (6.2) est vérifiée pour $\hat{F}_{x_j\mathrm{CAL}}(t)$.

Sans référence explicite à un modèle, Harms et Duchesne (2006) spécifient l'information disponible pour le calage comme étant une taille de population connue, N, et les quantiles de population connue $Q_{x_j\alpha}$ pour $j=1,2,...,J$. L'information auxiliaire complète, avec les valeurs $\mathbf{x}_k=(x_{k1},...,x_{kJ})'$ connues pour $k\in U$, n'est pas nécessaire. (Toutefois, dans la pratique, l'information complète serait généralement requise, parce qu'il est peu probable que les valeurs exactes des quantiles de plusieurs variables x puissent être importées de sources externes.) Ils déterminent les poids w_k de façon à minimiser la distance du chi-carré $\sum_s(w_k-d_k)^2/2d_kq_k$, pour la valeur spécifiée de q_k, sous la contrainte des équations de calage

$$\sum_s w_k=N;\ \hat{Q}_{x_j\mathrm{CAL},\alpha}=Q_{x_j\alpha},\ j=1,2,...,J$$

pour des estimations définies convenablement $\hat{Q}_{x_j\mathrm{CAL},\alpha}$. Maintenant, si nous décidions de spécifier $\hat{Q}_{x_j\mathrm{CAL},\alpha}$ comme $\hat{Q}_{x_j\alpha}=\inf\{t|\hat{F}_{x_j\mathrm{CAL}}(t)\geq\alpha\}$, il serait généralement impossible de trouver une solution exacte au problème de calage tel qu'il est énoncé. Harms et Duchesne choisissent plutôt de lui substituer des estimateurs lissés, qu'ils appellent « estimateurs interpolés de répartition » des fonctions de répartition $F_{x_j}(t)$, $j=1,2,...,J$. Ils remplacent $\Delta(\cdot)$ par une fonction légèrement modifiée. Il est alors possible d'obtenir les poids w_k, ainsi qu'une fonction de répartition estimée correspondante $\hat{F}_{y\mathrm{CAL}}(t)$; enfin, ils estiment $Q_{y\alpha}$ comme $\hat{Q}_{y\alpha}=\hat{F}^{-1}_{y\mathrm{CAL}}(\alpha)$.

Les poids calés résultants w_k nous permettent d'extraire les quantiles de population connus des variables auxiliaires. La chose est rassurante, car on s'attendrait à ce que ces poids produisent des estimateurs raisonnables des quantiles de la variable étudiée y. De surcroît, dans le cas d'une variable auxiliaire scalaire unique x, l'estimateur par calage résultant donne des quantiles de population exacts pour y quand la relation entre y et x est parfaitement linéaire, c'est-à-dire quand $y_k=\beta x_k$ pour tout $k\in U$. Une idée faisant intervenir des fonctions de répartition lissées est également mentionnée dans Tillé (2002).

La méthode mathématiquement plus simple de Rueda et coll. (2007) est une application du calage fondée sur un modèle en ce que leur calage se fait par rapport à un total de population des valeurs *prévues* de y. Elle requiert une information auxiliaire complète. Partant de la valeur connue $\mathbf{x}_k$, elle consiste à calculer d'abord les prédictions linéaires $\hat{y}_k=\hat{\boldsymbol{\beta}}'\mathbf{x}_k$ pour $k\in U$, avec $\hat{\boldsymbol{\beta}}=(\sum_s d_kq_k\mathbf{x}_k\mathbf{x}_k')^{-1}(\sum_s d_kq_k\mathbf{x}_ky_k)$, où $d_k=1/\pi_k$ et les q_k sont des facteurs d'échelle spécifiés. Les poids w_k sont obtenus par minimisation de la distance du chi-carré sous la contrainte des équations de calage énoncées en fonction des prédictions, de façon à réaliser la convergence à J points choisis arbitrairement t_j, $j=1,...,J$:

$$\frac{1}{N}\sum_s w_k\Delta(t_j-\hat{y}_k)=F_{\hat{y}}(t_j),\ j=1,...,J$$

où $F_{\hat{y}}(t_j)$ est la fonction de répartition en population finie des prédictions $\hat{y}_k$, évaluées à t_j. Les auteurs pensent qu'un assez petit nombre de points sélectionnés arbitrairement t_j peut suffire, disons moins de 10. Une fois que les w_k sont déterminés, l'estimation du quantile α est obtenue d'après $\hat{F}_{y\mathrm{CAL}}(t)=(1/N)\sum_s w_k\Delta(t-y_k)$.

L'estimation des quantiles illustre bien le fait que la méthode de calage peut être exécutée de plus d'une façon lorsqu'on estime des paramètres un peu plus complexes. Les deux méthodes mentionnées ici donnent une estimation presque sans biais sous le plan. Les poids de Harms et Duchesne (2006) sont polyvalents, indépendants de la variable y. Par contre, la méthode de Rueda et coll. (2007) nécessite un nouvel ensemble de poids pour chaque nouvelle variable y. Des données empiriques, obtenues par simulation, donnent à penser que les deux méthodes se comparent favorablement aux méthodes antérieures d'estimation des quantiles, non fondées explicitement sur l'approche du calage (mais sur la même information auxiliaire).

L'extension de la méthode du calage à l'estimation d'autre paramètres complexes, tels que le coefficient de Gini, est esquissée dans Harms et Duchesne (2006).

6.2 Calage pour d'autres paramètres complexes

Plikusas (2006), ainsi que Krapavickaitė et Plikusas (2005) examinent l'estimation par calage de certaines fonctions de totaux de population. (Leur expression « calage non linéaire » signifie « fonction non linéaire de totaux » et n'est pas utilisée ici.) Un exemple simple est l'estimation du ratio de deux totaux, $R = \sum_U y_{1k} / \sum_U y_{2k}$, où y_{1k} et y_{2k} sont les valeurs, pour l'élément k, des variables y_1 et y_2, respectivement. (En fait, la fonction de répartition (6.1) est également du type ratio, avec $y_{2k} = 1$ et $N = \sum_U 1$ comme total au dénominateur.) Ces auteurs étudient l'estimateur par calage $\hat{R}_{\text{CAL}} = \sum_s w_k y_{1k} / \sum_s w_k y_{2k}$. Les poids w_k, communs au numérateur et au dénominateur, sont déterminés par calage sur l'information auxiliaire énoncée comme il suit : Il existe une variable auxiliaire, x_{1k}, pour y_{1k}, et une autre, x_{2k}, pour y_{2k}; le ratio des totaux $R_0 = \sum_U x_{1k} / \sum_U x_{2k}$ est une valeur connue, obtenue par un dénombrement complet antérieur ou provenant d'une autre source fiable. L'équation de calage proposée est $\sum_s w_k e_k = 0$, où $e_k = x_{1k} - R_0 x_{2k}$. Parce que $\sum_U e_k = 0$, selon la méthode de la distance du chi-carré minimale, les poids sont

$$w_k = d_k \left\{ 1 - \left(\sum\nolimits_s d_k e_k \right) \left(\sum\nolimits_s d_k e_k^2 \right)^{-1} e_k \right\}.$$

Ces poids extraient correctement la valeur connue du ratio R_0; en posant que $y_{1k} = x_{1k}$ et $y_{2k} = x_{2k}$ dans $\hat{R}_{\text{CAL}}$, nous obtenons

$$\frac{\sum_s w_k x_{1k}}{\sum_s w_k x_{2k}} - R_0 = \frac{\sum_s w_k e_k}{\sum_s w_k x_{2k}} = 0.$$

Les données empiriques présentées dans Plikusas (2006), ainsi que dans Krapavickaitė et Plikusas (2005) donnent à penser que l'estimateur par calage se compare favorablement (variance plus faible, tout en maintenant la quasi-absence de biais sous le plan) à d'autres estimateurs, établis suivant d'autres arguments que le calage, tout en s'appuyant sur la même information auxiliaire.

7. Comparaison du calage à d'autres approches

Comme beaucoup l'ont souligné, les utilisateurs voient dans le calage un moyen simple et convaincant d'intégrer l'information auxiliaire, pour des paramètres simples (section 4), ainsi que pour des paramètres plus complexes, tels que les quantiles, les ratios et d'autres (section 6). Sa simplicité et son caractère pratique sont des avantages indéniables, mais peut-on aussi affirmer que le calage est « supérieur du point de vue théorique »? Existe-t-il des situations où l'on peut montrer que le calage donne des réponses plus exactes et/ou plus satisfaisantes à des questions d'importance que d'autres approches fondées sur le plan de sondage?

La section 4.5 donne une raison de penser que l'approche du calage pourrait être supérieure à l'approche GREG, parce que le calage basé sur un modèle pourrait donner des estimations plus précises que l'estimateur GREG non linéaire, pour le même modèle auxiliaire. La section 7.1 qui suit offre un autre exemple où l'approche du calage et l'approche GREG produisent des réponses divergentes, l'avantage allant à la méthode de calage.

7.1 Un exemple d'estimation par domaine

L'exemple présenté ici, tiré d'Estevao et Särndal (2004), illustre, pour une situation pratique simple, un conflit entre les résultats de l'approche GREG et de celle du calage. Le contexte est l'estimation du total y d'une sous-population (domaine).

Un échantillon probabiliste s est tiré de $\{1, 2, ..., k, ..., N\}$; les poids de sondage connus sont $d_k = 1/\pi_k$. Soit U_a un domaine; $U_a \subset U$. L'indicateur de domaine est δ_{ak} dont la valeur est $\delta_{ak} = 1$ si $k \in U_a$ et $\delta_{ak} = 0$ autrement. La grandeur estimée est le total de domaine $Y_a = \sum_U y_{ak}$, où $y_{ak} = \delta_{ak} y_k$, et y_k est observé pour $k \in s$. L'estimateur $\hat{Y}_{\text{HT}} = \sum_s d_k y_{ak}$ de Horvitz-Thompson est sans biais sous le plan de sondage, mais sa précision est faible, surtout si le domaine est petit, et l'utilisation d'information auxiliaire permettra de l'améliorer. Nous spécifions une valeur du vecteur auxiliaire $\mathbf{x}_k$ pour chaque $k \in U$.

Comme il est fréquent en pratique, les éléments qui appartiennent à un domaine d'intérêt ne sont pas précisés dans la base de sondage (s'ils le sont, on dispose d'information très puissante dès le départ, mais souvent, les conditions réelles ne sont pas aussi favorables). Néanmoins,

supposons que les éléments d'un groupe plus grand U_C sont identifiables; $U_a \subset U_C \subset U$. Par exemple, supposons que y est le « revenu » et que U_C est un groupe professionnel spécifié pour les personnes énumérées dans la base de sondage, tandis que U_a est un sous-groupe professionnel non défini dans la base de sondage. Nous pouvons identifier les sous-ensembles d'échantillons $s_C = s \cap U_C$ et $s_a = s \cap U_a$ et tirer parti du fait que nous connaissons le total $\sum_U \mathbf{x}_{Ck}$, qui peut être estimé sans biais par $\sum_s d_k \mathbf{x}_{Ck}$, où $\mathbf{x}_{Ck} = \delta_{Ck}\mathbf{x}_k$, et δ_{Ck} est l'indicateur de groupe informatif : $\delta_{Ck} = 1$ si $k \in U_C$ et $\delta_{Ck} = 0$ autrement. Le total auxiliaire de domaine $\sum_U \mathbf{x}_{ak}$ n'est pas disponible, parce que U_a n'est pas défini. Le calage en vue de satisfaire $\sum_s w_k \mathbf{x}_{Ck} = \sum_U \mathbf{x}_{Ck}$ donne l'estimateur presque sans biais sous le plan $\hat{Y}_{a\text{CAL}} = \sum_U w_k y_{ak}$, où $w_k = d_k(1+\boldsymbol{\lambda}'\mathbf{z}_k)$, avec $\boldsymbol{\lambda}' = (\sum_U \mathbf{x}_{Ck} - \sum_s d_k \mathbf{x}_{Ck})' \, (\sum_s d_k \mathbf{z}_k \mathbf{x}'_{Ck})^{-1}$. L'instrument asymptotiquement optimal pour le vecteur donné $\mathbf{x}_k$ est (voir la section 4.3) $\mathbf{z}_k = \mathbf{z}_{0Ck} = d_k^{-1} \sum_{\ell \in s}(d_k d_\ell - d_{k\ell})\mathbf{x}_{C\ell}$.

Par contre, l'approche par la régression en s'appuyant sur la même information auxiliaire mène à $\hat{Y}_{a\text{GREG}} = \sum_s d_k y_{ak} + (\sum_U \mathbf{x}_{Ck} - \sum_s d_k \mathbf{x}_{Ck})'\mathbf{B}_{\tilde{s};d}$, également presque sans biais sous le plan, où le coefficient de régression $\mathbf{B}_{\tilde{s};d} = (\sum_{\tilde{s}} d_k \mathbf{x}_k \mathbf{x}'_k)^{-1} \sum_{\tilde{s}} d_k \mathbf{x}_k y_k$ est le résultat d'un ajustement par les moindres carrés pondérés à un niveau approprié, en utilisant tous les points de données $(y_k, \mathbf{x}_k)$ disponibles pour $k \in s$ (quand $\tilde{s} = s$) ou une partie d'entre eux (quand $\tilde{s} \subset s$).

Ainsi, le modélisateur pourrait opter pour un ajustement de la régression « s'étendant au-delà du domaine » (de sorte que $\tilde{s} \supset s_a = s \cap U_a$) pour essayer d'emprunter de l'information pour $\hat{Y}_{a\text{GREG}}$ en le laissant dépendre également des données sur y provenant de l'extérieur du domaine. En revanche, $\hat{Y}_{q\text{CAL}}$ dépend exclusivement des données sur y contenues dans le domaine, ce qui est effectivement mieux. Estevao et Särndal (2004) montrent que $\hat{Y}_{a\text{CAL}}$ avec $\mathbf{z}_k = \mathbf{z}_{0Ck}$ a une variance (asymptotique) plus faible que $\hat{Y}_{a\text{GREG}}$, quelle que soit la façon dont $\tilde{s}$ est choisi. L'apport de données sur y en provenance de l'extérieur n'aide pas; l'approche du calage et l'approche de la régression ne s'accordent pas.

8. Estimation par calage en présence d'information composite

Comme l'ont montré les sections précédentes, beaucoup d'auteurs ont choisi d'étudier l'estimation dans le cas d'un échantillonnage direct, à une seule phase, d'éléments, sans aucune non-réponse. L'information disponible pour le calage est simple. Le k^{e} élément de la population finie $U = \{1, 2, ..., k, ..., N\}$ est associé à une valeur $\mathbf{x}_k$ du vecteur auxiliaire.

Toutefois, dans une importante catégorie de situations, l'information auxiliaire a une *structure composite* et sa complexité augmente avec celle du plan d'échantillonnage. Dans les plans à deux phases ou plus, ou à deux degrés ou plus, l'information comporte habituellement plusieurs composantes, qui reflètent les caractéristiques du plan de sondage. Elle est énoncée en fonction de plus d'un vecteur de variables auxiliaires. Par exemple, dans l'échantillonnage à deux degrés, une partie de l'information pourrait être disponible au sujet des unités d'échantillonnage de premier degré (grappes), et une autre partie, au sujet des unités de deuxième degré (éléments).

Par conséquent, l'estimation par calage (ou par toute autre méthode) doit systématiquement tenir compte de la structure de l'information. L'information complète comporte plusieurs éléments et le calage peut être effectué de plus d'une façon. Tous les éléments pertinents doivent être pris en compte, afin d'obtenir les estimations les plus exactes possibles. Accomplir cela de façon générale ou « optimale » n'est pas une tâche banale. L'approche du calage offre un moyen de le faire.

L'approche de la régression, appuyée sur un modèle auxiliaire dûment formulé, en est un autre, mais certains utilisateurs pourraient la juger moins directe. Donc, les enquêtes qui permettent l'utilisation d'information auxiliaire composite apportent un éclairage supplémentaire sur ce qui distingue l'approche du calage de l'approche GREG.

Nous discutons de l'échantillonnage à deux phases et de l'échantillonnage à deux degrés à la présente section. À la section 9, nous examinons un autre exemple d'information composite, dans le contexte de la correction du biais de non-réponse.

Un autre aspect de l'information composite est celui où le but est de combiner des renseignements provenant de plusieurs enquêtes, ce qui constitue aussi un moyen de renforcer les estimations et d'en accroître l'exactitude. Ce facteur est l'un de ceux (outre le désir de réaliser la cohérence entre les enquêtes pour le bénéfice de l'utilisateur des données) qui motivent la méthode de pondération répétée mentionnée antérieurement, qui a été élaborée par l'organisme statistique des Pays-Bas. L'utilisation d'informations auxiliaires combinées dans le contexte de l'estimation GREG est examinée dans Merkouris (2004).

8.1 Information composite pour les plans d'échantillonnage à deux phases

Par échantillonnage double, on entend un plan de sondage comportant le tirage de deux échantillons probabilistes, s_1 et s_2, à partir de la même population $U = \{1, ..., k, ..., N\}$. Des données auxiliaires peuvent être relevées pour les deux échantillons, mais les valeurs de la variable étudiée y_k ne sont relevées que pour $k \in s_2$ dans le but d'estimer $Y = \sum_U y_k$. Hidiroglou (2001) distingue

plusieurs formes d'échantillonnage double. Dans le cas *hiérarchique* (*ou emboîté*) (échantillonnage à deux phases classique), l'échantillon de première phase s_1 est tiré de U, et l'échantillon de seconde phase s_2 est un sous-échantillon tiré de s_1, de sorte que $U \supset s_1 \supset s_2$. Nous pouvons distinguer deux *cas non hiérarchiques* (*ou non emboîtés*). Dans le premier, s_1 est tiré de la base de sondage U_1 et s_2, de la base de sondage U_2, où U_1 et U_2 couvrent la même population U, et les unités d'échantillonnage peuvent être définies différemment dans les deux bases. Dans le deuxième cas non hiérarchique, s_1 et s_2 sont tirés indépendamment de U.

Afin d'illustrer comment l'information composite intervient dans l'estimation, considérons le cas hiérarchique. Les poids de sondage sont $d_{1k} = 1/\pi_{1k}$ (s_1 échantillonné dans U); $d_{2k} = 1/\pi_{2k} (\pi_{2k} = \pi_{k|s_1}$ dans le sous-échantillonnage de s_2 à partir de s_1). Le poids de sondage combiné est $d_k = d_{1k} d_{2k}$. L'estimateur sans biais élémentaire $\hat{Y} = \sum_{s_2} d_k y_k$ peut être amélioré grâce à l'utilisation d'information auxiliaire, spécifiée ici à deux niveaux.

Niveau de la population : La valeur du vecteur $\mathbf{x}_{1k}$ est connue (donnée dans la base de sondage) pour chaque $k \in U$, de sorte qu'elle est connue pour chaque $k \in s_1$ et pour chaque $k \in s_2$; $\sum_U \mathbf{x}_{1k}$ est un total vectoriel de population connu.

Niveau du premier échantillon : La valeur du vecteur auxiliaire $\mathbf{x}_{2k}$ est connue (observée) pour chaque $k \in s_1$ et, par conséquent, pour chaque $k \in s_2$; le total inconnu $\sum_U \mathbf{x}_{2k}$ est estimé sans biais par $\sum_{s_1} d_{1k} \mathbf{x}_{2k}$.

Quel est le meilleur moyen de tenir compte de cette information composite? Dans une adaptation de l'approche GREG, Särndal et Swensson (1987) ont formulé deux modèles auxiliaires linéaires, le premier spécifié en fonction du vecteur $\mathbf{x}_{1k}$ et le second tenant compte aussi du vecteur $\mathbf{x}_{2k}$. Les deux modèles sont ajustés et les prédictions résultantes, de deux types, sont utilisées pour créer un estimateur GREG approprié $\hat{Y}_{\text{GREG}}$ de $Y = \sum_U y_k$.

Dupont (1995) fait le commentaire important que l'information composite donnée appelle « deux approches naturelles différentes ». Outre l'approche GREG, il existe une approche par calage qui donnera les poids finaux w_k pour un estimateur par calage $\hat{Y}_{\text{CAL}} = \sum_{s_2} w_k y_k$. Il est intéressant de comparer les résultats de ces deux approches. L'une et l'autre offrent plus d'une option. Dans l'approche GREG, il existe divers moyens de formuler les modèles auxiliaires linéaires et leur structure de variance respective. Dans l'approche par le calage, diverses formulations des équations de calage sont possibles.

Voici, par exemple, une option de *calage en deux étapes* : d'abord trouver les poids intermédiaires w_{1k} qui satisfont $\sum_{s_1} w_{1k} \mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$, puis utiliser ces poids à la deuxième étape pour calculer les poids finaux w_k qui satisfont

$$\sum_{s_2} w_k \mathbf{x}_k = \sum_{s_1} w_{1k} \mathbf{x}_k = \begin{pmatrix} \sum_U \mathbf{x}_{1k} \\ \sum_{s_1} w_{1k} \mathbf{x}_{2k} \end{pmatrix}$$

où $\mathbf{x}_k$ est le vecteur auxiliaire combiné

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_{1k} \\ \mathbf{x}_{2k} \end{pmatrix}.$$

Alternativement, dans une option *à une seule étape*, nous déterminons les poids w_k directement pour satisfaire

$$\sum_{s_2} w_k \mathbf{x}_k = \begin{pmatrix} \sum_U \mathbf{x}_{1k} \\ \sum_{s_1} d_{1k} \mathbf{x}_{2k} \end{pmatrix}.$$

En général, les poids finaux w_k ne sont pas identiques dans les deux options. Supposons que $\sum_U \mathbf{x}_{1k}$ est un total $\mathbf{x}_1$ importé. Si nous examinons la situation de plus près, nous constatons que l'option en deux étapes nécessite plus d'information, parce que les valeurs connues $\mathbf{x}_{1k}$ sont requises individuellement pour $k \in s_1$, tandis que dans l'option à une seule étape, il suffit qu'elle soit disponible pour $k \in s_2$. Nous pourrions donc nous attendre à un certain avantage de l'option à deux étapes en ce qui concerne la variance, puisque $\sum_{s_1} w_{1k} \mathbf{x}_{2k}$ est souvent plus exact (en tant qu'estimateur de $\sum_U \mathbf{x}_{2k}$) que $\sum_{s_1} d_{1k} \mathbf{x}_{2k}$ dans la procédure à une seule étape. Néanmoins, cette attente n'est pas toujours confirmée et la méthode à une seule étape peut être supérieure, par exemple quand $\mathbf{x}_1$ et $\mathbf{x}_2$ sont faiblement corrélés.

Dupont (1995), ainsi qu'Hidiroglou et Särndal (1998) examinent les liens qui existent, sans surprise, entre les deux approches. Un estimateur GREG, dérivé de modèles adjoints ayant des structures de variance spécifiques, pourrait être identique à l'estimateur par calage si les poids de ce dernier sont calés d'une certaine façon. Dans d'autres cas, les différences pourraient être faibles.

L'efficacité de diverses options dépend, de manière assez subtile, de la configuration des corrélations entre y_k, $\mathbf{x}_{1k}$ et $\mathbf{x}_{2k}$. Par exemple, dans quelle mesure $\mathbf{x}_1$ et $\mathbf{x}_2$ sont-ils complémentaires, dans quelle mesure sont-ils des substituts l'un de l'autre? Dans l'approche GREG, il est difficile, voire même futile, de définir avec précision une structure de variance qui reflète vraiment la « réalité » qui sous-tend les données. L'approche du calage est plus directe et Estevao et Särndal (2002, 2006) explorent certaines de ses possibilités.

8.2 Information composite dans les plans d'échantillonnage à deux degrés

Le point commun entre l'échantillonnage à deux degrés classique (grappes échantillonnées au premier degré,

éléments sous-échantillonnés dans les grappes sélectionnées au deuxième degré) et l'échantillonnage à deux phases tient au fait que l'information totale peut comporter plus d'une composante. Il peut exister a) de l'information au niveau de la grappe (au sujet des grappes), b) de l'information au niveau de l'élément pour toutes les grappes et c) de l'information au niveau de l'élément pour les grappes selectionnées uniquement. Ici encore, les auteurs appartiennent à deux écoles, certains exploitant l'information par l'approche du calage et les autres choisissant la route de la régression généralisée GREG.

Estevao et Särndal (2006) conçoivent l'estimation par calage sous échantillonnage à deux degrés classique, où l'information composite est spécifiée comme il suit : i) pour la population de grappes $U_{\rm I}$, il existe un total connu $\sum_{U_{\rm I}} \mathbf{x}_{(c)i}$, où $\mathbf{x}_{(c)i}$ est une valeur du vecteur auxiliaire associé à la grappe U_i, pour $i \in U_{\rm I}$; ii) pour la population d'éléments $U = \bigcup_{i \in U_{\rm I}} U_i$, il existe un total connu $\sum_U \mathbf{x}_k$, où la valeur du vecteur auxiliaire $\mathbf{x}_k$ est associée à l'élément $k \in U$. Supposons que l'on doit produire à la fois les statistiques de grappe et les statistiques d'élément dans une enquête, si bien qu'il faut estimer le total de population de grappes $Y_{\rm I} = \sum_{U_{\rm I}} y_{(c)i}$ et le total de population d'éléments $Y = \sum_U y_k$.

Si aucune relation n'est imposée entre les poids des grappes $w_{{\rm I}i}$ et les poids des éléments w_k, les premiers sont calés de manière à satisfaire $\sum_{s_{\rm I}} w_{{\rm I}i} \mathbf{x}_{(c)i} = \sum_{U_{\rm I}} \mathbf{x}_{(c)i}$ et les seconds, de manière à satisfaire $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. (Ici, $s_{\rm I}$ est l'échantillon de grappes tiré de $U_{\rm I}$; s_i est l'échantillon d'éléments tiré de U_i; et $s = \bigcup_{i \in s_{\rm I}} s_i$ est l'échantillon complet d'éléments.) Alors, $\hat{Y}_{\rm ICAL} = \sum_{s_{\rm I}} w_{{\rm I}i} y_{(c)i}$ estime le total de population de grappes $Y_{\rm I}$ et $\hat{Y}_{\rm CAL} = \sum_s w_k y_k$ estime le total de population d'éléments Y.

Dans la pratique, on recourt souvent à la *pondération intégrée*, qui consiste à imposer une relation commode entre les poids des grappes $w_{{\rm I}i}$ et les poids w_k des éléments compris dans les grappes sélectionnées. Estevao et Särndal (2006) examinent deux formes de pondération intégrée.

L'une d'elles consiste à imposer $w_k = d_{k|i} w_{{\rm I}i}$, où $d_{k|i}$ est l'inverse de la probabilité d'inclusion de l'élément *k* dans la grappe *i*. (Par exemple, dans l'échantillonnage en grappes à un degré, quand tous les éléments *k* d'une grappe échantillonnée sont sélectionnés, alors $d_{k|i} = 1$. Par conséquent, la relation imposée est $w_k = w_{{\rm I}i}$ et tous les éléments de la grappe reçoivent le même poids pour le calcul des statistiques d'élément, et ce même poids est également utilisé pour calculer les statistiques de grappe.) L'équation de calage $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$ se lit alors $\sum_{s_{\rm I}} w_{{\rm I}i} \sum_{s_i} d_{k|i} \mathbf{x}_k = \sum_U \mathbf{x}_k$. Les poids de grappe $w_{{\rm I}i}$ sont maintenant calculés en minimisant $\sum_{s_{\rm I}} (w_{{\rm I}i} - d_{{\rm I}i})^2 / d_{{\rm I}i}$ sous la contrainte de l'équation de calage qui tient compte des deux types d'information :

$$\begin{pmatrix} \sum_{s_{\rm I}} w_{{\rm I}i} \mathbf{x}_{(c)i} \\ \sum_{s_{\rm I}} w_{{\rm I}i} \sum_{s_i} d_{k|i} \mathbf{x}_k \end{pmatrix} = \begin{pmatrix} \sum_{U_{\rm I}} \mathbf{x}_{(c)i} \\ \sum_U \mathbf{x}_k \end{pmatrix}. \qquad (8.1)$$

Une fois les poids de grappe $w_{{\rm I}i}$ déterminés, le calcul des poids d'élément $w_k = d_{k|i} w_{{\rm I}i}$ s'ensuit.

Une autre méthode de pondération intégrée raisonnable consiste à imposer que $\sum_{s_i} w_k = N_i w_{{\rm I}i}$. Par exemple, pour l'échantillonnage en grappes à un degré, cela implique que le poids de grappe $w_{{\rm I}i}$ est égal à la moyenne des poids d'élément w_k contenus dans la grappe.

L'échantillonnage à deux degrés est également traité dans Kim, Breidt et Opsomer (2006). Ces auteurs supposent que des données auxiliaires existent pour les grappes, par la voie d'une variable de grappe quantitative unique $x_{(c)i}$, mais non pour les éléments. Ils développent et étudient un estimateur de type GREG du total de population d'éléments $Y = \sum_U y_k$, $\hat{Y} = \sum_{i \in U_{\rm I}} \hat{\mu}_i + \sum_{i \in s_{\rm I}} d_i (\hat{t}_i - \hat{\mu}_i)$, où $\hat{t}_i$ est sans biais sous le plan pour le total de population de grappes $t_i = \sum_{U_i} y_k$, et $\hat{\mu}_i$ est obtenu par un ajustement par régression polynomiale locale. L'estimateur peut être exprimé sous la forme linéairement pondérée en utilisant des poids qui sont calés sur les totaux de population des puissances de la variable de grappe $x_{(c)i}$.

8.3 Pondération des ménages et pondération des personnes

Dans certaines grandes enquêtes sociales, l'objectif est de produire des estimations au niveau du ménage ainsi qu'au niveau de la personne, si bien que certaines variables étudiées sont des variables du ménage (grappe) et d'autres, des variables de la personne (élément). Par conséquent, un certain nombre d'auteurs ont étudié la situation de l'échantillonnage en grappes à un degré ($d_{k|i} = 1$) et de la pondération intégrée qui consiste à attribuer le même poids à tous les membres d'un ménage sélectionné, poids qui est également utilisé pour produire les statistiques au niveau du ménage. Une solution générale de ce problème de pondération, si l'information au niveau du ménage et l'information au niveau de la personne sont l'une et l'autre spécifiées, consiste à obtenir les poids des ménages $w_{{\rm I}i}$, calés comme dans l'équation (8.1) avec $d_{k|i} = 1$, puis à prendre $w_k = w_{{\rm I}i}$.

Dans plusieurs articles, l'accent est mis sur les valeurs du vecteur auxiliaire $\mathbf{x}_k$ attribuées aux personnes. Alexander (1987) calcule des poids qui minimisent la distance du chi-carré, tandis que Lemaître et Dufour (1987) et Niewenbrook (1993) calculent les poids intégrés à l'aide d'un estimateur GREG. La méthode de Lemaître et Dufour procède par construction indirecte d'une « valeur de vecteur auxiliaire équipondérée » s'appliquant à toutes les personnes membres d'un ménage sélectionné. Leur résultat est calculable par la méthode directe exposée à la section 8.2.

Le auteurs d'articles plus récents reviennent sur la question de la double pondération ménage et personne. Certains adoptent l'approche du calage et d'autres, celle de la régression généralisée GREG. Isaki, Tsay et Fuller (2004) en font un problème de pondération par calage; ils appliquent des poids calés aux totaux de contrôle au niveau du ménage, ainsi qu'au niveau de la personne et ne formulent aucun modèle auxiliaire explicite. Par contre, Steel et Clark (2007) suivent l'approche GREG, avec spécification de modèles auxiliaires linéaires et des structures de variance connexes.

9. Calage pour corriger la non-réponse

9.1 Correction classique de la non-réponse

Nombre de bons articles théoriques ont pour contexte le cas simple de la section 2, qui inclut l'absence totale de non-réponse. Ce sont de bonnes théories, mais applicables dans des conditions qui ne se réalisent que rarement, voire jamais. (À titre d'auteur d'articles dans ce domaine, je ne suis moi-même pas irréprochable.) La non-réponse existe dans pratiquement toutes les enquêtes. Bien qu'elle ne soit pas souhaitable, il s'agit d'un phénomène naturel dont la théorie devrait tenir compte d'emblée, en adoptant une perspective de sélection à deux phases.

Dans de nombreuses enquêtes, les taux de non-réponse sont aujourd'hui très élevés si on les compare à ceux d'il y a 40 ans, qui étaient si faibles qu'on pouvait essentiellement ne pas en tenir compte. De nos jours, la théorie de l'échantillonnage doit s'attacher de plus en plus souvent aux conséquences indésirables de la non-réponse. En particulier, un objectif immédiat est d'examiner le biais et d'essayer de le réduire autant que possible

Soit un échantillon probabiliste s tiré de $U = \{1, 2, ..., k, ..., N\}$; le poids de sondage connu de l'élément k est $d_k = 1/\pi_k$. Il y a non-réponse, ce qui donne un ensemble de réponses r, qui est un sous-ensemble de s; la valeur de la variable étudiée y_k est observée pour $k \in r$ uniquement. La probabilité de réponse inconnue de l'élément k est $\Pr(k \in r|s) = \theta_k$. Nous écartons l'estimateur sans biais $\hat{Y} = \sum_r d_k \phi_k y_k$, parce que $\phi_k = 1/\theta_k$ est inconnue. Si nous voulons retenir la notion de somme linéairement pondérée, comment devons-nous construire les poids? La correction par pondération de la non-réponse totale, ou non-réponse d'une unité, en se fondant sur la « modélisation de la non-réponse » se fait depuis longtemps. Le calage offre une nouvelle perspective.

Dans ce que nous pourrions appeler la « méthode classique », les poids de sondage probabilistes $d_k = 1/\pi_k$ sont d'abord corrigés de la non-réponse et, éventuellement, d'autres imperfections, telles que les valeurs aberrantes. L'information utilisée à cette étape provient souvent du regroupement des éléments échantillonnés. Enfin, si des totaux de population fiables sont disponibles, les poids de sondage corrigés sont calés sur ces totaux.

Au Canada, la méthodologie de l'Enquête sur la population active, décrite dans Statistique Canada (1998), est un exemple de cette pratique répandue. Un poids de sondage (modifié) est d'abord calculé pour un ménage donné, par multiplication de trois facteurs. Le produit du poids de sondage et d'un facteur de correction de la non-réponse est appelé le sous-poids. À la dernière étape, les sous poids sont calés sur des estimations de population postcensitaire très précises par groupe d'âge, sexe et région infraprovinciale. Les poids finaux ont les propriétés souhaitées de convergence, dans les régions d'une province, avec les estimations postcensitaires. Le biais de non-réponse qui persiste dans les estimations résultantes est inconnu, mais considéré comme modeste.

La méthode classique est intégrée dans le type d'estimateur $\hat{Y} = \sum_r d_k (1/\hat{\theta}_k) y_k$, où θ_k a été estimé par $\hat{\theta}_k$ à une étape préliminaire, en modélisant la réponse (c'est- à-dire la propension à répondre). Ce que demande la théorie au statisticien, à savoir formuler le « modèle de réponse réelle », capable de fournir des valeurs $\hat{\theta}_k$ exactes, sans biais, n'est pas une tâche facile. Toutefois, dans de nombreuses enquêtes, les facteurs $1/\hat{\theta}_k$ sont appliqués machinalement, sans esprit critique, par exemple par extension directe dans les strates déjà utilisées pour la sélection de l'échantillon.

La méthode classique est appliquée, par exemple, dans Ekholm et Laaksonen (1991) et dans Rizzo, Kalton et Brick (1996).

Souvent, les praticiens agissent comme si l'estimateur résultant $\hat{Y} = \sum_r d_k (1/\hat{\theta}_k) y_k$ (découlant d'une modélisation plus ou moins pénétrante de la réponse pour essayer d'obtenir les $\hat{\theta}_k$) était essentiellement sans biais, ce qu'il n'est pas (à moins qu'on ait eu la chance de spécifier le modèle idéal). Ils se comportent (pour les besoins de l'estimation de la variance, par exemple) comme si $\pi_k \hat{\theta}_k$ était la probabilité de sélection réelle de l'élément k dans une étape unique de sélection, ce qui n'est définitivement pas le cas. Cette pratique, dont l'origine remonte à un passé idyllique, devient de plus en plus contestable à mesure que grimpent subrepticement les taux de non réponse.

Le remplacement de θ_k par $\hat{\theta}_k$ donne inévitablement lieu à un biais. Il y a plusieurs décennies, les taux de non-réponse n'étaient habituellement que de quelques pour cents et il était justifiable d'ignorer ce biais, mais leur croissance galopante rend cette pratique indéfendable aujourd'hui. Selon les premiers principes, l'objectif est de produire une estimation sans biais et non une estimation où le carré du biais est un contributeur dominant (et inconnu) à l'erreur

quadratique moyenne. Nous devons décider de limiter le biais autant que possible. L'approche du calage peut nous aider à construire un vecteur auxiliaire qui répond à cet objectif.

9.2 Calage pour la correction du biais de non-réponse

Faisant plus ou moins contraste avec la procédure classique, un certain nombre d'articles récents mettent l'accent sur l'approche du calage pour corriger la non-réponse. Certaines références récentes à cet égard sont Deville (1998, 2002), Ardilly (2006), chapitre 3, Skinner (1998), Folsom et Singh (2000), Fuller (2002), Lundström et Särndal (1999), Särndal et Lundström (2005), et Kott (2006).

L'approche du calage débute par l'évaluation de l'information auxiliaire totale disponible au niveau de l'échantillon (valeurs des variables auxiliaires observées pour les répondants et les non-répondants) et au niveau de la population (totaux auxiliaires connus de population). L'objectif est de tirer le meilleur parti possible des deux sources combinées, afin de réduire le biais ainsi que la variance. Les poids de sondage sont modifiés, en une ou en deux étapes de calage, de façon qu'ils reflètent i) le résultat de la phase de réponse, ii) les caractéristiques individuelles des répondants et iii) l'information auxiliaire spécifiée. L'information peut se résumer comme il suit.

Niveau de la population : La valeur du vecteur auxiliaire $\mathbf{x}_k^*$ est connue (spécifiée dans la base de sondage) pour chaque $k \in U$, donc est connue pour chaque $k \in s$ et chaque $k \in r$; $\sum_U \mathbf{x}_k^*$ est un total de population connu.

Niveau de l'échantillon : La valeur du vecteur auxiliaire $\mathbf{x}_k^\circ$ est connue (observée) pour chaque $k \in s$, et est donc connue pour chaque $k \in r$; le total inconnu $\sum_U \mathbf{x}_k^\circ$ est estimé sans biais par $\sum_s d_k \mathbf{x}_k^\circ$.

Le calage sur cette information composite peut se faire en deux étapes (calcul de poids intermédiaires pour commencer, puis utilisation de ces poids à la deuxième étape pour produire les poids finaux) ou directement en une seule étape. En principe, les différences de biais et de variance des estimations devraient être modestes. Dans l'option en une seule étape, le vecteur auxiliaire combiné et l'information correspondante sont

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}; \mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}.$$

En utilisant une extension de la méthode du vecteur instrumental décrite à la section 4.3, nous recherchons les poids calés $w_k = d_k v_k$, où $v_k = F(\boldsymbol{\lambda}' \mathbf{z}_k)$ est le facteur de correction de la non-réponse, avec un vecteur $\boldsymbol{\lambda}$ déterminé par l'équation de calage $\sum_r w_k \mathbf{x}_k = \mathbf{X}$; l'estimateur par calage résultant est $\hat{Y}_{\text{CAL}} = \sum_r w_k y_k$. Il suffit de spécifier la valeur du vecteur instrumental $\mathbf{z}_k$ pour les répondants; $\mathbf{z}_k$ peut être différent de $\mathbf{x}_k$. La fonction $F(\cdot)$ joue le même rôle qu'aux section 4.2 et 4.3. Ici, $F(\boldsymbol{\lambda}' \mathbf{z}_k)$ estime implicitement la probabilité de réponse inverse, $\phi_k = 1/\theta_k$ comme l'ont souligné Deville (2002), Dupont (1995) et Kott (2006). Dans le cas linéaire, $F(u) = 1 + u$ et $v_k = 1 + \boldsymbol{\lambda}' \mathbf{z}_k$, avec $\boldsymbol{\lambda}' = (\sum_U \mathbf{x}_k - \sum_r d_k \mathbf{x}_k)' (\sum_r d_k \mathbf{z}_k \mathbf{x}_k')^{-1}$.

Bien qu'elles ne soient observées que pour les éléments échantillonnés, les variables qui constituent le vecteur $\mathbf{x}_k^\circ$ peuvent être de la plus haute importance pour la réduction du biais de non-réponse (quoique moins importantes que les $\mathbf{x}_k^*$ pour la réduction de la variance). Par exemple, Beaumont (2005b) discute des variables du processus de collecte des données qui peuvent être utilisées pour construire la composante vectorielle $\mathbf{x}_k^\circ$.

9.3 Construction du vecteur auxiliaire

Dans certaines enquêtes, les variables auxiliaires possibles abondent, comme le signalent, par exemple, Rizzo, Kalton et Brick (1996), de même que Särndal et Lundström (2005). Ainsi, pour les enquêtes auprès des ménages et des particuliers en Scandinavie, on peut obtenir une foule de variables auxiliaires éventuelles par appariement des données d'enquête à celles des registres administratifs de haute qualité existants. Il convient ensuite de décider lesquelles de ces variables devraient être incluses dans le vecteur auxiliaire $\mathbf{x}_k$ afin que celui-ci soit aussi efficace que possible, notamment en ce qui concerne la réduction du biais. Comme le mentionnent Rizzo, Kalton et Brick (1996), le choix des variables auxiliaires est probablement plus important que celui de la méthode de pondération.

Examinons le biais quand $\mathbf{z}_k = \mathbf{x}_k$. Nous devons comparer divers vecteurs $\mathbf{x}_k$ afin de choisir, en dernière analyse, celui qui produira vraisemblablement le biais le plus faible. (Nous posons que $\mathbf{x}_k$ est tel que $\boldsymbol{\mu}' \mathbf{x}_k = 1$ pour tout k et un vecteur constant $\boldsymbol{\mu}$, ce qui est le cas de nombreux vecteurs $\mathbf{x}_k$, y compris les exemples 1 à 5 présentés au début de la section 2.) Nous obtenons une bonne approximation du biais de $\hat{Y}_{\text{CAL}}$ par linéarisation de Taylor de la forme $\text{biaisapprox}(\hat{Y}_{\text{CAL}}) = (\sum_U \mathbf{x}_k)' (\mathbf{B}_{U;\theta} - \mathbf{B}_U)$, où intervient la différence entre le coefficient de régression pondéré $\mathbf{B}_{U;\theta} = \left(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_U \theta_k \mathbf{x}_k y_k$ et le coefficient non pondéré $\mathbf{B}_U = (\sum_U \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_U \mathbf{x}_k y_k)$. À moins que tous les θ_k soient égaux, le biais causé par l'écart entre les deux vecteurs de coefficients de régression pourrait être important, même si $\mathbf{x}_k$ semble être un « bon vecteur auxiliaire ». Cette expression du biais approximatif (*nearbias*) est donnée dans Särndal et Lundström (2005). Des expressions apparentées du biais (sous des conditions

différentes) sont proposées dans Bethlehem (1988) et dans Fuller, Loughin et Baker (1994). Nous pouvons aussi écrire $\text{biaisapprox}(\hat{Y}_{\text{CAL}}) = \sum_U (\theta_k M_k - 1) y_k$, où $M_k = (\sum_U \mathbf{x}_k)'(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k$. Dans la comparaison des variantes possibles de $\mathbf{x}_k$, une référence commode est le « vecteur auxiliaire primitif », $\mathbf{x}_k = 1$ pour tout $k \in U$, qui donne $\hat{Y}_{\text{CAL}} = N \bar{y}_r = N \sum_r y_k / n_r$, où n_r est le nombre de répondants, avec $\text{biaisapprox}(N \bar{y}_r) = N(\bar{y}_{U;\theta} - \bar{y}_U)$, où $\bar{y}_{U;\theta} = \sum_U \theta_k y_k / \sum_U \theta_k$ et $\bar{y}_U = \sum_U y_k / N$. Le ratio

$$\text{biaisrel}(\hat{Y}_{\text{CAL}}) = \frac{\text{biaisapprox}(\hat{Y}_{\text{CAL}})}{\text{biaisapprox}(N \bar{y}_r)} = \frac{\sum_U (\theta_k M_k - 1) y_k}{N(\bar{y}_{U;\theta} - \bar{y}_U)}$$

indique dans quelle mesure un vecteur candidat $\mathbf{x}_k$ réussit à contrôler le biais, comparativement au vecteur primitif. Nous recherchons un vecteur $\mathbf{x}_k$ qui donne un biais faible. Cependant, $\text{biaisrel}(\hat{Y}_{\text{CAL}})$ n'est pas un indicateur de biais calculable, car il dépend de y_k non observé et de θ_k inobservable. Il nous faut un indicateur calculable qui s'approche de $\text{biaisrel}(\hat{Y}_{\text{CAL}})$ et dépend du vecteur $\mathbf{x}$, mais non des variables y, qui peuvent être nombreuses dans l'enquête.

Il est facile de voir que $\text{biaisrel}(\hat{Y}_{\text{CAL}}) = 0$ dans le cas d'un vecteur $\mathbf{x}$ idéal (probablement inexistant) tel que $\phi_k = 1/\theta_k = \boldsymbol{\lambda}' \mathbf{x}_k$ pour tout $k \in U$ et un vecteur constant $\boldsymbol{\lambda}$.

Pour un vecteur $\mathbf{x}$ qui peut effectivement être construit dans le contexte de l'enquête, nous pouvons au moins obtenir les prédictions de ϕ_k. Si nous déterminons $\boldsymbol{\lambda}$ de manière à minimiser $\sum_U \theta_k (\phi_k - \boldsymbol{\lambda}' \mathbf{x}_k)^2$, nous obtenons $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}_U$, où $\hat{\boldsymbol{\lambda}}_U' = (\sum_U \mathbf{x}_k)'(\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1}$; la valeur prévue de ϕ_k est $\hat{\phi}_{kU} = \hat{\boldsymbol{\lambda}}_U' \mathbf{x}_k = M_k$. Les premier et deuxième moments (pondérés par thêta) des prédictions $\hat{\phi}_{kU} = M_k$ sont, respectivement, $\bar{M}_{U;\theta} = \sum_U \theta_k M_k / \sum_U \theta_k = N / \sum_U \theta_k = 1/\bar{\theta}_U$ et

$$Q = \frac{1}{\sum_U \theta_k} \sum_U \theta_k (M_k - \bar{M}_{U;\theta})^2 = (1/\bar{\theta}_U)(\bar{M}_U - 1/\bar{\theta}_U)$$

où $\bar{M}_U = \sum_U M_k / N$. Särndal et Lundström (2007) montrent que, dans certaines conditions, la relation entre $\text{biaisrel}(\hat{Y}_{\text{CAL}})$ et Q est approximativement linéaire,

$$\text{biaisrel}(\hat{Y}_{\text{CAL}}) \approx 1 - \frac{Q}{Q_0}$$

où $\bar{\phi}_U = \sum_U \phi_k / N$ et $Q_0 = (1/\bar{\theta}_U)(\bar{\phi}_U - 1/\bar{\theta}_U)$ est la valeur maximale de Q. Donc, si Q est calculable, il pourrait servir d'indicateur pour comparer les différents vecteurs $\mathbf{x}_k$ candidats. Nous obtenons plutôt un analogue calculable $\hat{Q}$ de Q en tant que variance des prédictions basées sur l'échantillon correspondantes $\hat{\phi}_{ks} = \hat{\boldsymbol{\lambda}}_s' \mathbf{x}_k = (\sum_s d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k = m_k$, de sorte que

$$\hat{Q} = \frac{1}{\sum_r d_k} \sum_r d_k (m_k - \bar{m}_{r;d})^2 = \bar{m}_{r;d}(\bar{m}_{s;d} - \bar{m}_{r;d})$$

où

$$\bar{m}_{r;d} = \frac{\sum_r d_k m_k}{\sum_r d_k} = \frac{\sum_s d_k}{\sum_r d_k}; \bar{m}_{s;d} = \frac{\sum_s d_k m_k}{\sum_s d_k}.$$

Nous nous attendons à ce que le biais relatif *(relbias*) diminue de manière approximativement linéaire à mesure que $\hat{Q}$ augmente. Par conséquent, indépendamment des variables y, $\hat{Q}$ peut être utilisé comme outil pour classer les divers vecteurs $\mathbf{x}$ en fonction de leur capacité à réduire le biais.

Nous pouvons nous servir de $\hat{Q}$ comme outil de sélection des variables x qu'il convient d'inclure dans le vecteur $\mathbf{x}_k$, par exemple par régression multiple ascendante de sorte que les variables soient ajoutées à $\mathbf{x}_k$ une à la fois, celle qui entre à une étape donnée étant celle qui produit l'accroissement le plus important de $\hat{Q}$. La méthode est décrite dans Särndal et Lundström (2007).

10. Calage pour tenir compte d'autres erreurs non dues à l'échantillonnage

Les erreurs dues à la non-réponse jouent un rôle déterminant dans la qualité des statistiques publiées. En revanche, si nous examinons la place que pourrait tenir l'approche du calage dans le traitement des erreurs non dues à l'échantillonnage d'autres sources que la non-réponse, il n'est pas surprenant que la littérature soit jusqu'à présent moins abondante. Néanmoins, plusieurs auteurs esquissent une approche de calage en vue d'intégrer également les erreurs de base de sondage, les erreurs de mesure et les valeurs aberrantes. Le calage pourrait offrir une théorie plus générale de l'estimation dans le contexte des enquêtes qui engloberaient les diverses erreurs non dues à l'échantillonnage.

Comme le souligne Deville (2004), le concept de calage s'applique aisément et efficacement à une grande variété de problèmes posés par les sondages. Selon lui, sa portée dépasse celle de l'estimation par la régression, une notion à laquelle certains semblent vouloir réduire l'approche du calage. Il expose sommairement comment cette approche permet de traiter plusieurs erreurs dues à la non-réponse.

Folsom et Singh (2000) présentent une méthode de calage des poids s'appuyant sur ce qu'ils appellent le modèle exponentiel généralisé (GEM). Elle comporte trois volets : le traitement des valeurs extrêmes, la correction de la non-réponse et le calage par poststratification. Elle fournit des contrôles intégrés pour les valeurs extrêmes. Le calage pour corriger à la fois les erreurs de couverture (sous- ou

surdénombrement dans la base de sondage) et la non-réponse est exposé dans Särndal et Lundström (2005) et dans Kott (2006). Skinner (1998) discute de l'utilisation du calage en présence de non-réponse et d'erreurs de mesure. Son commentaire sur la nécessité de poursuivre les travaux de recherche en vue d'étudier les propriétés des estimations par calage en présence d'erreur non due à l'échantillonnage demeure un défi près de dix ans plus tard.

11. Conclusion

Si une question doit être choisie en vue de formuler la conclusion du présent exposé, il s'agit selon moi du concept d'information auxiliaire, qui est le concept central de l'article. Sans information auxiliaire, il n'y a pas de calage, car il n'existe rien sur quoi l'exécuter. J'ai mentionné par ailleurs que l'estimation par la régression est un autre moyen, qui suit un raisonnement différent, de tenir compte de l'information auxiliaire dans l'estimation.

L'un des objectifs du présent article était de brosser le tableau de deux types de raisonnement et de souligner en quoi ils se distinguent. Des exemples montrent comment la réalisation d'un objectif d'estimation essentiellement semblable est abordée par certains auteurs suivant la logique du calage et par d'autres, suivant celle de la régression généralisée GREG (ou du moins principalement suivant l'un ou l'autre de ces types de raisonnement). Les estimateurs respectifs que ces auteurs finissent par recommander peuvent ou non donner des résultats concordants. Que les écarts aient ou non des conséquences importantes (pour ce qui est de la variance, du biais, de questions pratiques telles que la convergence et la transparence) dépend de la situation. Le présent article arrivera peut-être à mieux faire comprendre ce qui distingue deux courants de pensée qui ont guidé les chercheurs spécialisés en échantillonnage.

Bibliographie

Alexander, C.H. (1987). Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages. *Techniques d'enquête*, 13, 193-209.

Andersson, P.G., et Thorburn, D. (2005). Une distance de calage optimale menant à un estimateur par la régression optimal. *Techniques d'enquête*, 31, 103-107.

Ardilly, P. (2006). *Les techniques de sondage*. Paris : Éditions Technip.

Bankier, M.D., Rathwell, S. et Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. Document de travail, Census Operations Section, Social Surveys Methods Division, Statistique Canada.

Bankier, M., Houle, A.M. et Luc, M. (1997). Calibration estimation in the 1991 and 1996 Canadian censuses. *Proceedings*, *Section on Survey Research Methods*, American Statistical Association, 66-75.

Beaumont, J.-F. (2005a). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society* B, 67, 445-458.

Beaumont, J.-F. (2005b). L'utilisation de renseignements sur le processus de collecte des données pour traiter la non-réponse totale au moyen de l'ajustement de poids. *Techniques d'enquête*, 31, 249-254.

Beaumont, J.-F., et Alavi, A. (2004). Estimation robuste par la régression généralisée. *Techniques d'enquête*, 30, 217-231.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

Bethlehem, J.G., et Keller, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

Breidt, F.J., et Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28, 1026-1053.

Breidt, F.J., Claeskens, G. et Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92, 831-846.

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

Chambers, R.L., Dorfman, A.H. et Wehrly, T.E. (1993). Bias robust estimation in finite populations nonparametric calibration. *Journal of the American Statistical Association*, 88, 268-277.

Deming, W.E. (1943). *Statistical Adjustment of Data*. New York : John Wiley & Sons, Inc.

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.-C. (1998). La correction de la nonréponse par calage ou par échantillonnage équilibré. Article présenté aux Congrès de l'ACFAS, Sherbrooke, Québec.

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. *Actes des Journées de Méthodologie*, I.N.S.E.E., Paris.

Deville, J.-C. (2004). Calage, calage généralisé et hypercalage. Document interne, I.N.S.E.E., Paris

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

Deville, J.-C., Särndal, C.-E. et Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Duchesne, P. (1999). Estimateurs de calage robustes. *Techniques d'enquête*, 25, 47-60.

Dupont, F. (1995). Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire. *Techniques d'enquête*, 21, 141-150.

Ekholm, A., et Laaksonen, S. (1991). Weighting via response modeling in the Finish Household Budget Survey. *Journal of Official Statistics*, 3, 325-337.

Estevao, V.M., et Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, 16, 379-399.

Estevao, V.M., et Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics*, 18, 233-255.

Estevao, V.M., et Särndal, C.-E. (2004). Borrowing strength is not the best technique within a wide class of design-consistent domain estimators. *Journal of Official Statistics*, 20, 645-660.

Estevao, V.M., et Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *Revue Internationale de Statistique*, 74, 127-147.

Firth, D., et Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society*, 60, 3-21.

Folsom, R.E., et Singh, A.C. (2000). The generalized exponential model for design weight calibration for extreme values, nonresponse and poststratification. *Proceedings*, *Section on Survey Research Methods*, American Statistical Association, 598-603.

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

Fuller, W.A., Loughin, M.M. et Baker, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20, 79-89.

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.

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

Harms, T. (2003). Extensions of the calibration approach: Calibration of distribution functions and its link to small area estimators. Chintex document de travail numéro 13, Federal Statistical Office, Germany.

Harms, T., et Duchesne, P. (2006). On calibration estimation for quantiles. *Techniques d'enquête*, 32, 37-52.

Hidiroglou, M.A. (2001). L'échantillonnage double. *Techniques d'enquête*, 27, 157–169.

Hidiroglou, M.A., et Särndal, C.-E. (1998). Emploi des données auxiliaires dans l'échantillonnage à deux phases. *Techniques d'enquête*, 24, 11-20.

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*, *Social Statistics Section*, American Statistical Association, 300-305.

Isaki, C.T., Tsay, J.H. et Fuller, W.A. (2004). Pondération de données d'échantillon reposant sur des contrôles indépendants. *Techniques d'enquête*, 30, 39-49.

Kalton, G., et Flores-Cervantes, I. (1998). Weighting methods. Dans *New Methods for Survey Research* (Éds. A.Westlake, J. Martin, M. Rigg et C. Skinner), Berkeley, U.K.: Association for Survey Computing.

Kim, J., Breidt, F.J. et Opsomer, J.D. (2005). Nonparametric regression estimation of finite population totals under two-stage sampling. Manuscript non publié.

Knottnerus, P., et van Duin, C. (2006). Variances in repeated weighting with an application to the Dutch Labour Force Survey. *Journal of Official Statistics*, 22, 565-584.

Kott, P.S. (2004). Commentaire sur Demnati et Rao : Estimateurs de la variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 29-30.

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 149-160.

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

Krapavickaitė, D., et Plikusas, A. (2005). Estimation of a ratio in the finite population. *Informatica*, 16, 347-364.

Lavallée, P. (2007). *Indirect Sampling*. New York: Springer-Verlag.

LeGuennec, J., et Sautory, O. (2002). CALMAR2 : une nouvelle version de la macro CALMAR de redressement d'échantillon par calage. *Actes des Journées de Méthodologie*, INSEE, Paris.

Lehtonen, R., et Veijanen, A. (1998). Estimateur de régression généralisés logistiques. *Techniques d'enquête*, 24, 53-58.

Lehtonen, R., Särndal, C.-E. et Veijanen, A. (2003). L'effet du choix d'un modèle dans l'estimation par domaine, dont les petits domaines. *Techniques d'enquête*, 29, 37-49.

Lehtonen, R., Särndal, C.-E. et Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7, 649-674.

Lemaître, G., et Dufour, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.

Lundström, S., et Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.

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

Montanari, G.E. (1987). Post-sampling efficient prediction in large-scale surveys. *Revue Internationale de Statistique*, 55, 191-202.

Montanari, G.E., et Ranalli, M.G. (2003). On calibration methods for design-based finite population inferences. Bulletin of the International Statistical Institute, 54^{e} session, volume LX, articles contribute, livre 2, 81-82.

Montanari, G.E., et Ranalli, M.G. (2005). Nonparametric model-calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100, 1429-1442.

Myrskylä, M. (2007). Generalised regression estimation for domain class frequencies. Statistics Finland Research Reports 247.

Nieuwenbroek, N.J. (1993). An integrated method for weighting characteristics of persons and households using the linear regression estimator. Rapport, Central Bureau of Statistics, Pays-Bas.

Nieuwenbroek, N.J., et Boonstra, H.J. (2002). Bascula 4.0 for weighting sample survey data with estimation of variances. The Survey Statistician, Software Reviews, Juillet 2002.

Nieuwenbroek, N.J., Renssen, R.H. et Hofman, L. (2000). Towards a generalized weighting system. Dans *Proceedings*, *Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria VA.

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

Plikusas, A. (2006). Non-linear calibration. *Proceedings, Workshop on Survey Sampling*, Venspils, Latvia. Riga: Central Statistical Bureau of Latvia.

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

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, 368-374.

Renssen, R.H., Kroese, A.H. et Willeboordse, A.J. (2001). Aligning estimates by repeated weighting. Rapport, Central Bureau of Statistics, Pays-Bas.

Rizzo, L., Kalton, G. et Brick, J.M. (1996). Comparaison de quelques méthodes de correction de la non-réponse d'un panel. *Techniques d'enquête*, 22, 43-53.

Rueda, M., Martínez, S., Martínez, H. et Arcos, A. (2007). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, 137, 435-448.

Särndal, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*, 7, 155-170.

Särndal, C.-E., et Swensson, B. (1987). A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse. *Revue Internationale de Statistique*, 55, 279-294.

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

Särndal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York : John Wiley & Sons, Inc.

Särndal, C.-E., et Lundström, S. (2007). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. Statistics Sweden : Research and development - methodology report 2007:2.

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, S., Horn. S. et Yu, F. (1998). Estimation de la variance de l'estimateur général de régression : approche de calage à niveau élevé. *Techniques d'enquête*, 24, 43-52.

Skinner, C. (1998). Calibration weighting and non-sampling errors. *Proceedings International Seminar on New Techniques for Statistics,* Sorrento, novembre 4-6, 1998, 55-62.

Statistique Canada (1998). Methodology of the Canadian Labour Force Survey. Statistique Canada, Division des méthodes d'enquêtes auprès des ménages. Ottawa : Minister of Industry, catalogue no. 71-526-XPB.

Statistique Canada (2003). Quality Guidelines (quatrième édition). Ottawa : Minister of Industry, numéro de catalogue 12-539-XIE.

Steel, D.G., et Clark, R.G. (2007). Estimation par la régression au niveau de la personne et au niveau du ménage dans les enquêtes ménages. *Techniques d'enquête*, 33, 59-69.

Stukel, D.M., Hidiroglou, M.A. et Särndal, C.-E. (1996). Estimation de la variance des estimateurs de calage : comparaison des méthodes du jackknife et de la linéarisation de Taylor. *Techniques d'enquête*, 22, 117-126.

Théberge, A. (1999). Extension of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.

Théberge, A. (2000). Calage et poids restreints. *Techniques d'enquête*, 26, 113-122.

Tillé, Y. (2002). Estimation sans biais par calage sur la répartition dans les plans simples sans remise. *Techniques d'enquête*, 28, 83-91.

Tracy, D.S., Singh, S. et Arnab, R. (2003). Note sur le calage sous échantillonnage stratifié et double. *Techniques d'enquête*, 29, 111-116.

Vanderhoeft, C. (2001). Generalised calibration at Statistics Belgium. SPSS Module g-CALIB-S and current practices. Statistics Belgium : Document de travail no. 3. Disponible à : www.statbel.fgov.be/studies/paper03_en.asp.

Vanderhoeft, C., Waeytens, E. et Museux, J.M. (2001). Generalised calibration with SPSS 9.0 for Windows baser. Dans *Enquêtes, Modèles et Applications* (Éds. J.J. Droesbeke et L. Lebart), Paris : Dunod.

Webber, M., Latouche, M. et Rancourt, E. (2000). Harmonised calibration of income statistics. Statistique Canada, document interne, avril 2000.

Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, 90, 937-951.

Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Zheng, H., et Little, R.J.A. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.

Zieschang, K.D. (1990). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2007
Vol. 33, N° 2, pp. 137-147
Statistique Canada, N° 12-001-X au catalogue

Pondération de données d'enquête recueillies en deux phases

Seppo Laaksonen [1]

Résumé

L'état de données manquantes peut prendre diverses formes. Dans cet exposé, nous nous attardons à la non-réponse des unités et tentons de corriger cet état par une pondération appropriée. Le cas empirique que nous présentons vise l'échantillonnage à deux phases. En première phase, on a fait enquête auprès d'un grand échantillon à l'aide d'un questionnaire plutôt général. Au terme de cette phase, l'intervieweur a demandé aux répondants s'ils désiraient participer à une seconde phase où, avec un questionnaire plus détaillé, on se concentrerait sur un certain nombre de thèmes venant de la première phase. Cette procédure crée trois mécanismes de données manquantes. La difficulté est de savoir comment pondérer le plus exactement possible les répondants de la seconde phase par souci de cohérence des données issues des deux phases. Nous analyserons d'abord les différences de données manquantes propres à un tel scénario en trois temps en nous reportant à des données d'enquête auprès des personnes et nous comparerons ensuite divers modes de pondération. Notre recommandation est d'utiliser toutes les données auxiliaires disponibles le mieux possible. On obtient un bon résultat en mêlant les deux méthodes classiques avec d'abord une pondération de propension à répondre et ensuite un calage sur la distribution connue de la population.

Mots clés : Calage; variables auxiliaires internes et externes; méthode de modélisation de la propension à répondre; sous-échantillonnage sélectif.

1. Introduction

L'enquête type se fait en une étape ou phase. Dans un tel cas, on aura sélectionné en premier lieu des unités en appliquant un certain plan d'échantillonnage et on aura tâché de prendre contact avec les personnes choisies et de les interviewer le mieux possible. Ce faisant, on aura créé une certaine non-réponse à un degré variable ou encore d'autres formes de données manquantes ou lacunaires. Pour corriger un état de données manquantes, on recourt habituellement à des méthodes postenquête de correction plus ou moins avancées qui font intervenir les variables auxiliaires disponibles à diverses sources (voir, par exemple, Laaksonen 1999, ou une version élargie dans Laaksonen 2006b, et Lundström et Särndal 2001). Il reste que les données servant à la pondération seront normalement tirées de registres ou d'autres sources administratives ou enquêtes. On pourra parler de variables auxiliaires *externes* à distinguer des variables *internes* de même nature; elles seront internes en ce sens que l'information viendra de la même enquête ou de celle qui l'a précédée.

Les variables auxiliaires internes servent en particulier à l'imputation si des valeurs manquent. Elles sont aussi beaucoup utilisées dans les enquêtes longitudinales s'il y a réponse à un cycle d'enquête et non-réponse à un autre. Dans ce cas, l'information auxiliaire interne peut servir à la fois à la repondération et à l'imputation.

Notre exposé ne porte pas sur une enquête type comme celle que nous venons d'évoquer. Il vise deux caractéristiques :

(i) Une enquête peut être à deux (ou à trois à certains égards) étapes ou phases. La première phase est celle de l'enquête type à laquelle répondent un certain nombre d'unités. En seconde phase, nous gardons seulement dans la base de sondage les répondants qui sont disposés à répondre à un questionnaire plus détaillé. Il faut d'abord distinguer les répondants en première phase qui disent être disposés à participer de ceux qui, dans ce même groupe, répondront effectivement au second questionnaire. Ce sous-groupe se trouvera donc à remplir les deux questionnaires.

(ii) Dans nos corrections après enquête, nous aurons la possibilité d'exploiter des variables auxiliaires tant externes qu'internes en seconde phase. Les variables internes viendront de la première enquête.

Notre propos est l'examen de la seule repondération, mais certaines des idées que nous avançons sont aussi applicables à l'imputation. Ce que nous décrirons n'a pas beaucoup été utilisé dans le cas des enquêtes transversales, mais le même problème s'est souvent posé. Il arrive fréquemment, par exemple, que l'intervieweur interroge directement l'enquêté en premier lieu et que, à la fin de cette interview initiale, il demande au répondant de remplir lui-même un questionnaire par la suite. Si l'intéressé est d'accord, l'intervieweur lui remet immédiatement le questionnaire à remplir ou l'envoie ultérieurement. Dans l'une ou l'autre de ces situations, il recevra les réponses par

1. Seppo Laaksonen, Département de mathématiques et de statistique, case postale 68, FIN-00014 Université de Helsinki, Finlande; courriel : Seppo.Laaksonen@ Helsinki.Fi.

la poste ou par courrier électronique. Un exemple récent en est l'Enquête sociale européenne où le questionnaire supplémentaire porte plus particulièrement sur les valeurs de la vie (voir www.europeansocialsurvey.com). Bien sûr, ce ne sont pas tous les répondants de la première interview qui vont remplir le questionnaire qui suit.

Les questions posées en seconde phase n'abordent pas nécessairement le même thème que celles du questionnaire de base. Une autre stratégie courante est de prendre d'abord un questionnaire général sur un thème et de revenir en seconde phase avec des questions plus détaillées sur le même thème. Il peut aussi y avoir une certaine rétroaction de la première phase pour le second questionnaire et même un échantillon qui dépend de la distribution des variables clés de la première enquête (c'est là un exemple d'échantillonnage adapté). On procède fréquemment de la sorte lorsqu'on n'a pas vraiment d'expérience avec ce type d'enquête. La première phase sert aussi d'enquête pilote. Avec ce qu'on appelle les échantillons-maîtres, on est proche de l'idée que l'enquête de première phase (ce qui peut comporter un échantillonnage à des sources administratives comme microrecensement dans certains pays) puisse servir à élaborer une base de sondage appropriée. Dans ce cas, les variables de l'échantillon-maître sont plutôt limitées, ne livrant habituellement que des données de fait ou des renseignements généraux.

Avec les échantillons-maîtres, on vise à ce que la base de sondage obtenue représente bien la population visée. En cas d'échantillonnage, la base de sondage pourrait procurer les données auxiliaires nécessaires à la vérification, à l'imputation et à la pondération pour l'enquête bien réelle qui doit suivre (enquête de seconde phase). Chaque enquête réelle exploite donc un sous-échantillon de l'échantillon-maître. Nous examinerons ici un cas plus complexe qu'illustre le tableau 1.

Tableau 1 Illustration de l'échantillon initial avec trois ensembles de données en suivi

Échantillon avec variables auxiliaires X (*sexe*, *âge*, *région et saison*) Poids *selon le plan* pour 12 554 unités sans surdénombrement	Répondants en première phase avec variables Y_1 (*santé*, *activités de plein air*, *etc.*) Poids *pour base* de 10 666 unités	Volontaires en seconde phase avec variables Y_1 Poids pour 8 481 unités	Répondants en seconde phase avec variables Y_2 (*patinage, nautisme*, *etc.*) Poids pour 5 480 unités

On applique d'abord une procédure type d'échantillonnage avec un certain nombre de variables auxiliaires X. En l'occurrence, on a obtenu un taux de réponse relativement élevé (10 666 unités sur 12 554, soit un pourcentage approximatif de 85 %) à la première enquête. On a observé une certaine déperdition d'effectif parce que les répondants n'étaient pas tous disposés à participer à la seconde enquête (il ne reste plus alors que 68 % de l'échantillon initial). Comme le taux de non-réponse est plutôt élevé en seconde phase (malgré le caractère volontaire de la participation des répondants), le sous-échantillon restant n'est plus que de 44 % de l'échantillon initial. Nous disposons des trois ensembles de données qui suivent à des fins d'analyse :

A. Répondants en première phase avec les variables d'enquête Y_1.
B. Répondants en seconde phase avec les variables d'enquête Y_2.
C. Répondants tant en première qu'en seconde phase avec les variables d'enquête Y_1 et Y_2.

La plupart des utilisateurs recevront deux fichiers, A et B, qu'ils pourront réunir en un fichier C s'ils peuvent trouver un identificateur commun. À quoi s'attend l'utilisateur au reçu des deux fichiers? Il s'attend, bien sûr, à ce que les estimations d'un même paramètre soient identiques dans les deux fichiers, c'est-à-dire que les résultats soient homogènes, sachant que l'estimation d'un paramètre par un fichier C de moindre taille est moins précise qu'une estimation par un fichier plus complet. En principe, il est possible d'imputer les valeurs manquantes des variables Y_2, mais nous pensons qu'il est impossible de bien le faire, aussi vaut-il mieux pour nous procéder par pondération, le but étant de trouver des corrections de pondération du fichier B pour la meilleure analyse possible sur les variables Y_1 et Y_2.

Plusieurs stratégies sont applicables en pareil cas. Les aspects généraux d'intérêt sont présentés entre autres par Kalton et Kasprzyk (1986), Little (1986), Särndal, Swensson et Wretman (1992), Fuller, Loughin et Baker (1994), Wu et Sitter (2001) et Lundström et Särndal (2001). Si nous posons que l'état de données manquantes ne dépend que du plan d'échantillonnage, nous pouvons pondérer les fichiers A et B en ce sens. En cas d'échantillonnage aléatoire stratifié par exemple, la même stratification vaudrait naturellement pour les deux phases. En cas de poststratification, une stratégie analogue pourrait s'employer.

Dans l'exemple précis de notre enquête, la base de sondage comprenait le premier groupe de renouvellement en 12 mois de l'EPA finlandaise. Chaque mois, l'échantillon est prélevé au hasard. Notre EPA se fait par échantillonnage aléatoire simple, mais à cause de la non-réponse, on repondère par une technique type de calage (Deville, Särndal et Sautory 1993) en prenant le *sexe*, la *tranche d'âge* (six catégories) et la *région* (cinq catégories) comme variables auxiliaires. C'est ce que nous allons appeler les poids de sondage dans la suite de notre exposé. Nous avons

établi en conséquence la pondération de base des répondants de la première phase en ajoutant la variable *saison* ($4 \times 2 = 8$ catégories sur deux ans) à nos variables auxiliaires. Dans les enquêtes réalisées en Finlande, on emploie rarement la variable « *saison* », mais le caractère « saisonnier » de notre enquête rendait cette utilisation nécessaire (voir la section 2). Nous n'examinerons pas cet aspect en détail. Les trois premières variables reviennent souvent dans nos enquêtes auprès des personnes, cette information se prêtant d'emblée à la validation par un registre de la population mis à jour. Cela étant, nous supposerons disposer des meilleures estimations possible pour les répondants de la première phase lorsque nous utilisons cette repondération par calage. De toute manière, nous n'avons pas d'autre accès à des sources externes d'information susceptibles de nous être utiles.

Il est possible de se reporter aux estimations obtenues des répondants de première phase pour le calage de seconde phase. C'est une stratégie qui n'est pas difficile à appliquer en soi, mais on devrait inclure dans cette application toutes les variables X et autant de variables Y_1 que possible. Il faut également trouver pour une telle stratégie des niveaux d'ensemble ou de domaine qui soient le plus précis possible, tâche ardue que nous n'entreprenons pas dans cette étude. Rien ne garantit non plus que les estimations d'autres agrégats seront suffisamment exemptes de biais (les résultats de Laaksonen (1999) livrent certaines indications à l'appui de cette conclusion).

La stratégie que je propose est plus simple et fonctionne sans problèmes techniques pour tous les domaines, bien qu'on n'ait nullement la garantie, bien sûr, qu'un éventuel biais sera largement réduit pour l'ensemble des domaines. Je n'ai donc opté pour aucune stratégie avancée de calage, bien que la chose soit praticable. J'espère que d'autres auteurs en montreront les avantages possibles. Une référence utile à ce propos est Dupont (1995), qui parle de calage de données d'enquête en deux phases, mais sans fournir d'information empirique à cette fin. À noter donc que je procède par calage, mais sans prendre quelque chose de très avancé (voir la section 3).

La méthode que je propose fait largement appel à une modélisation de propension à répondre qui a été employée avec succès dans d'autres cas (voir, par exemple, Ekholm et Laaksonen (1991), Laaksonen (1999), Duncan et Stasny (2001) et Laaksonen et Chambers (2006)). Le cas de Rizzo, Kalton et Brick (1996) rappelle assez notre enquête en deux phases, mais il s'agit d'une enquête longitudinale. Le cadre méthodologique appliqué par ces auteurs a un certain nombre de caractéristiques semblables. Une grande différence est le mécanisme de réponse qui intervient ici en deux phases (participation volontaire et réponse à la deuxième phase). Nous analysons ces stades séparément nous aussi. Bien entendu, nous comparons les résultats obtenus par diverses techniques. À la section 2, nous décrivons sommairement nos questionnaires et nos ensembles de données; à la section 3, nous détaillons les principes qui sous-tendent nos méthodes; à la section 4, nous présentons des données de comparaison et, à la section 5, nous tirons nos conclusions.

2. Principes sous-tendant les ensembles de données

Les données viennent d'une enquête spéciale menée auprès de Finlandais de 15 à 74 ans (pour plus de renseignements à ce sujet, voir Virtanen, Pouta, Sievänen et Laaksonen 2001). Le thème abordé est celui des activités en période de loisirs et notamment des activités et passe-temps de plein air. Il y a d'abord eu une interview sur place assistée par ordinateur (IPAO) où on a posé des questions diverses sur les loisirs et les hobbies des gens (cyclisme, motocyclisme, randonnée pédestre, jogging, nautisme, natation, chasse, pêche, photographie d'observation de la nature, ski, patinage, équitation, *etc.*). Dans tous les cas, la période de référence était l'année précédente. À la fin de cette enquête, on a demandé aux répondants s'ils seraient disposés à recevoir un questionnaire spécial d'enquête postale qui leur permettrait de décrire plus en détail certaines de ces activités. L'enquête en question aurait lieu quelques semaines après.

La période d'enquête a été de deux ans (1998-2000), le but étant d'alléger le fardeau de réponse et les charges d'interview. Autre facteur : les activités visées étaient saisonnières dans une certaine mesure, et on pouvait s'attendre à ce que les réponses subissent l'influence des saisons (on ne répond pas, par exemple, de la même manière à des questions sur le ski si on est en été ou en hiver). Après correction de surdénombrement (104 unités), l'échantillon initial s'est établi à 12 554 personnes.

Nous avons choisi les variables binaires suivantes pour l'analyse présentée à la section 4 : *activités de plein air* (l'enquêté exerce régulièrement certaines de ces activités dans la nature), *santé* (l'enquêté est-il assez en santé pour mener de telles activités?), *ski*, *pêche*, *patinage*, *nautisme*, *cyclisme* et *jogging*. Dans tous les cas, nous attribuons la valeur 1 si on a exercé l'activité en question l'année précédente et la valeur 0 si on ne l'a pas fait. Toutes ces variables étaient comprises dans le questionnaire de première phase; nous savions donc à quoi nous attendre après les deux phases. Précisons que notre ensemble de données comportait des variables plus complexes, mais que nous avons jugé bon de simplifier pour que les résultats soient d'une interprétation plus facile. Les grandes conclusions seraient les mêmes si notre choix avait été autre.

À la section 4, nous faisons deux types de comparaisons (i) en fonction de l'information connue de la première phase et (ii) sans nous reporter à l'information connue. Dans l'un et l'autre cas, nous pouvons heureusement fort bien savoir si nous avons réussi à réduire le biais, puisque nous disposons des estimations « vraies » (ou des meilleures estimations possible). Ajoutons que nous analysons un certain nombre de variables qui figurent seulement dans le second questionnaire, mais qu'il est impossible d'établir le degré d'efficacité de chaque méthode dans ce cas. Nous ne présentons pas non plus les derniers résultats en détail, mais on sait par observation que le tout se comporte comme dans le deuxième cas.

3. Modélisation de la propension à répondre : méthode et calage

Notre étude comprend trois phases avec les spécifications suivantes de pondération :

D'abord, on met au point une bonne pondération calibrée pour les répondants de première phase avec les variables *région*, *sexe*, *tranche d'âge* et *saison* (voir aussi la section 1). Le symbole w_k désignera le poids de sondage du répondant k. Ces poids sont basés sur le calage et sont aussi appelés poids « de base ». À noter que, préalablement, il y a eu des poids de sondage de créés pour l'ensemble de données en fonction d'un plan d'échantillonnage aléatoire stratifié. Nous disposons de telles valeurs des poids de sondage pour les non-répondants en première phase.

En second lieu, nous modélisons les probabilités de participation volontaire et de réponse à l'aide de la fonction de combinaison la plus courante (la fonction logit n'est pas nécessairement la meilleure de ces fonctions, nous le savons depuis (2006a), mais c'est elle que nous utilisons ici). Dans ce cas, $\text{logit} = \log(\pi/(1-\pi))$, où π est la probabilité binaire de réponse (1 = participation volontaire et 0 = non-participation volontaire ou 1 = réponse et 0 = non-réponse). Les variables explicatives employées sont les variables X et un certain nombre de variables Y_1 considérées comme « bonnes ». Ce modèle donne les probabilités prévues de réponse p_k qui servent de la manière suivante à la repondération dans chaque cas :

$$w_k(\text{res}) = \frac{w_k}{p_k} g_c.$$

Dans ce cas, $g_c = a$ est un facteur d'échelle qui étalonne en fonction de certains agrégats connus au niveau c. Il existe plusieurs solutions de rechange à ce mode de calage, mais ce que nous décrivons pourrait être considéré comme un calage type. Dans notre étude, nous utilisons les agrégats h en poststratification (c'est la cellule de recoupement des trois variables X de la *tranche d'âge*, du *sexe* et de la *région*; le nombre total de cellules est de 6*2*5 = 60). La technique est simple :

$$g_h = \frac{\sum_h w_k}{\sum_h w_k / p_k}.$$

Comme nous l'avons mentionné, on obtient en Finlande une haute qualité pour ce genre d'agrégats de poststratification, mais non pas nécessairement pour les autres.

Comme la seconde phase est double, nous avons les trois possibilités suivantes de modélisation qui sont toutes exploitées à la section 4 :

a) modèle relatif à la participation volontaire;
b) modèle relatif à la réponse des personnes qui se sont portées volontaires (modélisation « *deux étapes* »);
c) modèle de réponse en une étape (modélisation « *directe* » ou « *une étape* »).

À noter que, combinées, les phases a) et b) donnent les valeurs de pondération du fichier B, d'où la formulation suivante (vol = participation volontaire; p_1 = probabilité estimée de réponse à la phase 1; p_2 = probabilité estimée de réponse à la phase 2; facteurs d'échelle g_1 et g_2 respectivement) :

Phase 1 : $$w_k(\text{vol}) = \frac{w_k}{p_{1k}} g_{1h},$$

Phase 2 : $$w_k(\text{res}) = \frac{w_k(\text{vol})}{p_{2k}} g_{2h}.$$

Il a fallu utiliser les bons poids de sondage pour chaque tâche de modélisation. Pour les modèles a) et c), ce sont les poids w_k, mais pour le modèle b), les poids $w_k(\text{vol})$. Dans nos tests de comparaison, nous avons également modélisé la réponse en première phase et utilisé les poids de sondage. En pondérant dans la modélisation, nous obtenons des estimations plus fidèles, puisque nous tâchons de rendre notre analyse représentative de la population cible. Parfois, la pondération est d'une grande incidence comme dans les enquêtes auprès des entreprises où, souvent, les valeurs de pondération sont plus variables que dans les enquêtes types auprès des ménages. Dans le cas qui nous occupe, les résultats ne différaient pas outre mesure selon que les modèles étaient en pondération ou non, bien qu'une pondération s'impose en pareil cas (c'est ce que démontrent nettement Laaksonen et Chambers (2006) où l'incidence de la pondération est appréciable; Rizzo et coll. (1996) emploient aussi la pondération). Nous présenterons plus loin à la section 4 les données empiriques (estimations avec leurs erreurs-types et les probabilités de réponse) pour les solutions en pondération.

Il n'y a pas que les techniques clés dont nous venons de parler; à la prochaine section, nous livrons aussi en pondération w_k nos *meilleures estimations possible* pour les paramètres qui sont connus (variables Y_1).

De plus, nous mettons nos résultats bien précis en comparaison par rapport au seul calage de poststratification sans modélisation (symbole « *cal* »' dans les sections qui restent de notre exposé). Il s'agit là d'une manière tout à fait normale d'aborder le problème de la pondération (c'était le « style maison » avant cette proposition méthodologique). Il convient toutefois de noter que, si un modèle de réponse comprend seulement les variables (et les mêmes catégories) de poststratification, la pondération en modélisation de la propension à répondre correspond parfaitement à celle que l'on obtient par la technique de poststratification.

4. Résultats empiriques

Cette section décrit les résultats de l'application de différentes méthodes. Nous exposons les résultats de divers modèles de réponse, puis comparons les valeurs de pondération entre elles pour enfin confronter un certain nombre d'estimations des paramètres issues des différentes techniques.

4.1 Modélisation de la participation volontaire et de la réponse

Pour comprendre à fond le comportement de données manquantes (par absence de réponse et de participation volontaire) aux trois stades dégagés pour l'enquête, nous présentons au tableau 2 des résultats fondés sur les variables auxiliaires X disponibles à chaque phase (dans la pratique, nous employons aussi la variable « saison », mais sans tenir compte de ses effets dans cette analyse, puisqu'il ne s'agit pas là d'un sujet d'intérêt premier pour notre propos).

Tableau 2
Régressions logistique par les trois variables explicatives communes des trois phases, c'est-à-dire de la réponse en première phase et de la participation volontaire et de la réponse réelle en seconde phase. Nous présentons les estimations sous forme de rapports de probabilité; les intervalles de confiance à 95 % figurent entre parenthèses

Variables explicatives	Modèle 1 Réponse en première phase	Modèle 2a Participation volontaire	Modèle 3a Réponse des volontaires	Modèle 4a Réponse en seconde phase
Sexe (réf. : sexe féminin)				
Hommes	0,71 (0,65; 0,78)	0,84 (0,76; 0,93)	0,75 (0,68; 0,83)	0,77 (0,71; 0,83)
Tranche d'âge (réf. : 65 ans et plus)				
24 ans ou moins	1,00 (0,83; 1,21)	5,57 (4,65; 6,68)	0,51 (0,41; 0,64)	1,49 (1,28; 1,73)
25-34	0,96 (0,79; 1,15)	4,76 (4,00; 4,81)	0,65 (0,52; 0,81)	1,73 (1,49; 2,00)
35-44	0,85 (0,71; 1,02)	4,08 (3,46; 4,81)	0,64 (0,52; 0,80)	1,62 (1,40; 1,88)
45-54	0,89 (0,74; 1,07)	3,16 (2,71; 3,69)	0,86 (0,69; 1,06)	1,82 (1,58; 2,10)
55-64	1,18 (0,96; 1,45)	2,05 (1,74; 2,41)	1,15 (0,90; 1,47)	1,75 (1,49; 2,04)
Région (réf. : nord)				
sud-est	0,55 (0,46; 0,66)	2,12 (1,79; 2,50)	0,96 (0,79; 1,16)	1,35 (1,17; 1,55)
sud-ouest	0,76 (0,64; 0,91)	1,83 (1,57; 2,14)	1,04 (0,86; 1,25)	1,35 (1,18; 1,55)
ouest central	1,14 (0,91; 1,42)	2,14 (1,77; 2,59)	1,16 (0,93; 1,43)	1,56 (1,33; 1,83)
est central	0,96 (0,78; 1,18)	1,20 (1,01; 1,44)	1,15 (0,92; 1,43)	1,19 (1,02; 1,40)
Nombre d'observations	12 554	10 666	8 481	10 666
-2 log L	10 904	10 296	8 569	14 618

Nombre de résultats intéressants sont à relever dans ces modèles de comportement échelonné de données manquantes. Les résultats de la première enquête sont plutôt normaux; par exemple, les hommes répondent plus médiocrement que les femmes aux deux phases. La propension à répondre est également plus faible dans le sud que dans le reste du pays. Les différences entre tranches d'âge sont quelque peu étonnantes, puisque les groupes d'âge intermédiaire sont ceux qui répondent le moins bien.

Les estimations de la participation volontaire sont différentes. Dans l'est central et le nord, les gens sont le moins disposés à participer à la seconde phase, mais les résultats de réponse une fois que les gens se sont portés volontaires ne diffèrent pas nettement. Selon l'âge, il semblerait que les gens plus jeunes sont plus disposés à participer, mais sans très bien répondre par la suite. Ainsi, les gens plus âgés seraient en ce sens plus prêts à s'engager à participer, mais on peut clairement voir que les plus âgés seront sous-représentés hors de toute correction.

Si nous considérons les deux premières variables auxiliaires internes (tableau 3), nous pouvons observer que les gens qui ne sont pas relativement en santé (variable *santé*) et qui ne s'adonnent pas activement aux loisirs en milieu naturel (*activités de plein air*), ne sont pas disposés non plus à recevoir tout nouveau questionnaire, à en juger par le rapport de probabilité très élevé dans ce cas. Fait intéressant, les rapports de probabilité sont proches pour la variable *santé* et la participation volontaire. En d'autres termes, la personne qui n'est pas en santé ne sera pas très encline à se porter volontaire, mais si elle le fait, sa réponse ressemblera à celle de la personne en santé. Une tendance semblable se dégage en ce qui concerne la variable *activités de plein air*. Il convient de noter que les domaines « non en santé » et « non en plein air » ne sont pas très étendus et que, bien que jouant un grand rôle dans la modélisation de la propension à répondre, leur incidence sur les estimations obtenues à la fin n'est pas considérable (section 4.3).

Si on ajoute les deux autres variables auxiliaires internes, celles du *ski* et de la *pêche*, la même sélectivité est constatée, mais elle n'est pas aussi appréciable. En conclusion, on discerne clairement que le mécanisme de réponse à la seconde enquête serait très peu « non informatif ». Il faut s'attendre à certains effets, par conséquent, sur la repondération et les estimations d'enquête, ce que nous examinerons dans les deux sous-sections qui suivent.

4.2 Comparaison entre les poids

Comme nous l'avons exposé, nous fournissons plusieurs poids. Le tableau 4 les récapitule avec les statistiques descriptives de manière à expliquer les changements consécutifs à chaque correction de pondération. Les poids de sondage ne peuvent intervenir dans nos comparaisons, car nous ne disposons pas de données sur les variables Y pour l'échantillon initial. On peut cependant voir que c'est pour cette pondération que la variation relative mesurée ici est la plus faible ($1 + cv^2$, cv étant le coefficient de variation). Cette formule sert également d'approximation de l'effet de plan. Rizzo et coll. (1996) se servent aussi de cet indicateur lorsqu'ils comparent les pondérations.

Les changements ne sont pas considérables pour la première phase, c'est-à-dire lorsqu'on passe du poids de sondage au poids de base (sauf pour la moyenne liée à une taille d'échantillon qui décroît), mais aux deux phases qui suivent, l'effet de plan est plus marqué. On constate en outre que la variation est moindre pour les deux pondérations par calage que pour la pondération basée sur la modélisation de propension à répondre. Pour chaque pondération, la distribution est asymétrique à droite; elle l'est le moins pour les poids de sondage, bien sûr. On s'étonnera quelque peu que l'asymétrie soit la plus grande pour la pondération de la participation volontaire. On trouvera aux figures 1 à 3 plus de détails sur les distributions en question et les différences entre les pondérations.

La figure 1 illustre bien que certaines valeurs de pondération s'accroissent nettement à cause de la modélisation de propension à répondre (modèle 2b). Il est possible de voir en détail quels types d'unités sont sous les courbes à valeurs de pondération largement en croissance. Ainsi, derrière les courbes du côté gauche où les valeurs sont de plus de 700 pour la pondération par la propension à répondre, on trouve des gens qui ne sont pas en santé et qui ne raffolent pas des activités de plein air, mais qui n'en figurent pas moins au fichier de la participation volontaire. De même, on découvre d'autres groupes intéressants en se reportant aux résultats des estimations par modélisation. La majeure partie des courbes sont cependant concentrées et, par conséquent, on peut s'attendre à moins de changements dans les estimations que là où la pondération varierait plus amplement.

La dispersion est quelque peu plus grande à la figure 2 qu'à la figure 1, mais le profil demeure le même. On peut donc trouver des sous-groupes intéressants derrière des courbes distinctes.

À la figure 3 enfin, nous comparons les deux pondérations de seconde phase. Ce diagramme de dispersion diffère considérablement des deux précédents, car la relation est plutôt linéaire. Les valeurs maximales sont plus élevées pour la pondération « deux étapes » que pour la pondération « une étape », mais les valeurs de pondération « une étape » sont dans bien des cas nettement supérieures. Ainsi, les gens qui ne sont pas en santé et qui se livrent cependant à des activités de plein air reçoivent une valeur « une étape » relativement élevée, mais il n'y a pas d'effet net de l'âge. Par ailleurs, les membres des tranches d'âge supérieures qui

n'ont guère d'activités de plein air reçoivent des valeurs « deux étapes » relativement hautes, mais la santé n'a rien à y voir. On ne s'attend toutefois pas à relever de grandes différences dans les estimations respectives « une étape » et « deux étapes », mais une de ces pondérations aurait alors été à introduire. Avec un choix plus simple, celui d'une pondération « une étape », il serait toujours bon d'analyser séparément les deux pondérations avec leurs propensions respectives à répondre de manière à mieux comprendre les causes des deux types d'état de données manquantes.

Tableau 3
Régressions logistique à l'aide de certaines variables auxiliaires de la réponse en première étape qui s'ajoutent à celles du tableau 2 (les numéros de modèle aux tableaux 3 et 2 se correspondent et la variable de réponse et les ensembles de données sont les mêmes)

Variables explicatives et autres statistiques	Modèle 2b Participation volontaire	Modèle 3b Réponse des volontaires	Modèle 4b Réponse en seconde phase	Modèle 4c Réponse en seconde phase
Sexe (réf. : sexe féminin)				
Hommes	0,94 (0,85, 1,04)	0,77 (0,69; 0,85)	0,82 (0,75; 0,88)	0,75 (0,68; 0,83)
Tranche d'âge (réf. : 65 ans et plus)				
24 ans ou moins	4,92 (4,07; 5,97)	0,52 (0,41; 0,65)	1,30 (1,12; 1,52)	0,51 (0,41; 0,64)
25-34	3,83 (3,18; 4,60)	0,65 (0,52; 0,81)	1,46 (1,25; 1,70)	0,65 (0,52; 0,81)
35-44	3,26 (2,74; 3,88)	0,64 (0,51; 0,80)	1,37 (1,18; 1,58)	0,64 (0,52; 0,80)
45-54	2,59 (2,20; 3,05)	0,85 (0,68; 1,06)	1,56 (1,34; 1,81)	0,86 (0,69; 1,06)
55-64	1,73 (1,45; 2,05)	1,18 (0,89; 1,46)	1,55 (1,32; 1,81)	1,15 (0,90; 1,47)
Région (réf. : nord)				
sud-est	2,15 (1,81; 2,55)	0,96 (0,79; 1,16)	1,34 (1,16; 1,54)	0,96 (0,79; 1,16)
sud-ouest	1,92 (1,64; 2,26)	1,04 (0,86; 1,25)	1,36 (1,19; 1,56)	1,04 (0,86; 1,25)
ouest central	2,09 (1,71; 2,54)	1,15 (0,93; 1,43)	1,52 (1,29; 1,78)	1,16 (0,93; 1,43)
est central	1,17 (0,98; 1,41)	1,15 (0,91; 1,42)	1,18 (1,00; 1,38)	1,15 (0,92; 1,43)
Activités de plein air	3,04 (3,43; 2,71)	1,24 (1,43; 1,07)	1,93 (2,15; 1,74)	1,77 (1,97; 1,59)
Santé	3,61 (4,61; 2,82)	1,02 (1,58; 0,66)	2,71 (3,51; 2,09)	2,49 (3,24; 1,92)
Ski				1,36 (1,47; 1,25)
Pêche				1,27 (1,38; 1,17)
Nombre d'observations	10 666	8 481	10 666	10 666
-2 log L	9 721	8 560	14 342	14 244

Tableau 4 Statistiques descriptives de différentes valeurs de pondération d'échantillonnage (*PR* = propension à répondre)

Pondération	Phase	Nombre d'unités	Moyenne	Asymétrie	$1 + cv^2$
Poids de sondage	Zéro	12 658	308	0,94	1,30
Poids de base par calage	Premier	10 666	365	1,30	1,39
Poids en calage	Volontaires	8 481	460	2,52	1,63
Pondération de propension à répondre (PR), modèle 2b	Volontaires	8 481	460	4,60	1,82
Poids par calage	Deuxième	5 480	712	1,64	1,62
Pondération PR « deux étapes », modèles 2b et 3b	Deuxième	5 480	712	3,60	1,84
Pondération PR « une étape », modèle 4b	Deuxième	5 480	712	2,56	1,80

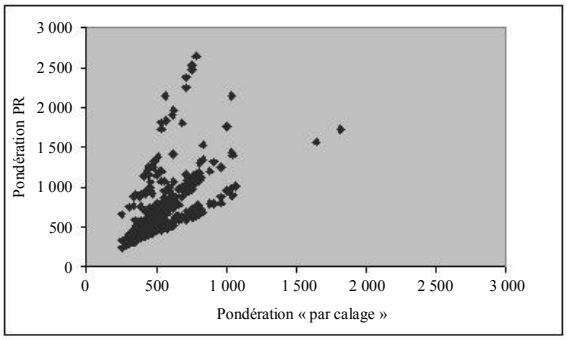


Figure 1 Diagramme de dispersion entre les deux pondérations de la participation volontaire

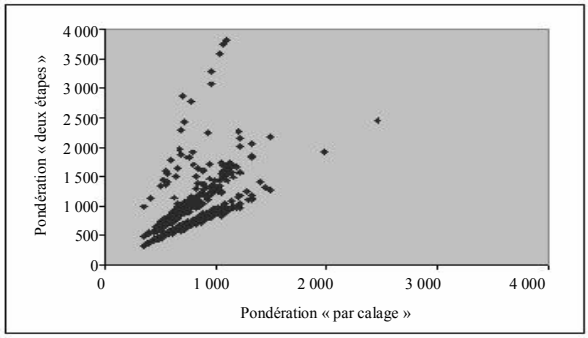


Figure 2 Diagramme de dispersion de la réponse en seconde phase entre la pondération par calage et la pondération « deux étapes » de la propension à répondre

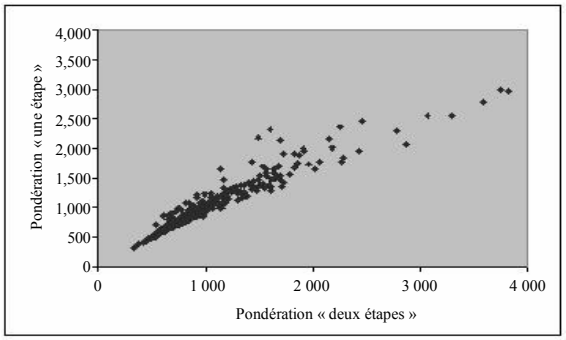


Figure 3 Diagramme de dispersion de la réponse en seconde phase entre les deux pondérations en modélisation de propension à répondre

4.3 Comparaison des estimations des paramètres

Nous n'avons pu faire d'études complètes de simulation avec des hypothèses variées et ainsi analyser quel type de méthode conviendrait le mieux dans chaque cas. Nous pouvons heureusement faire un grand pas dans cette direction en comparant les effets sur les estimations dans trois perspectives. Premièrement, nous modélisons la réponse et la participation volontaire en utilisant à la fois les variables X et un certain nombre de variables Y_1. Nous connaissons donc les meilleures estimations des paramètres par ces valeurs Y_1 de la première enquête. Deuxièmement, nous ajoutons des variables auxiliaires Y_1, mais en excluant certaines. Nous connaissons néanmoins les « meilleures » valeurs dans ce cas et pouvons donc faire des comparaisons précises. Troisièmement, nous pouvons comparer des estimations sans information connue et, dans ce dernier cas, il est seulement possible de déduire les valeurs qui pourraient être les meilleures.

Nous présentons les données explicites selon les variables décrites à la section 2. À noter que nous n'avons pas jugé important de présenter les erreurs-types des diverses estimations, parce que nous nous concentrons sur les biais de ces dernières. On notera cependant que les erreurs-types se situent dans l'intervalle 0,2-0,4 point pour l'ensemble de données première phase et dans l'intervalle 0,3-0,5 point pour l'ensemble de données seconde phase (les valeurs sont les plus basses pour *santé*, deuxièmes pour *jogging* et les plus hautes pour *activités de plein air* et *ski*).

La figure 4 livre les résultats avec une pondération sans correction aucune (tel est souvent le cas dans la pratique malheureusement). Nous pouvons voir que le biais est appréciable dans la plupart des estimations; il est le plus faible pour *jogging,* qui n'est pas une catégorie de forte activité si on la compare à la variable *activités de plein air,* par exemple. En règle générale, la plupart des utilisateurs n'aiment pas voir de tels biais importants et hautement significatifs statistiquement sauf pour *jogging* en seconde phase (l'intervalle de confiance à 95 % va de 1,7 à 2,3 points pour le biais de la variable *santé*). Dans ces résultats comme dans ceux qui suivent, biais veut dire *surestimation* : plus s'accentue l'état de données manquantes, plus les estimations augmentent et deviennent excessives. Faute de corrections appropriées, les résultats pécheront donc par optimisme, c'est-à-dire que les gens sembleraient faire trop de tous les exercices. À noter que la même tendance ressort dans les estimations de la première phase, mais nous ne pouvons l'expliquer d'une manière satisfaisante. Des différences étonnantes subsistent entre les deux jeux d'estimations; le biais sera supérieur tantôt pour les données « participation volontaire », tantôt pour les données « réponse en seconde phase ». Nous n'interprétons pas ces écarts en détail, mais il faut, bien sûr, y voir des différences d'état de données manquantes, d'où la possibilité de mises en garde à l'intention de l'utilisateur des données.

À des fins de comparaison, nous présentons à la figure 5 les mêmes résultats non corrigés pour la participation volontaire qu'à la figure 4, mais nous ajoutons les estimations correspondantes par calage de poststratification et pour modélisation de propension à répondre. Ce graphique indique clairement que la poststratification est source d'un certain avantage par rapport à toute solution hors correction. La méthode de la propension à répondre demeure la meilleure dans tous les cas cependant. Elle est extrêmement bonne pour les variables *santé* et *activités de plein air* qui

servent de variables auxiliaires dans les modèles qui la supportent.

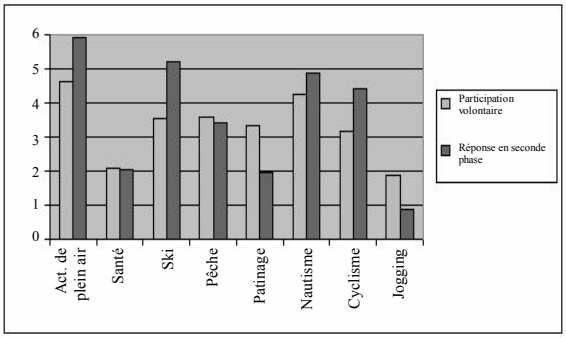


Figure 4 Biais des estimations en points de pourcentage en fonction de la pondération d'échantillonnage sans correction pour la participation volontaire et la réponse en seconde phase

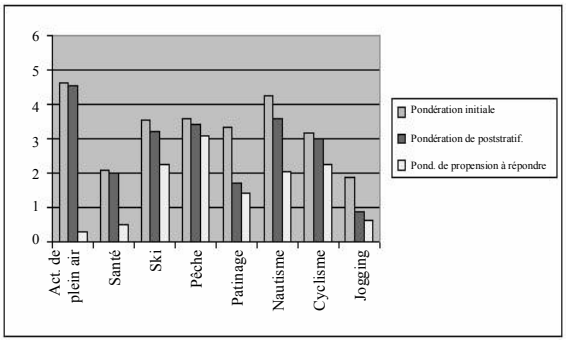


Figure 5 Biais des estimations en points de pourcentage de la participation volontaire en fonction de la pondération d'échantillonnage sans correction (symbole = « initial »), du calage de post-stratification et de la modélisation de propension à répondre où on emploie les variables auxiliaires *activités de plein air* et *santé* (modèle 2b au tableau 3)

Les figures 6 et 7 visent les estimations de la dernière phase et sont donc les plus importantes. La figure 6 dégage la même conclusion que la figure 5, en ce sens que la technique de modélisation de propension à répondre se révèle supérieure à la technique de calage par poststratification, bien que toutes les différences ne soient pas hautement significatives statistiquement (en particulier pour la variable *jogging*). La différence entre la méthode « une étape » et la méthode « deux étapes » est plutôt ténue et le biais change d'une variable à l'autre. D'après cette étude, il est donc impossible de dire laquelle de ces deux spécifications est meilleure.

La figure 7 fait un certain nombre de comparaisons lorsqu'on ajoute deux variables au modèle de la propension à répondre. Les résultats sont tout à fait prévisibles, puisque le biais de ces estimations se trouve réduit, tout comme le biais de toutes les autres estimations dans une certaine mesure. Aux yeux de bien des utilisateurs, le biais demeure excessif pour les variables *nautisme* et *cyclisme*. Nous pouvons le réduire, bien entendu, en ajoutant des variables auxiliaires au modèle. Jusqu'où pourrions-nous aller dans cette direction? Nous n'examinons pas la question plus avant. Par ailleurs, nos outils sont encore pires s'il s'agit de réduire le biais de telles variables en fonction de la seule seconde phase. Nous avons vérifié plusieurs de ces estimations et nous observons certains changements dans les estimations correspondantes, et ce, au même niveau que pour les variables *nautisme* et *cyclisme* à la figure 7. Dans ce cas, il est cependant impossible de vérifier le biais. Nous pouvons uniquement supposer, en nous fondant sur les exercices précédents, que les résultats sont moins biaisés que ceux qui correspondent à des valeurs encore moins bien corrigées.

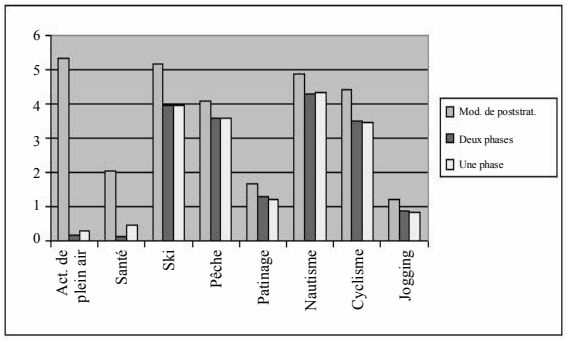


Figure 6 Biais des estimations en points de pourcentage de la réponse après les deux phases en fonction de la pondération de poststratification et des deux méthodes de la propension à répondre (pondération « deux étapes » et « une étape ») où les variables auxiliaires sont *activités de plein air* et *santé* (la méthode « deux étapes » est fondée sur les deux modèles échelonnés 2b et 3b au tableau 3 et la méthode « une étape » est directement fonction de la réponse en seconde phases (modèle 4b au tableau 3))

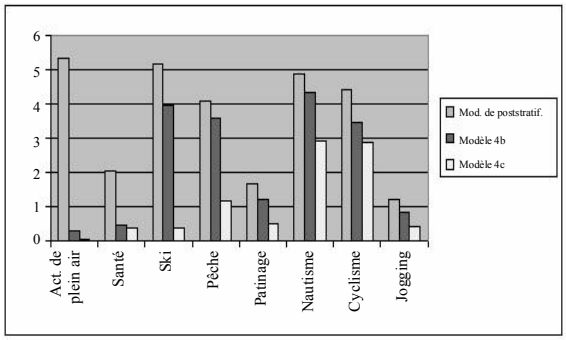


Figure 7 Biais des estimations en points de pourcentage de la réponse après les deux phases en fonction de la pondération de poststratification et des deux méthodes « une étape » de la propension à répondre où les variables auxiliaires sont *ski* et *pêche* (modèle 4c du tableau 3 en comparaison avec le modèle 4b)

5. Examen

Le problème dont traite le présent document est fréquent dans les enquêtes. Un grand nombre d'enquêtes se déroulent par phases et des incohérences s'y présentent à cause des données manquantes et d'autres divergences. Un exemple bien connu sur le plan international est l'Enquête sociale européenne (ESE) qui comporte deux questionnaires, à savoir un questionnaire de base et un questionnaire supplémentaire. Il va sans dire que le nombre de répondants est moindre pour le second de ces questionnaires que pour le premier, d'où une certaine sélectivité, par exemple, si le second questionnaire est en corrélation positive avec l'action politique. La situation sera étrange aux yeux de l'utilisateur lorsqu'une estimation varie selon qu'elle vient d'un grand ou d'un petit ensemble de données, bien que portant sur la même variable et la même période.

Comme pour l'ESE, notre étude vise des données d'enquête recueillies en deux phases. Le taux de réponse à la seconde enquête a été bien moindre que dans l'ESE. L'effet de sélectivité est également plus marqué. En appliquant les modèles de la propension à répondre, nous avons pu prévoir cette sélectivité, exploiter les résultats dans des corrections de pondération et, en dernier lieu, calibrer la somme des poids sur certains sommations agrégats connus de la population.

Cette stratégie vise à tirer tout le parti voulu de l'information auxiliaire disponible tant dans les registres qu'à d'autres sources extérieures, de même qu'à la première phase de l'enquête au niveau des microdonnées.

Dans notre exemple, la seconde phase de l'enquête est double, mais il n'existe qu'une collecte de données. Il y a d'abord le fait de se porter volontaire pour la seconde phase et ensuite la participation effective à l'enquête. Nous avons examiné les deux phases séparément pour trouver des données intéressantes sur les mécanismes respectifs de réponse. De plus, les résultats de cette analyse ont servi à la repondération. À des fins de comparaison, nous avons regardé les deux phases à la fois, élaboré un modèle applicable et continué la repondération d'une manière analogue. Enfin, nous avons comparé les estimations. Il était quelque peu étonnant que les deux ensembles de résultats diffèrent si peu dans nos exemples. C'est une bonne nouvelle cependant, puisqu'il est plus facile de travailler avec une pondération « une étape », laquelle pourrait alors être introduite.

Nous proposons un certain cadre méthodologique pour la pondération « deux étapes », mais sans pouvoir affirmer quelle spécification serait la meilleure dans chaque cas. Notre méthode est d'une application plutôt facile, mais les avantages qu'elle offre dépendent dans une large mesure, bien sûr, de la disponibilité de bonnes données auxiliaires externes et internes. Sans variables auxiliaires directes, on ne sait au juste quelle sera la qualité des estimations après correction. D'après nos exemples, elles seraient d'emblée moins entachées d'un biais que les estimations initiales. La technique que nous recommandons paraît cependant quelque peu « prudente » et, par conséquent, toutes nos meilleures estimations corrigées dans l'analyse seront en légère surestimation, bien que les différences ne soient pas statistiquement significatives. Voilà une occasion intéressante pour de futures recherches à prévoir tout particulièrement parce que ce problème devient plus fréquent dans le monde des enquêtes. Un autre thème d'intérêt pour l'avenir est la question de savoir comment choisir optimalement les variables auxiliaires en cas d'enquête en deux phases.

Remerciements

L'auteur aimerait remercier le rédacteur en chef de Techniques d'enquête et les examinateurs anonymes de leurs observations précises et utiles.

Bibliographie

Deville, J.-C., Särndal, C.-E. et Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Duncan, K.B., et Stasny, E.A. (2001). Utilisation de scores de propension pour contrôler le biais de couverture dans les enquêtes téléphoniques. *Techniques d'enquête*, 27, 2, 131-141.

Dupont, F. (1995). Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire. *Techniques d'enquête*, 21, 141-150.

Ekholm, A., et Laaksonen, S. (1991). Weighting via response modelling in the Finnish household budget survey. *Journal of Official Statistics*, 7, 2, 325-337.

Fuller, W.A., Loughin, M.M. et Baker, H. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20, 79-89.

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

Laaksonen, S. (1999). Weighting and auxiliary variables in sample surveys. Dans "Enquêtes et Sondages. Méthodes, modèles, applications, nouvelles approches" (Éds., G. Brossier et A.-M. Dussaix). Dunod, Paris. 168-180.

Laaksonen, S. (2006a). Does the choice of link function matter in response propensity modelling? *Model Assisted Statistics and Applications*, *An International Journal*. Éditeur: IOS Press. 1, 95-100.

Laaksonen, S. (2006b). Need for high quality auxiliary data service for improving the quality of editing and imputation. Dans *United Nations Statistical Commission*, "*Statistical Data Editing*", 3, 334-344.

Laaksonen, S., et Chambers, R. (2006). Survey estimation under informative non-response with follow-up. *Journal of Official Statistics*, 81-95.

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

Lundström, S., et Särndal, C.-E. (2001). Estimation in the presence of nonresponse and frame imperfections. Statistics Sweden.

Rizzo, L., Kalton, G. et Brick, J.M. (1996). Comparaison de quelques méthodesde correction de la non-réponse d'un panel. *Techniques d'enquête*, 22, 43-53.

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

Virtanen, V., Pouta, E., Sievänen, T. et Laaksonen, S. (2001). Luonnon virkistyskäytön kysyntätutkimuksen aineistot ja menetelmät. (Données et méthodes d'enquête sur l'utilisation récréative de la nature). Dans Luonnon Virkistyskäyttö (l'utilisation récréative de la nature). Finnish Forest Research Institute, 802.

Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2007
Vol. 33, N° 2, pp. 149-156
Statistique Canada, N° 12-001-X au catalogue

Pondération dans les échantillons rotatifs : le cas de l'enquête SILC en France

Pascal Ardilly et Pierre Lavallée [1]

Résumé

L'Enquête européenne sur le revenu et les conditions de vie (*Statistics on Income and Living Conditions*, SILC) a remplacé le Panel européen à partir de 2004. Elle permet de produire des statistiques annuelles sur la répartition des revenus, ainsi que sur la pauvreté et l'exclusion sociale. Cette enquête longitudinale, dont la collecte a eu lieu pour la première fois en France en mai 2004, touche tous les individus de plus de 15 ans occupant les 16 000 logements tirés dans l'échantillon-maître et la base de sondage des logements neufs. Tous ces individus sont suivis au cours du temps, même lorsqu'ils changent de logement. L'enquête doit aussi fournir des estimations transversales de qualité.

Afin de limiter la charge des enquêtés, le plan de sondage préconisé pour SILC par Eurostat est un schéma rotatif basé sur quatre panels d'une durée de quatre ans chacun avec remplacement d'un panel tous les ans. La France a néanmoins choisi de porter la durée de ses panels à neuf années. Le plan de sondage rotatif permet de répondre aux besoins longitudinaux et transversaux de l'enquête. Cependant, il pose des défis en matière de pondération.

Après un rappel du contexte de l'inférence lorsqu'on pratique une enquête longitudinale, l'article traite des pondérations longitudinales et transversales, qui sont conçues de manière à produire des estimateurs approximativement sans biais.

Mots clés : Enquête longitudinale; panel; méthode du partage des poids; pondération longitudinale; pondération transversale.

1. Introduction

L'enquête SILC (*Statistics on Income and Living Conditions*) est une enquête européenne portant sur la mesure du revenu et sur l'évaluation des conditions de vie des personnes vivant en ménage ordinaire (on exclut donc les personnes vivant en communauté). Elle a remplacé, à partir de l'année 2004, le panel communautaire. Bien que l'enquête soit européenne, et donc sous la tutelle d'Eurostat, elle est menée de façon indépendante à l'intérieur de chaque état membre de l'Union européenne. Les états membres - comme ici la France - peuvent ainsi adapter le plan de sondage proposé par Eurostat afin de répondre à leurs besoins nationaux. Le traitement des données est aussi effectué à l'intérieur de chaque état membre, comme c'est le cas habituellement pour les enquêtes d'Eurostat au sein de l'Union européenne. Le présent article se restreint au cas de l'enquête SILC en France, mais il pourrait aussi concerner les autres états membres de l'Union européenne.

L'enquête SILC est une *enquête longitudinale*, menée en mai de chaque année, qui s'intéresse aux individus physiques beaucoup plus qu'aux ménages, et qui s'effectue en face-à-face auprès de toutes les personnes résidant dans les logements échantillonnés. Cette enquête peut se voir comme la version européenne de l'Enquête sur la dynamique du travail et du revenu (EDTR) menée par Statistique Canada (voir Lavallée 1995, ainsi que Lévesque et Franklin 2000).

L'échantillon de l'enquête SILC est rotatif : chaque année, à partir de 2004 , il est constitué par la réunion de neuf sous-échantillons panels, tirés dans des conditions identiques en régime stationnaire, pour partie dans l'échantillon-maître, pour partie dans la base de sondage des logements neufs (BSLN). L'échantillon-maître et la BSLN sont deux bases de sondage de logements construites respectivement à partir du Recensement de la population (RP) de la France et des données du Système d'information et de traitement automatisé des données élémentaires sur les logements et les locaux (SITADEL) (voir Ardilly 2006).

Chaque panel entrant dans SILC réunit tous les individus résidant dans l'ensemble des logements tirés. L'interrogation de l'ensemble des individus des ménages appartenant aux logements sélectionnés permet de produire des estimations à la fois au niveau des individus et des ménages de la population. Elle permet, de plus, d'optimiser les coûts de collecte en maximisant le nombre d'individus atteints pour chaque contact. Certaines estimations concernent néanmoins un champ réduit, défini par les personnes ayant 16 ans ou plus au 31 décembre de l'année d'enquête.

Chaque année, un sous-échantillon sort et un sous-échantillon entre pour le remplacer. La première année où l'enquête a eu lieu, soit 2004, chaque sous-échantillon comprenait 1 780 logements (à quelques unités près, du fait des procédures d'arrondi). À partir de la seconde année, donc dès 2005, la taille du sous-échantillon entrant de l'année a été fixée à 3 000 logements. Notons que

1. Pascal Ardilly, Divison « Échantillonnage et traitement statistique des données », INSEE, Paris, France. Courriel : pascal.ardilly@atih.sante.fr; Pierre Lavallée, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario), K1A 0T6 Canada. Couriel : pierre.lavallee@statcan.ca.

l'initialisation, en 2004, a conduit à un échantillon de 16 000 logements, divisé en neuf parties égales. L'une de ces parties a été interrogée une seule fois (en 2004), une autre deux fois (2004 et 2005), une autre trois fois (2004, 2005 et 2006), *etc*. En régime de croisière, un panel donné sera interrogé neuf années de suite. Durant la phase d'initialisation, qui s'achèvera en réalité en 2012 avec la sortie du neuvième et dernier sous-échantillon issu du tirage de 2004, les sous-échantillons seront évidemment sollicités à moins de neuf reprises.

La procédure d'échantillonnage proprement dite est la procédure standard utilisée lorsqu'on tire dans l'échantillon-maître et dans la BSLN (voir Ardilly 2006). Dans le cas présent, il n'y a aucune surreprésentation de catégories d'individus. C'est une enquête à taux uniforme - aux arrondis près - à l'exception des logements vacants ruraux et des résidences secondaires au RP de 1999 qui sont devenus principaux à la date de l'enquête, qui sont comme de tradition sous-représentés.

Le protocole de collecte permet de considérer chaque sous-échantillon comme un véritable panel d'individus : en effet, on suit physiquement les personnes qui quittent leur logement, les différentes directions régionales de l'INSEE se transmettant les dossiers des individus du panel qui déménagent. Pour plus de détails sur le plan de sondage de SILC, on peut consulter le *Journal Officiel de l'Union Européenne* du 17 novembre 2003 et les documents internes de l'INSEE décrivant l'échantillonnage pratiqué en France.

Comme SILC est une enquête longitudinale où il y a chevauchement de panels dans le temps, la pondération de l'échantillon amène une problématique particulière. Cet article présente en détail les deux types de pondération utilisés pour SILC. On discutera en premier lieu de principes généraux reliés au plan de sondage de SILC. En deuxième lieu, on présentera la pondération longitudinale et finalement, on s'attardera sur la pondération transversale.

Il est à noter que l'on n'abordera pas les questions de correction de la non-réponse ni du redressement des estimations. Le traitement de ces questions renvoie à ce que l'on retrouve en général dans les enquêtes longitudinales comme, par exemple, l'EDTR (voir Lavallée 1995, ainsi que Lévesque et Franklin 2000).

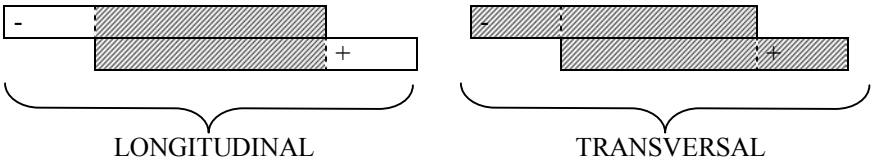
2. Principes généraux

2.1 Deux approches : la vision longitudinale et la vision transversale

Chaque année, on dispose d'un échantillon d'individus tous panélisés dont les huit neuvièmes ont déjà été interrogés au moins une fois les années passées (en l'absence de non-réponse).

On peut s'intéresser en particulier à deux types de paramètres : des totaux annuels Y_t (ou leurs satellites), ou des évolutions de totaux $\Delta_{t,t+1}$ entre deux années données, consécutives ou non. Pour simplifier, on considérera désormais qu'il s'agit de différences de totaux entre deux années consécutives. Quand on parle d'évolutions, il faut préciser les populations d'inférence qui entrent en jeu. Il y a alors deux façons de voir les choses : soit on raisonne sur des populations évolutives avec le temps et l'approche est dite transversale, soit on raisonne à population fixe et l'approche est dite longitudinale. Si on note Ω_t la population complète du champ de l'enquête l'année t, le total annuel à l'année t est donné par $Y_t = \sum_{i\in\Omega_t} Y_i^t$, où Y_i^t est une variable d'intérêt mesurée pour l'individu i. Lorsqu'on parle d'évolution, l'objectif peut être d'estimer la différence $\Delta^*_{t,t+1}$ entre le total Y_{t+1} à $t+1$ sur Ω_{t+1} et le total Y_t à t sur Ω_t, c'est-à-dire $\Delta^*_{t,t+1} = Y_{t+1} - Y_t$. Ceci correspond à une vision transversale. Sinon, l'objectif peut être d'estimer la différence $\Delta_{t,t+1}$ entre les totaux définis sur les unités communes aux populations Ω_{t+1} et Ω_t, les différences d'effectifs entre les deux populations s'expliquant par les unités entrant (naissances) et sortant (morts) de ces populations. Ceci correspond à une vision longitudinale. Soit $\Omega_{t,t+1} = \Omega_t \cap \Omega_{t+1}$, la population commune entre t et $t+1$. On définit alors $\Delta_{t,t+1}$ selon $\Delta_{t,t+1} = \sum_{i\in\Omega_{t,t+1}} (Y_i^{t+1} - Y_i^t)$.

Les schémas suivants synthétisent les deux approches. Le rectangle du haut symbolise la population complète à t et celui du bas la population complète à $t+1$. La partie « moins » représente les morts au sens large (décès, émigration, passage de l'individu en communauté,…) et la partie « plus » représente les naissances au sens large (nouveau-nés, immigration, entrée dans le champ par le franchissement d'un seuil d'âge,…). La partie grisée représente, à chaque date, la population d'inférence.



2.2 Enquêtes répétées dans le temps et stratégies envisageables

L'objectif est évidemment de pouvoir produire à la fois des estimations longitudinales et des estimations transversales. On peut envisager essentiellement trois stratégies :

1. Un échantillonnage « indépendant » chaque année. En fait, compte tenu de l'existence d'un échantillon-maître et d'une base de sondage de logements neufs (BSLN), les tirages s'effectuent tous les ans dans les mêmes communes, et par conséquent il n'y a pas de véritable indépendance entre les différents sous-échantillons. Cette solution est largement perfectible en terme de précision des évolutions.
2. Un échantillonnage intégralement panélisé, c'est-à-dire un tirage initial d'échantillon interrogé chaque année. Ce scénario pose en particulier un problème de charge, car l'opération SILC est engagée pour une durée indéterminée. De ce fait, il est irréaliste.
3. Un échantillon rotatif. C'est ce scénario qui a été choisi, compte tenu des avantages qu'il présente pour satisfaire les attentes en matière à la fois longitudinale et transversale.

Le tableau qui suit qualifie les trois plans de sondage envisageables en fonction des deux approches souhaitées.

TYPE d'échantillon	Approche TRANSVERSALE	Approche LONGITUDINALE
« Indépendant » chaque année	NATUREL	POSSIBLE mais moins efficace
Panel	IMPOSSIBLE sans tirage complémentaire	NATUREL
Rotatif	POSSIBLE	POSSIBLE

La stratégie rotative présente quatre grands atouts :

i. Elle réduit l'erreur d'échantillonnage associée à la mesure des évolutions (sur le principe, comme pour les panels, même si elle est moins efficace en théorie que le panel « pur »).

ii. Elle limite la charge des enquêtés par rapport au panel « pur ». En la circonstance, s'agissant pour la France d'un panel de neuf années, cet argument doit être utilisé avec modestie. Il a cependant plus de force dans le scénario préconisé par Eurostat, qui donne lieu à une enquête annuelle durant quatre années consécutives.

iii. Elle permet de prendre en compte d'une manière très « naturelle » l'évolution de la population avec le temps. Ce point sera plus compréhensible lorsqu'on abordera la question de la couverture des populations nouvelles.

iv. Elle permet de réduire les erreurs d'observation (comme les panels).

En revanche, on peut lui trouver au moins trois défauts :

i. Elle nécessite un suivi des individus dans le temps, ce qui occasionne des coûts de dépistage et des non-réponses du fait des déménagements.

ii. Par nature, la longueur des séries individuelles se limite à neuf années, ce qui est déjà fort appréciable, mais évidemment moins riche qu'un pur panel.

iii. La technique de pondération longitudinale/transversale n'est pas simple…

3. La pondération longitudinale

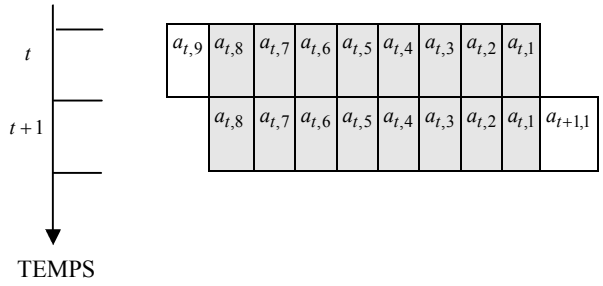
C'est une pondération a priori un peu plus simple à concevoir que la pondération transversale parce qu'il n'y a pas à tenir compte de l'évolution de la population dans le temps (en dehors des « morts », qui par convention disparaissent du champ avec le temps, mais ce point ne pose pas vraiment de problème technique). On rappelle que le principe de l'estimation longitudinale consiste à pratiquer une inférence sur la base d'une unique population considérée à une date initiale.

C'est clairement le caractère rotatif du plan de sondage qui complique la pondération, puisque entre deux années consécutives t et $t+1$, on va mobiliser huit panels distincts, tirés dans des populations différentes (il s'agit de populations d'individus physiques qui sont, bien entendu, différentes d'une année sur l'autre). Si on ne manipulait qu'un seul panel, il suffirait de s'en tenir à l'utilisation des poids de sondage associés aux individus du panel encore dans le champ à la date t, ni plus ni moins, puisque ces poids sont calculés une fois pour toutes au moment du tirage et permettent chaque année, sur toute la durée de vie du panel, une inférence sur la population initiale.

La difficulté essentielle consiste à représenter la population Ω_t à la date t à partir de huit sous-échantillons panels tirés à des dates différentes, donc dans des populations différentes. On peut comprendre intuitivement qu'un individu physique donné ait *in fine* une probabilité de sélection à la date t qui dépend du nombre de sous-échantillons panels dans lesquels il est susceptible d'être tiré. On suppose dans cette partie qu'il n'y a pas de non-réponse. La situation peut être formalisée de la manière suivante, en notant :

$a_{t,k}$ = sous-échantillon panel à enquêter l'année t en $k^{\text{ième}}$ interrogation, et $s_{t,t+1} = \bigcup_{k=1}^{8} a_{t,k}$.

On notera qu'on peut écrire $a_{t+1,k+1} = a_{t,k}$ $(\forall t, \forall k \neq 9)$ puisque par principe on reprend intégralement chaque sous-échantillon panel (non sortant) d'une année sur l'autre. Schématiquement, on a :



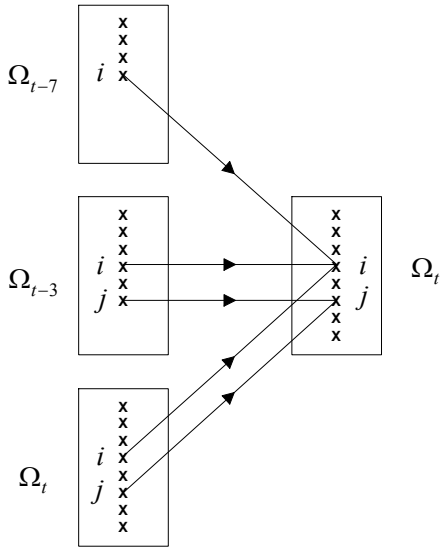
La partie grisée représente $s_{t,t+1}$ qui est l'échantillon exploité dans cette approche longitudinale. C'est en effet sur les individus de $s_{t,t+1}$ que l'on peut obtenir à la fois les informations Y_i^t et Y_i^{t+1} sur l'individu i définies respectivement aux dates t et $t+1$.

Soit un individu i quelconque de Ω_t, dans le champ de l'enquête à t. On note L_i le nombre d'années parmi $\{t-7, t-6, ..., t-1, t\}$ durant lesquelles l'individu i se trouvait dans le champ de l'enquête, donc était susceptible d'être tiré dans un panel « entrant ». Notons que l'on suppose ici que chaque année la base de sondage couvre exactement le champ de l'enquête. On a $L_i \in \{1, 2, 3, \ldots, 8\}$. Par ailleurs, on note K_i l'ensemble des indices k parmi 1, 2, 3, …, 8 pour lesquels on a $i \in a_{t,k}$. Il s'agit donc, à la date t, des numéros des panels dans lesquels on retrouve l'individu i. Pour tout i de $s_{t,t+1}$, K_i sera par construction un ensemble contenant au moins un élément. La plupart du temps, K_i ne comprendra en fait qu'un seul indice, mais parfois il pourra en comprendre deux, voire davantage : en effet, ce cas surviendra si i est tiré dans un panel, qu'il déménage et que son nouveau logement est échantillonné dans un autre panel, une année ultérieure. Il est à noter que notre contexte exclut qu'un logement donné soit tiré deux fois, parce qu'il y a un principe de non-réinterrogation des logements de l'échantillon-maître et de la BSLN. Mais ce n'est qu'une convention de nature pratique, la théorie s'accommodant fort bien d'un système dans lequel on pourrait retirer les logements.

Si $i \in a_{t,k}$, appelons $W_i(t,k)$ son poids de sondage « brut » : il s'agit du poids de sondage du logement dans lequel se trouve i à la date de son tirage en tant qu'individu-panel, donc lors du tirage annuel dans Ω_{t-k+1}. Ce système de poids permet une inférence directe du sous-échantillon $a_{t,k}$ vers la population complète Ω_{t-k+1}. En particulier, $\sum_{i \in a_{t,k}} W_i(t,k)$ estime sans biais le nombre total d'individus appartenant au champ de l'enquête et à la population Ω_{t-k+1}. Pour SILC en France, il s'agit d'un ordre de grandeur de 60 millions. Le poids longitudinal à affecter à tout individu i de $s_{t,t+1}$ sera *in fine* :

$$W_i^{t,t+1} = \frac{1}{L_i} \sum_{k \in K_i} W_i(t,k). \qquad (1)$$

Cette expression découle de l'application de la méthode de partage des poids (voir Lavallée 1995, ainsi que Lavallée 2002) où on définit la population initiale (celle des unités d'échantillonnage) comme réunion des populations $\Omega_{t-7}, ..., \Omega_{t-1}, \Omega_t$ et la population finale (celle des unités d'observation) comme Ω_t. Le schéma ci-dessous illustre le contexte, où, pour plus de clarté, nous n'avons pas reproduit les huit sous-populations initiales, mais seulement trois d'entres-elles. Le nombre de liens apparaît alors clairement égal à L_i (ici, par exemple, i a exactement huit liens, mais j en a strictement moins de huit parce qu'il n'apparaît pas dans les bases de sondage les plus anciennes). Pratiquement, il est réaliste de faire comme si on avait $\Omega_{t-7} \subset \Omega_{t-6} \subset \ldots \subset \Omega_{t-1} \subset \Omega_t$. On peut raisonner sur des populations emboîtées parce que, sauf exception, les individus qui sortent du champ au cours du temps avant t ne seront pas présents dans $s_{t,t+1}$.



La formule (1) fournit l'expression la plus générale possible du poids longitudinal « brut ». On peut ensuite la simplifier dans différents contextes. Si par exemple on néglige les cas où un individu-panel peut être tiré deux fois ou plus, on a

$$W_i^{t,t+1} = \frac{W_i}{L_i} \qquad (2)$$

où W_i est le poids de i relatif à l'unique sous-échantillon panel dans lequel il figure à la date t. Dans le cas de la France, compte tenu des tailles d'échantillon en jeu, adopter

in fine cette expression paraît tout à fait opportun. Si on se place dans un cadre idéal - qui paraît néanmoins trop simplifié dans notre contexte - où la population n'évolue pas dans le temps, on aura $L_i = 8$ pour tout *i*. Notons cependant que la population évolue beaucoup en neuf années, mais avec des durées de panélisation plus courtes, ce cas idéal peut être une approximation acceptable. Si, de plus, les panels sont tirés à probabilités égales, W_i sera égal à une constante *W* et alors

$$W_i^{t,t+1} = \frac{W}{8}. \qquad (3)$$

Ce cas de figure reste très peu probable dans le cas de la France. D'une part, jusqu'en 2012, il y a coexistence de sous-échantillons tirés avec des poids bruts nettement distincts (voir l'introduction). D'autre part, on aura tendance à concevoir l'échantillonnage en fixant le nombre total de logements à tirer (alors même que le nombre total de logements augmente) et non pas en raisonnant sur un objectif de taux de sondage constant.

Notons que la formule (3) est intuitive : finalement, tout se passe « comme si » n'importe quel individu de l'échantillon longitudinal $s_{t,t+1}$ avait une probabilité de sélection égale à huit fois celle qui caractérise chaque sous-échantillon panel composant $s_{t,t+1}$.

Ce qui précède s'applique au régime stationnaire et doit être légèrement adapté durant la phase d'initialisation du processus, c'est-à-dire jusqu'en 2012. La première opération de nature longitudinale porte sur les données conjointes 2004-2005, pour estimer des évolutions entre 2004 et 2005 avec la population de référence 2004 (privée des « morts » en 2005). Dans ce contexte, il suffit de diviser tous les poids W_i des huit sous-échantillons $a_{2004,1}$ à $a_{2004,8}$ par huit - autrement dit $L_i = 8$ pour tout *i*. En 2006, lorsqu'on s'intéressera aux évolutions 2005-2006, le dénominateur L_i pourra prendre deux valeurs seulement. Dans le premier scénario, l'individu-panel *i* était dans la base de sondage utilisée en 2004 (donc potentiellement tirable en 2004) et alors $L_i = 8$. Ceci vient du fait que tout se passe comme si, en 2004, on avait effectué les sept tirages des panels $a_{2005,2}$ à $a_{2005,8}$ exactement dans les mêmes conditions Dans le second scénario, l'individu *i* n'était pas dans la base de sondage de 2004 - mais alors il est dans la base 2005 et il se trouve nécessairement dans $a_{2005,1}$ - et $L_i = 1$. Pour mesurer les évolutions 2006-2007, L_i pourra être égal à 1, 2 ou 8, et ainsi de suite. Pour retrouver l'ensemble des valeurs possibles de L_i parmi $\{1, 2, 3, \ldots, 8\}$, il faudra attendre la mesure des évolutions 2011-2012.

Passée cette étape de pondération longitudinale, on obtient des poids longitudinaux $W_i^{t,t+1}$ et on forme l'estimateur de la différence $\Delta_{t,t+1}$ selon

$$\hat{\Delta}_{t,t+1} = \sum_{s_{t,t+1}} W_i^{t,t+1}.(Y_i^{t+1} - Y_i^t). \qquad (4)$$

A priori, les poids $W_i^{t,t+1}$ ne sont utilisés que dans le cadre d'une estimation d'évolution. Pour des estimations ponctuelles, ils apparaissent sans intérêt parce que la population d'inférence n'a pas grande signification à date donnée. Rappelons que jusqu'ici les $W_i^{t,t+1}$ n'ont fait l'objet d'aucune correction pour non-réponse, ni redressement. En pratique, l'estimateur (4) pour l'enquête SILC sera sujet à de tels ajustements.

L'estimation de la différence $\Delta^*_{t,t+1} = Y_{t+1} - Y_t$ correspond à une vision transversale : elle fait ainsi appel à la pondération qui est présentée dans la section suivante.

4. La pondération transversale

Il s'agit de pratiquer une inférence sur la population globale Ω_t du champ de l'enquête à la date courante *t*. La difficulté essentielle tient au fait qu'un sous-échantillon (panélisé) donné ne couvre correctement, en théorie, la population que l'année de son tirage. Passée cette année, le sous-échantillon panel ne représente plus la population nouvelle des « naissances », c'est-à-dire ceux qui entrent dans le champ de l'enquête. Cela concerne en particulier les nouveau-nés, les immigrants, les individus dont l'âge atteint certains seuils, les personnes anciennement sans domicile qui retrouvent un logement ordinaire, les retours de communautés, *etc*. Si en pratique on peut imaginer s'en satisfaire pendant quelque temps, ce défaut de couverture devient assez vite excessif (cela est vrai chaque année pour la plupart des sous-échantillons panels) et il faut d'une façon ou d'une autre obtenir un échantillon complémentaire au panel. Il est à noter que la problématique de l'évolution dans le temps de la population est fortement dissymétrique parce que la sous-population qui disparaît d'une année sur l'autre (les « morts ») ne pose pas de problème particulier en terme de pondération.

Dans l'enquête SILC, l'échantillon complémentaire est obtenu en appliquant la méthodologie suivante : on décide, pour chaque individu-panel enquêté lors du processus de suivi longitudinal, d'interroger l'ensemble des individus du ménage dans lequel se trouve l'individu-panel. Ainsi, tout ménage enquêté dans l'optique transversale est composé de deux types de personnes : des individus panel et des cohabitants (on nomme ainsi toute personne enquêtée qui n'est pas individu-panel). Cette méthodologie permet de couvrir une grande partie des « naissances » (au sens large) au sein de la population. Cependant, elle ne permet pas d'atteindre les ménages constitués seulement de « naissances » comme, par exemple, les ménages contenant seulement des immigrants. Précisons que la détermination

du statut de « naissance » se fait généralement en demandant, pour les nouveaux-nés, la date de naissance, et pour les immigrants, la date d'entrée au pays. On ajoute qu'en pratique, le défaut de couverture des naissances est en général considéré comme négligeable parce qu'il est en partie corrigé par l'utilisation du redressement.

La technique centrale utilisée pour produire les poids transversaux est la méthode de partage des poids (Lavallée 2002). Rappelons qu'à l'année *t*, on dispose de neuf sous-échantillons panels $a_{t,k} (1 \le k \le 9)$. On présente ici deux approches possibles pour l'application de la méthode du partage des poids. Notons que les informations à recueillir dans le questionnaire sont identiques pour mettre en oeuvre les deux méthodes.

4.1 Méthode 1

L'approche la plus rigoureuse consiste à relier l'ensemble des neuf sous-échantillons $a_{t,k}$ à l'échantillon transversal de l'année *t*, que nous noterons $\tilde{u}_t$ (Merkouris 2001). L'échantillon $\tilde{u}_t$ correspond donc à $s_{t,t} = \bigcup_{k=1}^{9} a_{t,k}$. Il faut tout d'abord commencer par définir les liens associés à ce schéma : lorsqu'un individu-panel quelconque de l'un des neuf sous-échantillons $a_{t,k}$ a été désigné par le sort, il pointe sur lui-même en tant qu'individu de l'échantillon transversal à *t* (schéma voisin de celui du 3.1). Dans ces conditions et en régime stationnaire, le poids transversal $W_i^{t(1)}$ d'un individu quelconque *i* de $\tilde{u}_t$ s'obtient de la façon qui suit. On note *m* le ménage auquel appartient *i*. On a

$$W_i^{t(1)} = \frac{\sum_{k=1}^{9} \sum_{\substack{j \in m \\ j \in a_{t,k}}} W_j(t,k)}{\sum_{k=1}^{9} \sum_{\substack{j \in m \\ j \in \Omega_{t-k+1}}} 1} \qquad (5)$$

où $W_j(t,k)$ est le poids de sondage qui découle de l'échantillonnage $a_{t,k}$.

Cette expression montre que tous les individus d'un même ménage ont à la fin le même poids. Au numérateur, on trouve la somme de tous les poids « bruts » (ceux qui reflètent l'échantillonnage) de tous les individus-panels du ménage, étant entendu qu'en général un individu-panel n'apparaît que dans un seul sous-échantillon mais qu'il peut y avoir des cas où un individu-panel a été tiré deux fois ou même davantage sur une période de neuf années consécutives (pour cause de déménagement, essentiellement). Il est à noter que l'échantillonnage de logements dans l'échantillon-maître et la BSLN s'appuie sur un principe de non-réinterrogation des logements déjà tirés et ainsi, dans le cas de SILC, la probabilité qu'un individu n'ayant pas changé de ménage apparaisse dans deux panels distincts est nulle.

Comme dans le cas longitudinal (voir 3.1), la pondération ne peut s'effectuer que si le système informatique de gestion des données est en mesure de rattacher chaque individu-panel de $\tilde{u}_t$ à l'ensemble des échantillons panels $a_{t,k}$ dans lesquels il se trouve. Au dénominateur, on dénombre pour chacune des neuf années $t-8$ à *t* considérées, les individus du ménage (qu'ils soient individus-panels ou cohabitants) qui se trouvent dans la base de sondage utilisée pour le tirage du sous-échantillon panel entrant l'année en question. Ce calcul nécessite évidemment la disponibilité de l'information via le questionnaire.

Cette approche a un double atout : d'une part elle est parfaitement générale, et d'autre part elle donne immédiatement lieu à des poids transversaux sans biais parce que tout ménage transversal est nécessairement relié à l'un quelconque des neuf sous-échantillons considérés. Le fait qu'il y ait chaque année un sous-échantillon entrant permet de représenter l'intégralité de la population transversale Ω_t, c'est-à-dire, dans un langage plus technique, assure l'existence d'au moins un lien pour chaque ménage considéré à *t*. C'est une propriété intéressante de l'échantillonnage rotatif que nous avons déjà mentionnée à la section 2.2. En contrepartie, la formule de pondération a un inconvénient qui est sa (relative) complexité, à la fois sur le plan théorique et lors de la phase de programmation informatique.

Dans la phase d'initialisation (donc jusqu'en 2011 compris), cette expression doit être adaptée : le numérateur ne change pas mais le dénominateur dénombre les individus échantillonnables à partir de 2004, première année de réalisation de l'enquête. En 2004, la pondération est évidente puisqu'il n'y a pas de partage des poids, mais en 2005 on prendra

$$W_i^{t(1)} = \frac{\sum_{k=1}^{9} \sum_{\substack{j \in m \\ j \in a_{t,k}}} W_j(t,k)}{\left(\sum_{\substack{j \in m \\ j \in \Omega_{2005}}} 1 \right) + 8 \cdot \left(\sum_{\substack{j \in m \\ j \in \Omega_{2004}}} 1 \right)}. \qquad (6)$$

En 2006, ce sera

$$W_i^{t(1)} = \frac{\sum_{k=1}^{9} \sum_{\substack{j \in m \\ j \in a_{t,k}}} W_j(t,k)}{\left(\sum_{\substack{j \in m \\ j \in \Omega_{2006}}} 1 \right) + \left(\sum_{\substack{j \in m \\ j \in \Omega_{2005}}} 1 \right) + 7 \cdot \left(\sum_{\substack{j \in m \\ j \in \Omega_{2004}}} 1 \right)}. \qquad (7)$$

4.2 Méthode 2

On peut avoir une vision alternative de la pondération transversale qui conduit à une expression de poids « un peu » plus simple et qui peut se programmer plus

facilement, mais qui se heurte à une difficulté qui n'apparaissait pas dans la méthode précédente et qui risque en pratique de rendre la pondération définitive un peu moins rigoureuse. L'idée est de raisonner non pas sur l'ensemble des sous-échantillons, mais sous-échantillon par sous-échantillon. On considère un quelconque des neuf sous-échantillons $a_{t,k}$ ainsi que l'échantillon de ménages auquel il mène. On applique alors le partage des poids, ce qui donne en régime stationnaire une pondération individuelle égale à

$$\tilde{W}_i(t,k) = \frac{\sum_{\substack{j\in m \\ j\in a_{t,k}}} W_j(t,k)}{\sum_{\substack{j\in m \\ j\in \Omega_{t-k+1}}} 1} \qquad (8)$$

pour tout individu *i* du ménage *m*. On vérifie très facilement que si $k=1$ (cas du sous-échantillon entrant), $\tilde{W}_i(t,1)$ est le poids de tirage du ménage *m*.

La difficulté associée à cette approche est liée à l'existence (a priori) à la date *t* d'individus non enquêtables parce qu'ils appartiennent à des ménages qui ne sont pas du tout « atteignables » au travers de l'échantillonnage $a_{t,k}$ (dès lors que $k \geq 2$), c'est-à-dire qui ont une probabilité nulle d'être enquêtés à *t*. Ce phénomène perturbateur n'existait pas dans la méthode précédente grâce à la prise en compte globale de l'ensemble des sous-échantillons puisqu'à la date *t* tout ménage a une probabilité strictement positive d'être sélectionné, au moins au travers de $a_{t,1}$. C'est une nouvelle occasion de souligner un des atouts essentiels de l'échantillonnage rotatif qui constitue une technique permettant chaque année de couvrir l'intégralité de la population. Dans notre approche, il est clair que si on considère $a_{t,k}$ pour $k \geq 2$, on ne couvre pas la population des ménages constitués exclusivement d'« immigrants » (au sens large) entre $t-k+1$ et *t*. Pour formaliser le contexte et aboutir au poids transversal final, on notera $\Omega_{\alpha,t}^{\text{immig}}$ la population d'« immigrants » (au sens large) présente à *t* dans des ménages ne comprenant que des immigrants échantillonnables après l'année α, avec $t-8 \leq \alpha \leq t-1$. Notons que, plus précisément, il faudrait dire « échantillonnables à partir d'une date strictement postérieure à la date de collecte de l'année α ».

A la date *t*, la population complète Ω_t est partitionnée en neuf composantes : les huit sous-populations $\Omega_{\alpha,t}^{\text{immig}}$, avec α variant de $t-8$ à $t-1$, et la sous-population constituée par les individus, soit qui étaient déjà enquêtables à $t-8$, soit qui sont devenus enquêtables à une date ultérieure à $t-8$ (donc des immigrants au-delà de $t-8$) mais qui sont intégrés à *t* dans un ménage comprenant au moins une personne enquêtable à $t-8$. Notons que l'on considère que si le ménage à *t* comprend au moins une personne échantillonnable à $t-8$, il en sera de même à toute date comprise entre $t-8$ et $t-1$. Cela revient à négliger les situations où un individu dans le champ à une date donnée en sort durant quelque temps (émigration, par exemple), puis y revient ensuite.

Par ailleurs, on note $\tilde{u}_{t,k}$ l'échantillon transversal à *t* issu du panel $a_{t,k}$, ce qui conduit à $\bigcup_{k=1}^{9} \tilde{u}_{t,k} = \tilde{u}_t$. Soit $Y_{\alpha,t}^{\text{immig}}$, le total des Y_i^t défini sur $\Omega_{\alpha,t}^{\text{immig}}$. On a alors, suite au partage des poids effectué pour tout $k = 2, \ldots, 9$:

$$E\left(\sum_{j\in\tilde{u}_{t,k}} \tilde{W}_j(t,k)\cdot Y_j^t\right) = \sum_{\Omega_t} Y_j^t - \sum_{\alpha=t-k+1}^{t-1} Y_{\alpha,t}^{\text{immig}}$$

$$= Y_t - \sum_{\alpha=t-k+1}^{t-1} Y_{\alpha,t}^{\text{immig}} \qquad (9)$$

et

$$E\left(\sum_{j\in\tilde{u}_{t,1}} \tilde{W}_j(t,1)\cdot Y_j^t\right) = \sum_{\Omega_t} Y_j^t = Y_t \qquad (10)$$

puisque $\tilde{u}_{t,1} = a_{t,1}$.

Avec un système de panels à courte durée, on pourrait peut-être négliger les $Y_{\alpha,t}^{\text{immig}}$ devant le vrai total sur Ω_t et alors le poids transversal final « brut » de tout individu *i* serait $\tilde{W}_i(t,k)/9$ si *i* est issu de $a_{t,k}$, ce qui conduirait à l'estimateur final

$$\hat{Y}_t = \frac{1}{9}\sum_{k=1}^{9}\sum_{i\in\tilde{u}_{t,k}} \tilde{W}_i(t,k)\cdot Y_i^t$$

$$= \frac{1}{9}\sum_{i\in\tilde{u}_t} \tilde{W}_i(t,k)\cdot Y_i^t. \qquad (11)$$

Les panels utilisés en France ont cependant une durée de vie longue, aussi il est fort possible que l'on ne puisse pas raisonner ainsi (l'examen des fichiers de collecte permettra d'en juger) et qu'il soit nécessaire de pondérer spécifiquement les individus des $\Omega_{\alpha,t}^{\text{immig}}$. Dans ces conditions, on vérifie que tout individu *i* de $\Omega_{\alpha,t}^{\text{immig}}$ qui se trouve finalement dans l'échantillon transversal $\tilde{u}_t$ aura un poids transversal brut $W_i^{t(2)}$ égal à la valeur $\tilde{W}_i(t,k)$ issue du partage des poids, divisée par $t-\alpha$ (et donc $1 \leq t-\alpha \leq 8$). Pour sa part, tout individu de Ω_t qui n'appartient à aucun des $\Omega_{\alpha,t}^{\text{immig}}$ (donc la grande majorité des cas) aura un poids final égal à $\tilde{W}_i(t,k)/9$. On remarquera par ailleurs que si *i* se trouve dans $\Omega_{\alpha,t}^{\text{immig}}$, il ne peut être enquêté qu'au travers de $a_{t,1}, a_{t,2}, \ldots, a_{t,t-\alpha}$. On obtient ainsi

$$W_i^{t(2)} = \begin{cases} \tilde{W}_i(t,k)/(t-\alpha) & \text{si } i \in \Omega_{\alpha,t}^{\text{immig}} \\ \tilde{W}_i(t,k)/9 & \text{sinon} \end{cases}. \qquad (12)$$

Durant la phase d'initialisation, il faut adapter les pondérations. En 2005, les individus de $\Omega^{\text{immig}}_{2004,2005}$ auront un poids final transversal directement issu du tirage du logement dans $a_{2005,1}$ (ils ne peuvent être atteints qu'au travers de ce panel entrant). En revanche, tous les autres individus sont « normalement » enquêtables à partir des neufs panels $a_{2005,k}$ $(1 \le k \le 9)$, si bien que leurs poids issus du partage des poids seront tous systématiquement divisés par 9. En 2006, les individus de $\Omega^{\text{immig}}_{2005,2006}$ auront un poids égal à celui du logement dans lequel ils résident et qui reflète directement le tirage de $a_{2006,1}$, ceux de $\Omega^{\text{immig}}_{2004,2006}$ auront leurs poids issus du partage des poids divisés par 2, et tous les autres individus auront leurs poids issus du partage des poids divisés par 9.

Ce traitement s'effectue bien sous-échantillon par sous-échantillon et ne doit pas tenir compte de ce qui survient dans les autres sous-échantillons. Si un individu est enquêté à t par l'intermédiaire de deux (ou plus) sous-échantillons $a_{t,k}$ distincts, on déroule le traitement complet associé à chacun des deux (ou plus) sous-échantillons. Ce peut être le cas, par exemple, d'un ménage composé de deux individus-panels provenant de deux sous-échantillons $a_{t,k}$ différents parce que ces individus se sont mariés et qu'avant leur mariage ils étaient suivis chacun séparément en formant un ménage de taille un. Dans cette configuration, chacun des deux individus est « formellement » enquêté deux fois, une fois en tant qu'individu-panel, une fois en tant que cohabitant.

Finalement, pour l'estimation de la différence $\Delta^*_{t,t+1} = Y_{t+1} - Y_t$, on pourra utiliser les poids $W_i^{t(1)}$ issus de la méthode 1, et ainsi calculer

$$\hat{\Delta}^*_{t,t+1} = \sum_{i \in \tilde{u}_{t+1}} W_i^{t+1(1)} Y_i^{t+1} - \sum_{i \in \tilde{u}_t} W_i^{t(1)} Y_i^t . \qquad (13)$$

Sinon, on pourra utiliser les poids $W_i^{t(2)}$ issus de la méthode 2. L'estimateur de la différence $\Delta^*_{t,t+1}$ sera alors donné par

$$\hat{\Delta}^*_{t,t+1} = \sum_{i \in \tilde{u}_{t+1}} W_i^{t+1(2)} Y_i^{t+1} - \sum_{i \in \tilde{u}_t} W_i^{t(2)} Y_i^t . \qquad (14)$$

Bibliographie

Ardilly, P. (2006). *Les techniques de sondage*, 2ième édition. Éditions Technip, Paris.

Lavallée, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27-35.

Lavallée, P. (2002). *Le sondage indirect, ou la méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles et Editions Ellipses.

Lévesque, I., et Franklin, S. (2000). Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics, 1997 Reference year. Document de recherche sur le revenu, Statistique Canada, Catalogue No. 75F0002MIE-00004, juin 2000.

Merkouris, T. (2001). Estimation transversale dans le cas des enquêtes auprès des ménages à panels multiples. *Techniques d'enquêtes*, 27, 189-200.

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

Singh, M.P., Drew, J.D., Gambino, J.G. et Mayda, F. (1990). *Méthodologie de l'Enquête sur la population active*. Statistique Canada, Catalogue 71-526.

Techniques d'enquête, décembre 2007
Vol. 33, N° 2, pp. 157-170
Statistique Canada, N° 12-001-X au catalogue

Regroupement de cellules lors de la poststratification

Jay J. Kim, Jianzhu Li et Richard Valliant [1]

Résumé

La poststratification est une méthode courante d'estimation dans le cas des enquêtes-ménages. Les cellules sont créées d'après les caractéristiques qui sont connues pour tous les répondants de l'échantillon et pour lesquelles il existe des dénombrements de contrôle externes provenant d'un recensement ou d'une autre source. Les inverses des corrections par poststratification sont habituellement appelés ratios de couverture. La couverture de certains groupes démographiques peut être sensiblement inférieure à 100 % et la poststratification est destinée à corriger les biais résultant d'une couverture insatisfaisante. Une méthode standard de poststratification consiste à regrouper ou à combiner certaines cellules lorsque les tailles d'échantillon sont inférieures à un minimum donné ou que les corrections des poids sont supérieures à un maximum donné. Le regroupement peut accroître ou réduire la variance d'une estimation, mais pourrait simultanément augmenter son biais. Nous étudions les effets, sur le biais et la variance, de ce type de regroupement dynamique des cellules du point de vue théorique et par simulation en utilisant une population basée sur la National Health Interview Survey de 2003. Nous proposons deux estimateurs possibles qui réduisent l'importance des corrections des poids lorsque les cellules sont regroupées.

Mots clés : Biais; combinaison de cellules; erreur de couverture; poststratification; sous-couverture; élagage des poids.

1. Introduction

La poststratification est une technique fréquemment adoptée dans la pondération des enquêtes afin 1) de réduire les variances ou 2) de corriger la couverture insuffisante de l'échantillon pour certains groupes de population cibles. Dans le cas des enquêtes-ménages aux États-Unis, le deuxième objectif est particulièrement important, parce que certains groupes démographiques, comme les jeunes hommes noirs, sont moins bien couverts que les autres (par exemple, voir Kostanich et Dippo 2000, chapitre 16). La correction pour la sous-couverture peut aboutir à des poids variables qui corrigent peut-être le biais, mais accroissent aussi les erreurs-types. Souvent, les praticiens évitent de procéder à des corrections extrêmes des poids, sacrifiant effectivement une certaine réduction du biais afin de maîtriser les variances.

Une façon de contrôler la grandeur des corrections des poids consiste à regrouper les cellules résultant de la poststratification initiale si la correction dans l'une d'elles est supérieure à une limite donnée. Little (1993), ainsi que Lazzeroni et Little (1998) s'intéressent aux méthodes de regroupement des catégories des variables de post-stratification ordinales. D'autres stratégies de regroupement des strates ou de construction d'estimateurs ont été proposées par Fuller (1966), Kalton et Maligalig (1991), ainsi que Tremblay (1986). Kim, Thompson, Woltman et Vajs (1982) décrivent certaines applications pratiques. Dans le présent article, nous étudions les effets de la combinaison des cellules sur le biais et la variance, en présumant que des cellules définies à un niveau plus fin de détail seraient préférables si les tailles d'échantillon et les corrections des poids se situaient dans certaines limites établies par les concepteurs de l'enquête.

Deux critères sont souvent utilisés pour décider si une cellule doit être ou non regroupée avec une autre. Le premier est le *ratio de couverture inverse* ou *facteur de correction initial* (FCI), qui est défini comme étant le ratio du dénombrement de contrôle au dénombrement d'échantillon initialement pondéré pour la cellule. Un ratio significativement différent de 1 indique que la couverture est trop faible ou trop élevée pour le groupe représenté par la cellule. Lorsque le FCI d'une cellule se situe en dehors de limites établies d'avance, on la combine à une autre. Par exemple, le seuil de regroupement pour un ratio « élevé » pourrait être égal à 2 et celui pour un ratio « faible », à 0,6, valeurs qui sont les seuils utilisés pour la Current Population Survey (CPS) réalisée par le Bureau of the Census des États-Unis (voir Kostanich et Dippo 2000, page 10-7). Le deuxième critère est la taille d'échantillon. Une cellule dont le dénombrement brut d'échantillon est trop faible peut être regroupée avec une autre en prenant pour motif que le FCI est instable. Nous dirons qu'une cellule possède un *faible effectif* si elle viole l'un des deux critères susmentionnés et est groupée avec une autre.

Les catégories de variables qui définissent les post-strates sont habituellement triées suivant un ordre naturel (par exemple, les catégories d'âge ou de revenu) ou un ordre commode (par exemple, race-appartenance ethnique). En pratique, on regroupe généralement une cellule avec une

1. Jay J. Kim, National Center for Health Statistics, Centers for Disease Control and Prevention; Jianzhu Li, Joint Program in Survey Methodology, University of Maryland; Richard Valliant, Survey Research Center, University of Michigan.

cellule adjacente dont les caractéristiques sont similaires, sans tenir compte des ratios de couverture différents des cellules individuelles.

Selon Kalton et Flores-Cervantes (2003, page 95), [traduction] « les méthodes qui restreignent automatiquement l'étendue des corrections redistribuent les corrections excessives de sorte que celles qui seraient autrement appliquées à certains répondants le sont à d'autres. Il conviendrait de déterminer si cette redistribution est appropriée. » Dans le présent article, nous examinons précisément cette question et déterminons les circonstances dans lesquelles la redistribution des poids due au regroupement de cellules risque d'être relativement nuisible.

Une faiblesse évidente des stratégies habituelles de regroupement de cellules est que, pour certains groupes, le biais de couverture est corrigé incomplètement. Par exemple, supposons que l'estimation par sondage du nombre d'unités dans un groupe ne soit égale qu'au tiers du dénombrement au recensement, de sorte que les poids initiaux devraient être multipliés par trois pour corriger le sous-dénombrement. Si la méthode de regroupement de cellules restreint la correction des poids des unités appartenant à ce groupe a un facteur deux, l'estimation par sondage du nombre d'unités dans le groupe ne correspondra qu'aux deux tiers du dénombrement au recensement. En outre, si l'on combine des cellules dont les moyennes sont fort différentes, on risque d'introduire un biais plutôt que d'en corriger un. La correction incomplète de la sous-couverture et le regroupement de cellules ayant des caractéristiques différentes peut causer un biais dans les totaux, les moyennes et d'autres types d'estimation.

Le tableau 1 donne certains exemples de ratios de couverture, c'est-à-dire les estimations par sondage calculées avant la poststratification divisées par les dénombrements au recensement, pour la Current Population Survey réalisée aux États-Unis en mars 2002 et pour la Behavioral Risk Factors Surveillance Survey (BRFSS) réalisée en 2003 dans un ensemble de 44 comtés du sud-ouest des États-Unis. Les ratios de couverture présentés pour le sous-ensemble de groupes démographiques varient de 0,70 à 0,93 pour la CPS, la valeur étant typiquement plus faible pour le groupe des Noirs uniquement que pour les autres groupes. La BRFSS est une enquête téléphonique pour laquelle les taux de réponse sont bas dans l'ensemble de comtés susmentionné et les ratios de couverture sont donc nettement plus faibles que pour la CPS. En outre, les écarts entre les ratios de couverture de divers groupes sont importants. Par exemple, le ratio est de 0,18 pour les hommes hispaniques de 35 à 44 ans, mais il est de 0,37 pour les hommes de la catégorie noir/ multiracial/autre appartenant au même groupe d'âge. Si ces deux groupes étaient fusionnés, la couverture incomplète serait sous-corrigée pour les Hispaniques, mais surcorrigée pour le groupe Noir/multiracial/autre. La National Health Interview Survey (NHIS) de 2003 fournit un autre exemple où les Amérindiens et les Asiatiques sont regroupés avec les Blancs dans les groupes d'âge. Dans la cellule pour le groupe des 25 à 29 ans, par exemple, le taux de couverture des Blancs, des Amérindiens et des Asiatiques était de 0,60, 0,44 et 0,31, respectivement (Tompkins et Kim 2006).

Nous démontrons dans le présent article la faiblesse des procédures courantes de regroupement de cellules et proposons certaines solutions de rechange. À la section 2, nous discutons du biais de certains estimateurs standard en cas de sous-couverture. À la section 3, nous présentons de nouveaux estimateurs qui retiennent une plus grande part de la correction pour la sous-couverture que la méthode standard lorsque les cellules sont regroupées. À la section 4, nous examinons les propriétés empiriques de la méthode standard et des méthodes de rechange proposées à l'aide de simulations. Enfin, à la section 5, nous concluons par un résumé et l'énoncé de certaines possibilités de recherche dans l'avenir.

2. Certains estimateurs standard

L'estimateur de Hájek, l'estimateur poststratifié et l'estimateur poststratifié avec regroupement des post-strates initiales au besoin sont trois estimateurs standard de la moyenne de population. Chacun est défini en détail ci-après. Lorsque le taux de couverture des unités de la population cible par la base de sondage varie, chaque estimateur peut être biaisé. Kim et coll. (2005) donnent certains exemples numériques des effets de la fusion de paires de cellules dont les taux de couverture diffèrent.

Afin d'élaborer la théorie pour d'autres estimateurs de la moyenne, nous modélisons le fait qu'une unité est couverte ou non par la base de sondage et le fait qu'une cellule possède ou non un faible effectif comme des événements aléatoires. Nous définissons trois variables indicatrices : $\delta_k = 1$ si l'unité k est sélectionnée dans l'échantillon et 0 autrement; $c_k = 1$ si l'unité k est couverte par la base de sondage et 0 autrement; $d_i = 1$ si la post-strate i est considérée comme ayant un faible effectif dans un échantillon particulier et 0 autrement $(i = 1, \ldots, I)$. Nous supposons que ces indicateurs sont mutuellement indépendants et que leurs espérances sont $\pi_k, \phi_k,$ et $p_i,$ respectivement. Considérons un plan de sondage probabiliste stratifié à deux degrés. Désignons une strate selon le plan par h; s_h est l'ensemble d'unités primaires d'échantillonnage (UPE) sélectionnées dans la strate h; $s_{hj(i)}$ est l'ensemble d'unités échantillonnées provenant de l'UPE j échantillonnée dans la strate h qui sont également présentes

dans la post-strate i; U_h est la population d'UPE dans la strate h; U_{hj} est la population d'unités dans l'UPE j dans la strate h; et $U_{hj(i)}$ est la population d'unités dans l'UPE j dans la strate h qui sont présentes dans la post-strate i. Pour les besoins de l'analyse exposée à la présente section, il n'est pas nécessaire de spécifier le plan d'échantillonnage plus en détail. Notons qu'aux sections 2 et 3, certaines sommations de la forme $\sum_h \sum_{j\in s_h} \sum_{k\in s_{hj(i)}}$ pour les unités comprises dans la post- strate i pourraient être simplifiées en $\sum_{k\in s_{(i)}}$ sans perte de généralité. Nous avons utilisé la notation plus élaborée afin de montrer clairement comment les degrés d'échantillonnage devraient être traités.

2.1 Estimateur de Hájek

En premier lieu, considérons l'estimateur de Hájek d'une moyenne, qui est

$$\hat{\bar{y}}_\pi = \frac{\left(\sum_{i=1}^{I}\sum_h \sum_{j\in s_h}\sum_{k\in s_{hj(i)}} y_k/\pi_k\right)}{\left(\sum_{i=1}^{I}\sum_h \sum_{j\in s_h}\sum_{k\in s_{hj(i)}} 1/\pi_k\right)} \equiv \hat{T}_\pi / \hat{N}_\pi. \qquad (1)$$

L'espérance de $\hat{T}_\pi$ par rapport à l'échantillonnage et au scénario de couverture est $E_c E_\pi(\hat{T}_\pi) = \sum_{i=1}^{I}\sum_h \sum_{j\in U_h} \sum_{k\in U_{hj(i)}} \phi_k y_k \equiv T^c$, où l'indice supérieur c signifie « couvert ». De même, l'espérance de $\hat{N}_\pi$ est $E_c E_\pi(\hat{N}_\pi) = \sum_{i=1}^{I}\sum_h \sum_{j\in U_h}\sum_{k\in U_{hj(i)}} \phi_k \equiv N^c$. Si nous développons $\hat{\bar{y}}_\pi$ autour de (T^c, N^c), son approximation linéaire est $\hat{\bar{y}}_\pi \doteq T^c/N^c + 1/N^c\,(\hat{T}_\pi - (T^c/N^c)\,\hat{N}_\pi)$. Considérons maintenant le biais de $\hat{\bar{y}}_\pi$ comme un estimateur de $\bar{Y} = \sum_{i=1}^{I} T_i/N$ où T_i est le total pour la population complète d'unités dans la post-strate i (et pas seulement la partie couverte). Après certains calculs, le biais est

$$\text{biais}\,(\hat{\bar{y}}_\pi) \doteq \frac{T^c}{N^c} - \frac{1}{N}\sum_{i=1}^{I} T_i = \frac{C_{\phi y}}{\bar{\phi}} \qquad (2)$$

où $\bar{\phi} = \sum \phi_k/N$, $C_{\phi y} = \sum(\phi_k - \bar{\phi})(y_k - \bar{Y})/N$, $\bar{Y} = T/N$, et $\sum$ dénote la somme sur $i, h, j \in U_h$, et $k \in U_{hj(i)}$. Par conséquent, $\hat{\bar{y}}_\pi$ est biaisé s'il existe toute corrélation entre la variable mesurée, y, et la probabilité de couverture, ϕ_k. Dans (2), le biais est $O(1)$, si bien qu'il reste important même dans les grands échantillons.

Si la probabilité de couverture est la même pour chaque unité de la post-strate i, c'est-à-dire $\phi_k = \phi(i)$ pour tout $k \in U_{hj(i)}$, alors le biais approximatif se réduit à $\text{biais}\,(\hat{\bar{y}}_\pi) \doteq \bar{\phi}^{-1}\sum_i W_i\ (\phi(i) - \bar{\phi})(\bar{Y}_i - \bar{Y})$ où $W_i = N_i/N$ et $\bar{Y}_i = \sum_{h, j\in U_h, k\in U_{hj(i)}} y_k/N_i$. S'il existe une corrélation entre les probabilités de couverture de post-strate et les moyennes de post-strate, l'estimateur de Hájek produira de nouveau un biais qui pourrait être positif ou négatif. Si les taux de couverture ou les moyennes ne varient pas selon la strate, c'est-à-dire $\phi(i) = \phi_o$ ou $\bar{Y}_i = \bar{Y}$, l'estimateur de Hájek sera sans biais, mais les post-strates ne sont habituellement pas formées de cette façon. En outre, le biais existe même si l'échantillonneur ne sait pas quel est l'ensemble approprié de post-strates qui subdivisent la population en groupes ayant des moyennes différentes.

Tableau 1
Ratios de couverture pour la Current Population Survey (CPS) et la Behavioral Risk Factors Surveillance Survey (BRFSS). Blanche uniquement et noire uniquement signifient que seules ces races ont été déclarées par un répondant à la CPS. Le groupe racial résiduel uniquement comprend les cas indiquant une seule race autre que blanche ou noire, ainsi que les cas déclarant deux races ou plus. Dans le tableau pour la BRFSS, les Hispaniques peuvent être de n'importe quelle race

Current Population Survey de mars 2002						
Âge	Blanche uniquement		Noire uniquement		Résiduel uniquement	
	Hommes	Femmes	Hommes	Femmes	Hommes	Femmes
0 à 15 ans	0,93	0,93	0,78	0,79	0,91	0,90
16 à 19ans	0,90	0,88	0,76	0,81	0,93	0,75
20 à 24 ans	0,79	0,85	0,72	0,77	0,75	0,72
25 à 34 ans	0,83	0,89	0,70	0,76	0,76	0,80
Tous âges confondus	0,90	0,92	0,78	0,83	0,85	0,84
BRFSS de 2003 : 44 comtés frontaliers en Arizona, en Californie, au Nouveau-Mexique et au Texas						
Âge	Blanche non-hispanique		Noire, multiraciale, autre		Hispanique	
	Hommes	Femmes	Hommes	Femmes	Hommes	Femmes
18 à 24 ans	0,19	0,26	0,12	0,24	0,15	0,22
25 à 34 ans	0,20	0,31	0,10	0,16	0,19	0,39
35 à 44 ans	0,28	0,31	0,37	0,25	0,18	0,30
Tous âges confondus	0,25	0,31	0,25	0,20	0,18	0,31

Sources : Bureau of the Census (2002), Gonzalez, Town et Kim (2005).

2.2 Moyenne poststratifiée sans regroupement de cellules

La moyenne poststratifiée est définie comme étant $\hat{\bar{y}}_{\mathrm{PS1}} = 1/N\sum_{i=1}^{I}(N_i/\hat{N}_{\pi i})\hat{T}_{\pi i}$ où $\hat{T}_{\pi i}$ et $\hat{N}_{\pi i}$ sont définis comme dans (1), mais en excluant la sommation sur i. Définissons $T_i^c = \sum_{h,\, j\in U_h,\, k\in U_{hj(i)}} \phi_k\, y_k$ et $N_i^c = \sum_{h,\, j\in U_h,\, k\in U_{hj(i)}} \phi_k$. Il s'agit des valeurs prévues (par rapport au scénario de couverture) du total et du nombre d'unités couvertes dans la post-strate i. Si nous développons $\hat{\bar{y}}_{\mathrm{PS1}}$ en série autour de $(T_i^c, N_i^c), i = 1, \ldots, I,$ son approximation linéaire est

$$\hat{\bar{y}}_{\mathrm{PS1}} \doteq \frac{1}{N}\left[\sum_i \frac{N_i}{N_i^c} T_i^c + \sum_i \frac{N_i}{N_i^c}\left(\hat{T}_{\pi i} - \frac{T_i^c}{N_i^c}\hat{N}_{\pi i}\right)\right].$$

Le biais de l'estimateur poststratifié est alors égal au premier terme de cette expression moins $\sum_{i=1}^{I} T_i/N$ et, après certaines manipulations, il peut s'écrire

$$\text{biais}\,(\hat{\bar{y}}_{\mathrm{PS1}}) \doteq \sum_{i=1}^{I} W_i \frac{C_{\phi y i}}{\bar{\phi}_i} \qquad (3)$$

où $\bar{\phi}_i = \sum \phi_k / N_i$, $C_{\phi y i} = \sum(\phi_k - \bar{\phi}_i)(y_k - \bar{Y}_i)/N_i$, et $\sum$ dénote la somme sur $h,\, j \in U_h,$ et $k \in U_{hj(i)}$. Donc, $\hat{\bar{y}}_{\mathrm{PS1}}$ est biaisé en cas de toute corrélation entre la variable y mesurée et la probabilité de couverture ϕ_k dans n'importe quelle post-strate. Si le taux de couverture est constant, à $\phi_k = \phi(i)$ dans la post-strate i, alors l'estimateur poststratifié est approximativement sans biais. De (3), il découle que les post-strates devraient être formées de façon que leurs taux de couverture ou bien les y soient homogènes dans chacune d'elles. Cette condition est comparable aux recommandations faites par Eltinge et Yansaneh (1997), Kalton et Maligalig (1991), ainsi que Little et Vartivarian (2005) pour la création des cellules pour la correction de la non-réponse. Dans les grandes enquêtes, l'ensemble initial de post-strates candidates est souvent plus étendu que ne le permet l'échantillon. Sauf rares exceptions, on regroupe certaines post-strates initiales afin de contrôler les corrections des poids. Habituellement, la raison pour laquelle aucun regroupement n'est effectué est que de petites catégories ont été combinées au préalable en s'appuyant sur l'expérience acquise antérieurement dans le cadre de la même enquête ou d'une enquête comparable. En ce sens, la poststratification sans regroupement de cellules, PS1, n'existe pas vraiment en pratique. La poststratification avec regroupement de cellules, PS2, décrite plus bas est une approche plus fréquente.

2.3 Moyenne poststratifiée avec regroupement de cellules

Examinons maintenant l'estimateur poststratifié avec regroupement de cellules, dans le cadre duquel les cellules à faible effectif sont repérées et combinées à d'autres cellules considérées comme étant leurs plus proches voisines. Il peut arriver que plus d'une cellule à faible effectif soit regroupée avec une cellule particulière dont l'effectif n'est pas faible. Les voisines peuvent être définies de diverses façons. Il pourrait s'agir de cellules ayant les mêmes taux de couverture estimés, $\hat{N}_{\pi i}/N_i$, de cellules qui sont adjacentes selon une caractéristique fondamentale, comme la proximité des classes de revenu, ou de cellules qui ont la même moyenne pour certaines variables étudiées importantes. L'algorithme général de regroupement, étant donné un ensemble initial de cellules, est le suivant :

1) calculer les critères de regroupement pour chaque cellule, c'est-à-dire le FCI, $N_i/\hat{N}_{\pi i}$, et la taille d'échantillon;
2) repérer les cellules à faible effectif, c'est-à-dire celles dont les critères tombent en dehors des limites pour le regroupement;
3) pour chaque cellule à effectif faible, déterminer la voisine à effectif non faible la plus proche et combiner les deux cellules.

La moyenne poststratifiée avec regroupement de cellules est alors donnée par $\hat{\bar{y}}_{\mathrm{PS2}} = 1/N\sum_g (N_g/\hat{N}_{\pi g})\hat{T}_{\pi g}$, où $\hat{T}_{\pi g} = \sum_{i\in A_g}\sum_{h,\, j\in s_h,\, k\in s_{hj(i)}} y_k/\pi_k$ et $\hat{N}_{\pi g}$ est défini de la même façon. Définissons $T_g^c = \sum_{A_g} T_i^c$ et $N_g^c = \sum_{A_g} N_i^c$, avec A_g l'ensemble de post-strates dans le groupe g. Le développement en série de $\hat{\bar{y}}_{\mathrm{PS2}}$ autour de (T_g^c, N_g^c) pour chaque groupe g donne

$$\hat{\bar{y}}_{\mathrm{PS2}} \doteq \frac{1}{N}\left[\sum_g \frac{N_g}{N_g^c} T_g^c + \sum_g \frac{N_g}{N_g^c}\left(\hat{T}_{\pi g} - \frac{T_g^c}{N_g^c}\hat{N}_{\pi g}\right)\right].$$

Il s'ensuit que

$$\text{biais}\,(\hat{\bar{y}}_{\mathrm{PS2}}) \doteq \sum_g W_g \frac{C_{\phi y g}}{\bar{\phi}_g} \qquad (4)$$

avec

$$W_g = N_g/N,\ \bar{\phi}_g = \sum \phi_k / N_g,\ C_{\phi y g} = \sum(\phi_k - \bar{\phi}_g)(y_k - \bar{Y}_g)/N_g,$$

et les sommations en $\bar{\phi}_g$ et $C_{\phi y g}$ sont faites sur $i \in A_g, h,$ $j \in U_h,$ et $k \in U_{hj(i)}$. Si ϕ_k est constant dans le groupe g, cet estimateur est sans biais, mais si $\phi_k = \phi(i),$ c'est-à-dire si le taux de couverture est constant dans la post-strate i, mais qu'il peut différer selon la post-strate, alors le biais devient

$$\text{biais}(\hat{\bar{y}}_{\mathrm{PS2}}) \doteq \sum_g W_g \frac{C^*_{\phi y g}}{\bar{\phi}_g} \qquad (5)$$

avec $\bar{\phi}_g = \sum W_{gi}\phi(i),\ C^*_{\phi y g} = \sum W_{gi}(\phi(i) - \bar{\phi}_g)(\bar{Y}_i - \bar{Y}_g),\ W_{gi} = N_i/N_g,$ et les sommations sont faites sur $i \in A_g$.

Donc, dans le cas où $\hat{\bar{y}}_{\mathrm{PS1}}$ est sans biais, $\hat{\bar{y}}_{\mathrm{PS2}}$ sera biaisé si l'on regroupe des post-strates dont les taux de couverture et les moyennes de population sont différents. Puisque $\bar{\phi}_g$ et $C^*_{\phi yg}$ sont $O(1)$, le biais ne diminue pas à mesure que l'échantillon s'accroît; donc, le carré du biais finira par être la partie dominante de l'erreur quadratique moyenne. Si les cellules sont regroupées, celles qui constituent un même groupe devraient avoir les mêmes taux de couverture, les mêmes moyennes, ou les deux, afin d'éviter le biais.

3. Estimateurs à poids restreints

Nous examinons deux méthodes alternatives de calcul de la pondération en cas de regroupement de post-strates qui s'inscrivent dans la foulée des travaux de Kim (2004). Les deux options sont conçues de façon à obtenir un compromis entre a) l'utilisation de toutes les post-strates et l'éventualité de corrections importantes des poids et b) le regroupement de post-strates produisant des poids moins variables, mais éventuellement des estimations biaisées. Nous les appelons méthodes de *restriction des poids* (WR pour *weight restriction*). Les deux options décrites à la présente section comportent le regroupement de cellules, mais conservent une part plus importante de la correction des poids des cellules individuelles que la méthode standard de regroupement de cellules.

La première option, que nous dénotons PS.WR1, consiste en l'algorithme suivant. Désignons la correction maximale permise des poids par $f_{\max}$, avec $f_{\max} > 1$.

1) Exécuter les étapes 1) à 3) de l'algorithme de la section 2.3 pour PS2.
2) Censurer tout FCI plus grand que $f_{\max}$ à $f_{\max}$ et ajuster chaque poids dans la cellule initiale correspondante à $\tilde{w}_k = w_k f_{\max}$, avec $w_k = 1/\pi_k$. Pour les unités comprises dans les cellules pour lesquelles $\mathrm{FCI} \le f_{\max}$, fixer $\tilde{w}_k = w_k$.
3) Calculer un facteur de correction par regroupement (FCR) pour un groupe *g* sous la forme
$$\tilde{f}_g = N_g \Big/ \sum_{i \in A_g} \sum_{h,\, j \in s_h,\, k \in s_{hj(i)}} \tilde{w}_k .$$
4) Le poids corrigé final est alors $\tilde{w}_k\, \tilde{f}_g$ pour l'unité *k* dans le groupe *g*.

Cette méthode réduira les valeurs les plus grandes de la correction finale des poids de façon à les rendre inférieures aux corrections sans regroupement de cellules, $N_i/\hat{N}_{\pi i}$, quoiqu'il puisse exister un ou plusieurs regroupements pour lesquels le FCR est plus grand que le seuil $f_{\max}$. Le total de contrôle pour le groupe *g*, N_g, est respecté en ce sens que $\sum_{i \in A_g} \sum_{h,\, j \in s_h, k \in s_{hj(i)}} \tilde{w}_k \tilde{f}_g = N_g$, mais les totaux de contrôle pour les cellules individuelles comprises dans A_g ne le sont pas.

Pour analyser les propriétés de PS.WR1, définissons $A_{g,sp}$ et $A_{g,\overline{sp}}$ comme étant les ensembles de post-strates à effectif faible et à effectif non faible dans le groupe de cellules *g*. PS.WR1 peut être exprimé par

$$\hat{\bar{y}}_{\mathrm{PS.WR1}} = \frac{1}{N} \sum_g \frac{N_g}{\hat{N}_{g\mathrm{WR1}}} \hat{T}_{g\mathrm{WR1}}$$

où

$$\hat{T}_{g\mathrm{WR1}} = \sum_{i \in A_{g,sp}} \sum_h \sum_{j \in s_h} \sum_{k \in s_{hj(i)}} f_{\max} w_k y_k + \sum_{i \in A_{g,\overline{sp}}} \sum_h \sum_{j \in s_h} \sum_{k \in s_{hj(i)}} w_k y_k$$

et $\hat{N}_{g\mathrm{WR1}}$ a la même définition avec y_k fixé à 1. L'espérance de $\hat{T}_{g\mathrm{WR1}}$ en fonction de la couverture, du degré de faible effectif (*sparseness*) et du plan d'échantillonnage est $E_c E_{sp} E_\pi(\hat{T}_{g\mathrm{WR1}}) = T_g^c + (f_{\max} - 1)\, \tilde{T}_g^c$ où $\tilde{T}_g^c = \sum_{i \in A_g} p_i T_i^c$. De même, $E_c E_{sp} E_\pi(\hat{N}_{g\mathrm{WR1}}) = N_g^c + (f_{\max} - 1)\, \tilde{N}_g^c$ avec $\tilde{N}_g^c = \sum_{i \in A_g} p_i N_i^c$. $\hat{\bar{y}}_{\mathrm{PS.WR1}}$ peut être développé en série autour des espérances de $(\hat{T}_{g\mathrm{WR1}}, \hat{N}_{g\mathrm{WR1}})$. Après certaines manipulations, le biais approximatif de $\hat{\bar{y}}_{\mathrm{PS.WR1}}$ devient

$$\mathrm{biais}\,(\hat{\bar{y}}_{\mathrm{PS.WR1}}) \doteq \sum_g W_g \frac{C_{\alpha\phi,\, y,\, g}}{(\overline{\alpha\phi})_g} \qquad (6)$$

où

$$(\overline{\alpha\phi})_g = \sum \alpha_i \phi_k / N_g,\ \alpha_i = 1 + (f_{\max} - 1) p_i,$$
$$C_{\alpha\phi, y, g} = \sum (\alpha_i \phi_k - (\overline{\alpha\phi})_g)(y_k - \bar{Y}_g)/N_g,$$

et les sommations sont faites sur $i \in A_g, h, j \in U_h$, et $k \in U_{hj(i)}$. Dans le cas d'une probabilité de couverture commune dans la post-strate *i*, c'est-à-dire $\phi_k = \phi(i)$, nous avons

$$(\overline{\alpha\phi})_g = \bar{\phi}_g + \left(f_{\max} - 1\right)(\overline{p\phi})_g$$

et

$$\alpha_i \phi(i) - (\overline{\alpha\phi})_g = (\phi(i) - \bar{\phi}_g) + (f_{\max} - 1)(p_i \phi(i) - (\overline{p\phi})_g)$$

où $(\overline{p\phi})_g = \sum_{A_g} W_{gi} p_i \phi(i)$. Il découle de cela que

$$C_{\alpha\phi,\, y,\, g} = C^*_{\phi yg} + (f_{\max} - 1) C_{p\phi,\, y,\, g}$$

avec

$$C_{p\phi,\, y,\, g} = \sum_{A_g} W_{gi} (p_i \phi(i) - (\overline{p\phi})_g)(\bar{Y}_i - \bar{Y}_g).$$

Si les moyennes de cellules $\bar{Y}_i$ sont toutes égales dans un groupe, alors $C^*_{\phi yg} = C_{p\phi,\, y,\, g} = 0$ et $\hat{\bar{y}}_{\mathrm{PS.WR1}}$ sera approximativement sans biais. Dans le cas particulier où la couverture est constante dans un groupe de cellules, c'est-à-dire $\phi(i) = \bar{\phi}_g$, alors $C_{\alpha\phi, y, g} = (f_{\max} - 1)\, \bar{\phi}_g \sum_{i \in A_g} W_{gi} (p_i - \bar{p}_g)(\bar{Y}_i - \bar{Y}_g)$ avec $\bar{p}_g = \sum_{A_g} W_{gi} p_i$. Donc, si p_i et $\phi(i)$ sont constantes dans A_g, alors $\hat{\bar{y}}_{\mathrm{PS.WR1}}$ sera presque sans biais, même si les moyennes de cellules $\bar{Y}_i, i \in A_g$

diffèrent. Il est presque certain que cette condition ne sera pas vérifiée aussi longtemps qu'une post-strate-strate d'un groupe présente une probabilité d'avoir un faible effectif sensiblement différente de celle des autres.

Dans le cas d'une probabilité de couverture commune dans la post-strate i, $\phi(i)$, nous pouvons aussi comparer le biais de l'estimateur à cellules regroupées, $\hat{\bar{y}}_{\mathrm{PS2}}$, à celui de $\hat{\bar{y}}_{\mathrm{PS.WR1}}$. En utilisant les résultats du paragraphe précédent, dans (6), le biais peut être exprimé comme étant

$$\mathrm{biais}(\hat{\bar{y}}_{\mathrm{PS.WR1}}) \doteq \sum_g W_g \left[\frac{C^*_{\phi yg}/\bar{\phi}_g + (f_{\max}-1)C_{p\phi,y,g}/\bar{\phi}_g}{1+(f_{\max}-1)(\overline{p\phi})_g/\bar{\phi}_g} \right].$$

Puisque $1+(f_{\max}-1)(\overline{p\phi})_g/\bar{\phi}_g \geq 1$, nous pouvons utiliser (5) pour obtenir

$$\begin{aligned} &\left|\mathrm{biais}(\hat{\bar{y}}_{\mathrm{PS.WR1}})\right| \\ &\leq \left|\sum_g W_g \left[C^*_{\phi y,g}/\bar{\phi}_g + (f_{\max}-1)C_{p\phi,y,g}/\bar{\phi}_g \right]\right| \\ &= \left|\mathrm{biais}(\hat{\bar{y}}_{\mathrm{PS2}}) + (f_{\max}-1)\sum_g W_g C_{p\phi,y,g}/\bar{\phi}_g\right|. \quad (7) \end{aligned}$$

Si $p_i\phi(i)$ et $\bar{Y}_i$ ne sont pas corrélées, le biais absolu de $\hat{\bar{y}}_{\mathrm{PS.WR1}}$ est inférieur ou égal à celui de $\hat{\bar{y}}_{\mathrm{PS2}}$, parce que $1+(f_{\max}-1)(\overline{p\phi})_g/\bar{\phi}_g \geq 1$. Si $p_i\phi(i)$ et $\bar{Y}_i$ sont corrélées, il faut considérer deux cas : i) $\mathrm{biais}(\hat{\bar{y}}_{\mathrm{PS2}}) \geq 0$ et ii) $\mathrm{biais}(\hat{\bar{y}}_{\mathrm{PS2}}) < 0$. Dans le premier, la dernière ligne de (7) sera inférieure ou égale au biais absolu de $\hat{\bar{y}}_{\mathrm{PS2}}$ si

$$\frac{-2\left|\mathrm{biais}(\hat{\bar{y}}_{\mathrm{PS2}})\right|}{f_{\max}-1} \leq \sum_g W_g C_{p\phi,y,g}/\bar{\phi}_g \leq 0.$$

Dans le cas ii), l'exigence est que

$$0 \leq \sum_g W_g C_{p\phi,y,g}/\bar{\phi}_g \leq \frac{2\left|\mathrm{biais}(\hat{\bar{y}}_{\mathrm{PS2}})\right|}{f_{\max}-1}.$$

Si la covariance entre les probabilités d'avoir un faible effectif et d'être couvert, $p_i\phi(i)$, et les moyennes de cellules, $\bar{Y}_i$, est faible dans tous les groupes et de signe opposé au biais $(\hat{\bar{y}}_{\mathrm{PS2}})$, alors $\hat{\bar{y}}_{\mathrm{PS.WR1}}$ aura un biais plus faible que $\hat{\bar{y}}_{\mathrm{PS2}}$.

La deuxième option, dénotée PS.WR2, a pour but d'exercer un plus grand contrôle sur la grandeur de la correction finale des poids que ne le fait l'option PS.WR1. Dans le cas de cette dernière, la correction finale peut être plus grande que $f_{\max}$. L'option PS.WR2 vise à limiter la correction finale à $f_{\max}=2$ ou à un autre maximum fixé d'avance. L'idée générale est de déterminer d'abord quelles cellules devraient être regroupées, comme dans le cas de PS.WR1. Ensuite, les poids dans les cellules à faible effectif sont multipliés par $f_{\max}$. Ceux dans la cellule à effectif non faible d'un regroupement sont ajustés par un facteur constant afin de faire concorder le chiffre estimé de population dans le groupe au dénombrement de contrôle. L'algorithme détaillé de calcul des poids pour l'option PS.WR2 est le suivant :

1) Exécuter les étapes (1) à (3) de l'algorithme de la section 2.3 pour PS2.
2) Dans un groupe contenant au moins une cellule à effectif non faible, calculer le total de contrôle dans le groupe g comme étant $N_g = \sum_{i\in A_g} N_i$ et le poids corrigé pour toutes les unités k dans $A_{g,sp}$ comme étant $\tilde{w}_k = w_k f_{\max}$.
3) Calculer le poids corrigé pour toutes les unités k dans $A_{g,\overline{sp}}$ comme étant $\tilde{w}_k = w_k(N_g - \hat{\tilde{N}}_{g,sp})/\hat{N}_{g,\overline{sp}}$ où $\hat{N}_{g,\overline{sp}} = \sum_{i\in A_{g,\overline{sp}}}\sum_{h,\,j\in s_h,k\in s_{hj(i)}} w_k$ et $\hat{\tilde{N}}_{g,sp} = \sum_{i\in A_{g,sp}} \sum_{h,\,j\in s_h,k\in s_{hj(i)}} \tilde{w}_k$.
4) Le poids corrigé final est alors $\tilde{w}_k$ pour l'unité k dans le groupe g.

Ce deuxième estimateur à poids restreints peut s'écrire sous la forme $\hat{\bar{y}}_{\mathrm{PS.WR2}} = (1/N)\sum_g \hat{T}_{g\mathrm{WR2}}$, où

$$\begin{aligned} \hat{T}_{gWR2} &= \sum_{i\in A_{g,sp}} \sum_{h,\,j\in s_h,k\in s_{hj(i)}} f_{\max} w_k y_k \\ &\quad + \sum_{i\in A_{g,\overline{sp}}} \sum_{h,\,j\in s_h,k\in s_{hj(i)}} \frac{N_g - f_{\max}\hat{N}_{g,sp}}{\hat{N}_{g,\overline{sp}}} w_k y_k \end{aligned}$$

où $\hat{N}_{g,sp} = \sum_{i\in A_{g,sp}}\sum_{h,\,j\in s_h,k\in s_{hj(i)}} w_k$. L'espérance de $\hat{T}_{g\mathrm{WR2}}$ en fonction de la couverture, du degré de faible effectif et du plan d'échantillonnage est

$$E_c E_{sp} E_\pi(\hat{T}_{g\mathrm{WR2}}) = f_{\max}\tilde{T}^c_g + \frac{N_g - f_{\max}\tilde{N}^c_g}{N^c_g - \tilde{N}^c_g}(T^c_g - \tilde{T}^c_g)$$

où $\tilde{N}^c_g = \sum_{i\in A_g} p_i N^c_i$. Après certains calculs, le biais approximatif de $\hat{\bar{y}}_{\mathrm{PS.WR2}}$ peut s'écrire

$$\begin{aligned} &\mathrm{biais}(\hat{\bar{y}}_{\mathrm{PS.WR2}}) \\ &\doteq \frac{1}{N} f_{\max} \sum_g \tilde{N}^c_g(\mu^c_{g,sp} - \mu^c_{g,\overline{sp}}) \\ &\quad + \sum_g W_g \frac{1}{\sum_{A_g} q_i N^c_i} \\ &\quad \times \left[\sum_{A_g} q_i T^c_i - N_g^{-1}\left(\sum_{A_g} q_i N^c_i\right)\left(\sum_{A_g} T_i\right)\right] \end{aligned}$$

où $\mu^c_{g,sp} = \tilde{T}^c_g \big/ \tilde{N}^c_g$ et $\mu^c_{g,\overline{sp}} = \sum_{A_g} q_i T^c_i / \sum_{A_g} q_i N^c_i$. Maintenant, notons que dans le cas d'une probabilité de couverture commune dans la cellule i, $\phi_k = \phi(i)$,

$$\begin{aligned} &\sum_{A_g} q_i T^c_i - N_g^{-1}\left(\sum_{A_g} q_i N^c_i\right)\left(\sum_{A_g} T_i\right) \\ &= \sum_{A_g}\sum_{h,\,j\in U_h,k\in U_{hj(i)}} (q_i\phi_k - (\overline{q\phi})_g)(y_k - \bar{Y}_g) \\ &= \sum_{A_g} N_i(q_i\phi(i) - (\overline{q\phi})_g)(\bar{Y}_i - \bar{Y}_g) \\ &\equiv N_g(C^*_{\phi yg} - C_{p\phi,y,g}) \end{aligned}$$

où $q_i = 1 - p_i$, $(\overline{q\phi})_g = \sum q_i \phi_k / N_g$, et où $C^*_{\phi yg}, C_{p\phi,y,g}$ ont été définis antérieurement. Ensuite, utilisons le fait que $\sum_{A_g} q_i N_i^c = N_g^c - \tilde{N}_g^c$ pour définir $P_g^c = N_g^c / N_g$, la proportion d'unités couvertes dans le groupe g, et $\tilde{P}_g^c = \tilde{N}_g^c / N_g$, la proportion d'unités couvertes prévue dans les cellules à faible effectif dans le groupe g. Alors, le biais peut aussi s'écrire sous la forme

$$\text{biais}(\hat{\bar{y}}_{\text{PS.WR2}}) \doteq \frac{1}{N} f_{\max} \sum_g \tilde{N}_g^c (\mu_{g,sp}^c - \mu_{g,\overline{sp}}^c) + \sum_g W_g \frac{C^*_{\phi yg} - C_{p\phi,y,g}}{P_g^c - \tilde{P}_g^c}. \quad (8)$$

À en juger par l'expression (8), $\hat{\bar{y}}_{\text{PS.WR2}}$ sera approximativement sans biais si la moyenne par unité pour les unités couvertes par la base de sondage dans chaque cellule regroupée est la même dans les cellules à faible effectif, $\mu_{g,sp}^c$, que dans les cellules à effectif non faible, c'est-à-dire $\mu_{g,sp}^c = \mu_{g,\overline{sp}}^c$, et que les covariances, $C^*_{\phi yg}$ et $C_{p\phi,y,g}$, sont toutes deux nulles. La dernière condition est réalisée en combinant des cellules ayant les mêmes moyennes, $\bar{Y}_i$. Le fait de combiner des cellules dont les taux de couverture sont égaux ne produira pas un $\hat{\bar{y}}_{\text{PS.WR2}}$ sans bais. La situation est plus limitée que pour $\hat{\bar{y}}_{\text{PS.WR1}}$, qui est sans biais si les taux de couverture ou les moyennes sont les mêmes dans toutes les cellules d'un groupement.

4. Une études empirique

Afin d'évaluer certaines idées présentées plus haut, nous avons réalisé une étude par simulation des propriétés de biais des diverses méthodes de poststratification proposée. Nous avons aussi examiné les propriétés d'un estimateur de variance qui est souvent utilisé en pratique.

4.1 Population étudiée

La population utilisée dans la simulation a été extraite du fichier de microdonnées à grande diffusion de la National Health Interview Survey (NHIS) de 2003. Nous avons créé un sous-ensemble du fichier de la NHIS contenant 21 664 personnes. Les strates et les UPE étaient fondées sur celles du fichier de données à grande diffusion de la NHIS, mais nous avons regroupé les strates par trois pour créer de nouvelles strates d'échantillonnage pour la population étudiée. Nous avons utilisé quatre variables binaires (caractéristiques ayant une valeur 0 ou 1) pour la simulation, fondées chacune sur une réponse autodéclarée par la personne, soit :

couverture par une assurance-maladie, c'est-à-dire si une personne était couverte ou non par tout type d'assurance-maladie;

limitations physiques, mentales ou émotionnelles, c'est-à-dire si une personne était limitée ou non de n'importe laquelle de ces façons;

retard des soins médicaux, c'est-à-dire si une personne avait retardé ou non des soins médicaux à cause de leur coût au cours des 12 derniers mois;

séjour d'au moins une nuit à l'hôpital, c'est-à-dire si une personne avait passé ou non une nuit à l'hôpital au cours des 12 derniers mois.

Le tableau 2 donne les pourcentages de personnes présentant ces quatre caractéristiques dans les cellules formées par l'âge et le sexe. Ces 16 cellules (âge × sexe) représentent l'ensemble initial de post-strates utilisées pour l'estimation. Les pourcentages peuvent varier considérablement d'une cellule à l'autre, selon les caractéristiques. Ainsi, les personnes de 18 à 24 ans sont nettement plus susceptibles que les autres de ne pas être couvertes par une assurance-maladie, tandis que les enfants de moins de cinq ans et les personnes de 65 ans et plus sont beaucoup plus susceptibles que les autres d'avoir été hospitalisés. Le regroupement de cellules dont les moyennes, ou les proportions dans ce cas-ci, diffèrent peut introduire un biais, comme nous l'avons mentionné plus haut.

Nous avons également créé une variable binaire artificielle ayant une moyenne commune de 0,20, quelle que soit l'appartenance de l'unité à la post-strate. Dans ces conditions, tous les estimateurs, y compris celui de Hájek, seront sans biais indépendamment des taux de couverture. En outre, l'idée conventionnelle selon laquelle le regroupement de cellules peut réduire les variances en lissant les corrections extrêmes des poids pourrait s'avérer vraie pour cette variable.

4.2 Plan d'échantillonnage

Deux UPE ont été sélectionnées dans chaque strate avec probabilité proportionnelle à la taille (PPT), la taille étant le nombre de personnes dans chaque UPE. L'échantillonnage des UPE a été fait avec remise afin de simplifier l'estimation de la variance. Si nous avions utilisé un échantillonnage sans remise, une méthode plus élaborée de sélection et d'estimation de la variance aurait été nécessaire (voir, par exemple, Särndal, Swensson et Wretman 1992, chapitre 3). Dans chaque UPE échantillonnée, 20 personnes ont été sélectionnées par échantillonnage aléatoire simple sans remise pour donner un total de 1 000 personnes dans chaque échantillon. Pour chaque combinaison de paramètres discutée plus loin, nous avons sélectionné 2 000 échantillons.

Nous avons utilisé 16 post-strates initiales obtenues par le croisement des huit groupes d'âge présentés au tableau 2 avec le sexe. Dans chaque échantillon, nous avons calculé

les estimateurs des proportions de population décrits aux sections 2 et 3, c'est-à-dire l'estimateur de Hájek, $\hat{\bar{y}}_{\pi}$, l'estimateur poststratifié $\hat{\bar{y}}_{\mathrm{PS1}}$, qui utilise l'ensemble des 16 post-strates, l'estimateur poststratifié avec regroupement de cellules, $\hat{\bar{y}}_{\mathrm{PS2}}$, et les deux estimateurs à poids restreint $\hat{\bar{y}}_{\mathrm{PS.WR1}}$ et $\hat{\bar{y}}_{\mathrm{PS.WR2}}$. Le code de simulation a été rédigé dans le langage R (R Development Core Team 2005) en se servant intensivement du progiciel pour données d'enquête R (Lumley 2004, 2005).

4.3 Scénarios de couverture

Cinq ensembles de scénarios de couverture, présentés au tableau 3, ont été utilisés pour filtrer la population avant d'échantillonner les UPE. Le ratio de couverture variait selon la post-strate et différait pour chacune des cinq caractéristiques pour lesquelles les proportions ont été estimées. Les ratios de couverture particuliers à chacune des cinq caractéristiques sont nommés de C1 à C5 dans le tableau 3. Ils ont été créés artificiellement d'après les moyennes de population pour chaque groupe âge-sexe. Une moins bonne couverture a été attribuée au groupe comportant une portion plus élevée d'unités possédant les caractéristiques de couverture par une assurance-maladie et de limitations fonctioinnelles; la situation était inverse pour le retard des soins médicaux et l'hospitalisation. Dans le scénario C5, les ratios de couverture sont assez différents afin de produire des corrections de la couverture qui varient sensiblement parmi dans l'ensemble initial de 16 post-strates. Quoique les taux présentés au tableau 3 soient faibles, ils sont comparables à ceux donnés pour la BRFSS au tableau 1, ou plus élevés que ceux-ci. En appliquant ces taux, nous avons sélectionné aléatoirement un sous-ensemble de la population devant figurer dans la base de sondage pour chaque échantillon qui avait été sélectionné. Par exemple, si le ratio de couverture dans la post-strate des garçons de moins de 5 ans était de 0,9, 90 % de la population de cette post-strate ont été sélectionnés aléatoirement pour demeurer dans la base de sondage, tandis que pour le reste de cette population, la probabilité d'être échantillonné était nulle.

4.4 Règles de regroupement

Nous avons établi des situations où les conditions pour l'absence de biais décrites aux sections 2 et 3 pouvaient être violées lorsque des cellules étaient regroupées dans les simulations. Chacun des estimateurs, $\hat{\bar{y}}_{\mathrm{PS2}}$, $\hat{\bar{y}}_{\mathrm{PS.WR1}}$ et $\hat{\bar{y}}_{\mathrm{PS.WR2,}}$ comporte un regroupement de cellules. Si le FCI (facteur de poststratification) dans une post-strate initiale, $N_i / \hat{N}_{\pi i}$, est supérieur à la correction maximale permise, $f_{\max}$, ou que la taille de l'échantillon de cellule est inférieure à un minimum fixé, $n_{i,\min}$, nous disons que cette post-strate est une cellule « à faible effectif » et nous la regroupons avec une cellule voisine. Nous avons utilisé deux méthodes de sélection des voisines, appelées ici « adjacence » et « moyenne proche ».

Dans le regroupement fondé sur l'adjacence, les voisines d'une cellule particulière sont définies comme étant les cellules qui lui sont horizontalement ou verticalement adjacentes dans le tableau âge × sexe. Par exemple, dans le tableau abrégé qui suit, les voisines de la cellule 3 sont les cellules ombrées 2, 4 et 7.

1	5
2	6
3	7
4	8

Dans le cas de la méthode d'adjacence, une cellule à faible effectif était regroupée à la voisine ayant le plus faible facteur de poststratification. Dans le cas de la méthode de la moyenne proche, une cellule à faible effectif était regroupée avec la cellule à effectif non faible dont la moyenne d'échantillon non pondérée était la plus proche de la sienne.

Tableau 2 Pourcentage de personnes possédant quatre caractéristiques liées à la santé dans les groupes formés selon l'âge et le sexe

	Chiffres de population			Pourcentage de personnes possédant la caractéristique											
				Non couverte par une assurance-maladie			Limitations physiques, mentales, émotionnelles			Retard des soins médicaux au cours des 12 derniers mois			Hospitalisation au cours des 12 derniers mois		
Âge	Hommes	Femmes	Total	Hommes	Femmes	Total	Hommes	Femmes	Total	Hommes	Femmes	Total	Hommes	Femmes	Total
< 5 ans	843	795	1 638	10	9	9	4	3	3	3	4	3	17	15	16
5 à 17 ans	2 271	2 082	4 353	13	14	13	10	6	8	4	4	4	2	1	2
18 à 24 ans	998	1 031	2 029	37	31	34	4	4	4	8	11	9	3	14	8
25 à 44 ans	2 971	3 207	6 178	28	23	25	7	7	7	9	10	9	3	10	6
45 à 64 ans	2 421	2 597	5 018	14	14	14	16	19	18	7	11	9	8	10	9
65 à 69 ans	305	384	689	2	1	2	24	29	27	3	8	6	15	14	14
70 à 74 ans	275	344	619	1	1	1	34	32	33	2	5	4	18	15	17
75 ans et plus	423	717	1 140	1	1	1	41	48	45	2	2	2	22	22	22
Total	10 507	11 157	21 664	18	16	17	12	13	13	6	8	7	7	10	8

Tableau 3 Ratios de couverture utilisés dans les simulations

	C1: Non couverte par une assurance-maladie		C2: Limitations physiques, mentales, émotionnelles		C3: Retard des soins médicaux au cours des 12 derniers mois		C4: Hospitalisation au cours des 12 derniers mois		C5: Moyenne commune *Y*	
Âge	Hommes	Femmes	Hommes	Femmes	Hommes	Femmes	Hommes	Femmes	Hommes	Femmes
< 5 ans	0,9	0,9	0,9	0,9	0,5	0,5	0,8	0,8	0,9	0,8
5 à 17 ans	0,8	0,8	0,9	0,6	0,5	0,5	0,5	0,5	0,7	0,2
18 à 24 ans	0,5	0,5	0,8	0,8	0,8	0,8	0,5	0,8	0,4	0,4
25 à 44 ans	0,5	0,5	0,8	0,8	0,8	0,8	0,5	0,8	0,6	0,5
45 à 64 ans	0,8	0,8	0,5	0,6	0,8	0,8	0,5	0,8	0,3	0,8
65 à 69 ans	0,9	0,9	0,5	0,6	0,5	0,5	0,8	0,8	0,4	0,4
70 à 74 ans	0,9	0,9	0,5	0,5	0,5	0,5	0,8	0,8	0,2	0,7
75 ans et plus	0,9	0,9	0,5	0,5	0,5	0,5	0,8	0,8	0,8	0,9

Nous avons utilisé deux valeurs différentes de $f_{\max}$ dans les simulations, soit $f_{\max} = 2$ et 1,8. L'utilisation de $f_{\max} = 1,8$ donne lieu à un plus grand nombre de regroupements de cellules que $f_{\max} = 2$ et à un plus grand nombre des biais (pour les autres caractéristiques que celle, artificielle, d'une moyenne commune) mentionnés aux sections 2 et 3 et causés par la combinaison de cellules possédant des moyennes ou des taux de couverture différents. La taille minimale de cellule a été fixée à $n_{i,\min} = 25$. Naturellement, en pratique, de nombreuses variations sont utilisées pour décider des combinaisons de cellules qui sont permises. Nous n'avons utilisé que deux des possibilités en guise d'illustration dans la simulation.

Une fois que toutes les cellules à faible effectif et leurs voisines sont repérées, le processus de regroupement se déroule séquentiellement à partir de la première cellule. Dans le cas d'une enquête comptant un grand nombre de post-strates possibles définies d'avance, il pourrait être nécessaire d'effectuer ces procédures itérativement afin d'éliminer toutes les cellules à faible effectif. Dans la présent simulation, nous n'avons procédé qu'à un seul cycle de regroupement.

4.5 Estimation de la variance

Pour chacun des estimateurs de proportion, nous avons calculé une estimation de la variance par linéarisation. Chaque estimateur de la variance est basé sur la méthode de substitution linéaire (par exemple, voir Särndal et coll. 1992, chapitre 5). Nous avons calculé des estimations de la variance pour tous les estimateurs de proportion en utilisant les fonctions `svydesign`, `poststratify` et `svymean` du progiciel pour données d'enquête R. L'approche théorique générale consiste à faire une approximation linéaire d'un estimateur particulier. L'approximation linéaire est réarrangée de façon que l'estimateur soit écrit sous la forme d'une somme de totaux pondérés d'UPE et l'estimateur de variance utilisé est celui applicable à l'échantillonnage d'UPE avec remise. Les estimateurs $\hat{\bar{y}}_{\mathrm{PSI}}$, $\hat{\bar{y}}_{\mathrm{PS2}}$, $\hat{\bar{y}}_{\mathrm{PS.WR1}}$ et $\hat{\bar{y}}_{\mathrm{PS.WR2}}$ sont traités comme des estimateurs poststratifiés standard pour les besoins de l'estimation de la variance. Pour $\hat{\bar{y}}_{\mathrm{PSI}}$, définissons les éléments suivants :

$u_k =$ $N_i / \hat{N}_{i\pi}(y_k - \hat{\bar{y}}_i), k \in s_i,$ avec $\hat{\bar{y}}_i = \hat{T}_{i\pi}/\hat{N}_{i\pi}$ et s_i l'ensemble de toutes les unités d'échantillon dans la post-strate i; u_k est appelé substitut linéaire;

$$\breve{u}_{hj+} = \sum_{i,k\in s_{hj(i)}} w_k u_k;$$

$$\bar{\breve{u}}_{h++} = \frac{1}{n_h}\sum_{j\in s_h} \breve{u}_{hj+}$$

L'estimateur de la variance de $\hat{\bar{y}}_{\mathrm{PSI}}$ est alors

$$v(\hat{\bar{y}}_{\mathrm{PSI}}) = \frac{1}{N^2}\sum_h \frac{n_h}{n_h - 1}\sum_{j\in s_h}(\breve{u}_{hj+} - \bar{\breve{u}}_{h++})^2. \qquad (9)$$

Pour l'estimateur à strates regroupées, $\hat{\bar{y}}_{\mathrm{PS2}}$, nous appliquons l'expression (9) en définissant le substitut linéaire comme étant

$$u_k = \frac{N_g}{\hat{N}_{g\pi}}(y_k - \hat{\bar{y}}_g), k \in s_g,$$

avec $\hat{\bar{y}}_g = \hat{T}_{g\pi}/\hat{N}_{g\pi}$ et s_g l'ensemble de toutes les unités d'échantillon dans le groupe g.

Pour PS.WR1 et PS.WR2, nous calculons les poids finals comme il est décrit à la section 3 et appelons la fonction `poststratify` dans R. Le substitut linéaire est alors calculé comme

$$u_k = \frac{N_g}{\hat{N}_g}(y_k - \hat{\bar{y}}_g) = y_k - \hat{\bar{y}}_g$$

parce que $\hat{N}_g = \sum_{i\in A_g}\sum_{k\in s_i}\tilde{w}_k = N_g$. La moyenne $\hat{\bar{y}}_g$ est calculée comme

$$\begin{aligned}\hat{\bar{y}}_g &= \sum_{i\in A_g}\sum_{k\in s_i}\tilde{w}_k y_k \Big/ \sum_{i\in A_g}\sum_{k\in s_i}\tilde{w}_k \\ &= \sum_{i\in A_g}\sum_{k\in s_i}\tilde{w}_k y_k \Big/ N_g.\end{aligned}$$

Le substitut linéaire pondéré est alors $u_k = \tilde{w}_k(y_k - \hat{\bar{y}}_g)$ et $\breve{u}_{hj+} = \sum_{i,k\in s_{hj(i)}}\tilde{w}_k u_k$.

Dans les cas de PS2, PS.WR1 et PS.WR2, ces estimateurs de variance ne tiennent pas compte de la nature dynamique du regroupement de cellules qui peut varier d'un échantillon à l'autre. Par conséquent, l'une des sources de variation n'est pas prise en compte et nous pouvons nous attendre à ce que les estimations de la variance soient un peu trop faibles comparativement aux variances empiriques obtenues par simulation.

4.6 Résultats des simulations

Les tableaux 4 à 7 résument les résultats pour les erreurs de correction de la couverture, les biais relatifs des proportions estimées, les variances des divers estimateurs étudiés et la couverture des intervalles de confiance calculée en utilisant les estimateurs de variance par linéarisation. Le tableau 4 donne l'erreur absolue moyenne de correction de la couverture, définie comme étant

$$\bar{e} = (DI)^{-1} \sum_{d=1}^{D} \sum_{i=1}^{I} \left| \hat{N}_{di} / N_i - 1 \right| \qquad (10)$$

où d est l'un des $D = 2000$ échantillons et $\hat{N}_{di}$ est le nombre estimé d'unités dans la post-strate i en se basant sur les poids finals pour un estimateur particulier (Hájek, PS2, PS.WR1 ou PS.WR2). La valeur de $\bar{e}$ est 0 pour l'estimateur poststratifié sans regroupement de cellules, PS1, puisqu'il corrige entièrement l'erreur de couverture dans chacune des 16 post-strates. Pour illustrer comment les erreurs moyennes de correction de la couverture peuvent varier, nous avons estimé les proportions pour les variables d'assurance-maladie et de moyenne commune Y en utilisant les ratios de couverture de la base de sondage C1 et C5. Pour la plupart des combinaisons de ratio de couverture, méthode de regroupement et seuil de correction, PS.WR1 corrige plus efficacement l'erreur de couverture que l'estimateur par regroupement standard, PS2. Par exemple, $\bar{e} = 0,086$ dans le cas de PS.WR1 pour (assurance-maladie, regroupement par adjacence, $f_{\max} = 2$), tandis que $\bar{e} = 0,120$ dans le cas de PS2. En revanche, la correction de la couverture est un peu moins bonne si l'on utilise PS.WR2 plutôt que PS2.

Le tableau 5 donne les biais relatifs, définis comme étant $100 \sum_{d=1}^{D} (\hat{\bar{y}}_d - \bar{Y}) / \bar{Y}$, où $\hat{\bar{y}}_d$ est l'une des estimations de la proportion pour l'échantillon d. Les estimations de Hájek sont fortement biaisées pour les quatre premières caractéristiques, puisqu'elles ne comportent aucune correction pour les différences de sous-couverture entre les cellules. Les biais relatifs varient de -12,1 % pour les limitations fonctionnelles à 13,4 % pour l'hospitalisation. Comme nous l'avons mentionné à la section 2, le biais peut être positif ou négatif, en fonction de la corrélation des taux de couverture et des moyennes de cellule.

La poststratification sans regroupement de cellules (PS1) produit des estimations presque sans biais, tandis que les autres options - PS2, PS.WR1 et PS.WR2 -introduisent toutes un biais lorsqu'on utilise le regroupement par adjacence pour les quatre premières caractéristiques. Le nombre de post-strates après le regroupement, donné au tableau 5, varie de 6 à 16 quand $f_{\max} = 2$ et de 5 à 13 quand $f_{\max} = 1,8$. Le biais relatif de PS2, si l'on utilise le regroupement par adjacence, varie de -4,4 % à 6,2 % quand $f_{\max} = 2$ et de -6,5 % à 9,4 % quand $f_{\max} = 1,8$. Dans le cas du regroupement par adjacence, les options de rechange, PS.WR1 et PS.WR2, produisent des biais compris entre ceux de PS1 (aucun regroupement) et de PS2. En particulier, PS.WR1 donne des résultats raisonnablement comparables à ceux de PS1 en ce qui concerne le biais avec regroupement par adjacence. En revanche, le regroupement selon la moyenne proche produit des estimations PS2, PS.WR1 et PS.WR2 essentiellement sans biais quand $f_{\max} = 2$. Dans le cas du regroupement selon la moyenne et $f_{\max} = 1,8$, les estimateurs PS2 et PS.WR2 présentent encore un léger biais, mais PS.WR1 se compare bien à PS1. Pour la cinquième caractéristique (moyenne commune Y), tel que prévu, tous les estimateurs sont presque sans biais, quelle que soit la méthode de regroupement des cellules.

L'une des justification conventionnelle du regroupement des cellules est que les poids extrêmes seront réduits et que les variances des estimations seront, à leur tour, plus faibles. Le tableau 6 donne les ratios des variances empiriques des proportions estimées par rapport à la variance de PS1. L'estimateur de Hájek donne des estimations dont les variances sont de 12 % à 18 % plus faibles que celles des estimations produites par PS1 pour l'assurance-maladie et les limitations fonctionnelles, mais qui sont plus variables que ces dernières pour le retard des soins médicaux et l'hospitalisation. Ces résultats montrent, en outre, clairement que le regroupement de cellules peut accroître ou réduire la variance d'un estimateur poststratifié. L'utilisation de PS2 produit de légères améliorations de la variance pour certaines combinaisons des quatre premières variables, mais sous le scénario (adjacence, $f_{\max} = 2$), la variance de PS2 pour l'hospitalisation est 17 % plus grande que si l'on utilise PS1. Dans le cas de (adjacence, $f_{\max} = 1,8$), PS2 donne des estimations 23 % plus variables pour l'hospitalisation. PS.WR1 ne produit pas de variances extrêmes de PS2 dans le cas du regroupement par adjacence; comme PS2, PS.WR2 possède une plus grande variance pour l'hospitalisation sous regroupement par adjacence. Si l'on recourt au regroupement selon la moyenne proche plutôt que l'adjacence, les variances de PS2, PS.WR1 et PS.WR2 sont sensiblement plus proches de celles de PS1. Toutefois, pour la variable de moyenne commune Y, le regroupement réduit systématiquement la variance. La réduction est de près de 20 % pour le regroupement par adjacence.

La partie de droite du tableau 6 donne les ratios des erreurs quadratiques moyennes (EQM) empiriques des proportions estimées, exprimées en pourcentage de l'EQM de PS1. À quelques exceptions près, PS2 est l'estimateur poststratifié donnant les pires résultats pour les quatre premières caractéristiques, quelle que soit la combinaison de variable, $f_{\max}$ et méthode de regroupement des cellules. Quand $f_{\max} = 1,8,$ c'est-à-dire le choix donnant lieu à un plus grand nombre de regroupements, les EQM de PS2 excède de 1,8 % à 44,2 % celles de PS1. Les EQM de PS.WR1 ainsi que PS.WR2 sont proches de celles de PS1, à l'exception du cas (hospitalisation, $f_{\max} = 1,8,$ adjacence) où le biais de 6,3 % de PS.WR2 donne lieu à une EQM supérieure de 25,6 % à celle de PS1. Le regroupement selon la moyenne proche est préférable au regroupement par adjacence, quoique pour les quatre premières caractéristiques, aucun des estimateurs ne donnent des EQM plus faibles que celles de PS1, qui ne comporte aucun regroupement de cellules.

De nouveau, les estimateurs donnent des résultats différents en ce qui concerne la variable de moyenne commune *Y*. Les EQM de Hájek, PS2, PS.WR1 et PS.WR2 sont toutes inférieures à celles de PS1. L'estimateur de Hájek est celui dont l'EQM est la plus faible, parce que la poststratification n'est pas nécessaire pour corriger le biais dans l'estimation de la moyenne.

Tableau 4 Erreur absolue moyenne de correction de la couverture telle qu'elle est définie dans l'expression (10)

Méthode de regroupement	Seuil de correction $f_{\max}$	Hájek	PS2 (regroupement standard)	PS.WR1 (troncature des poids puis regroupement)	PS.WR2 (correction maximale fixe des poids)
C1 Ratios de couverture pour la variable d'assurance-maladie					
Adjacence	2	0,257	0,120	0,086	0,221
Moyenne proche	2	0,257	0,080	0,127	0,281
Adjacence	1,8	0,256	0,150	0,085	0,020
Moyenne proche	1,8	0,256	0,101	0,109	0,258
C5 Ratio de couverture pour la variable de moyenne commune *Y*					
Adjacence	2	0,442	0,326	0,196	0,331
Moyenne proche	2	0,441	0,321	0,203	0,370
Adjacence	1,8	0,442	0,330	0,206	0,376
Moyenne proche	1,8	0,442	0,337	0,214	0,446

Tableau 5
Biais relatifs (en pourcentage) des proportions estimées (Les chiffres pour Hájek et PS1 ne sont pas affectés par le regroupement et sont répétés dans les quatre parties du tableau pour faciliter les comparaisons)

Caractéristiques	Fourchette du nombre de post-strates après le regroupement	Hájek	PS1 (pas de regroupement)	PS2 (regroupement standard)	PS.WR1 (troncature des poids puis regroupement)	PS.WR2 (correction maximale fixe des poids)
Regroupement par adjacence, seuil de correction = 2						
Assurance-maladie	(10, 16)	-11,5	0,1	-4,4	1,0	-1,4
Limitations	(8, 15)	-12,1	-0,3	-2,0	0,1	-1,0
Retard des soins	(6, 14)	8,2	-0,2	2,2	-0,6	0,9
Hospitalisation	(9, 16)	13,4	0,2	6,2	-0,7	2,8
Moyenne commune *Y*	(5, 11)	0,3	0,0	0,4	0,4	0,6
Regroupement selon la moyenne proche, seuil de correction = 2						
Assurance-maladie	(10, 16)	-11,5	0,1	-0,5	0,5	-0,3
Limitations	(8, 15)	-12,1	-0,3	-1,2	0,2	-1,1
Retard des soins	(6, 14)	8,2	-0,2	-0,3	-0,3	-0,2
Hospitalisation	(9, 16)	13,4	0,2	0,4	0,1	0,4
Moyenne commune *Y*	(6, 11)	0,3	0,0	0,2	0,1	0,2
Regroupement par adjacence, seuil de correction = 1,8						
Assurance-maladie	(7, 13)	-11,5	0,1	-6,5	0,7	-3,5
Limitations	(7, 12)	-12,1	-0,3	-3,4	0,3	-2,0
Retard des soins	(5, 11)	8,2	-0,2	3,5	-0,4	2,5
Hospitalisation	(5, 12)	13,4	0,2	9,4	0,0	6,3
Moyenne commune *Y*	(5, 9)	0,3	0,1	0,5	0,6	0,6
Regroupement selon la moyenne proche, seuil de correction = 1,8						
Assurance-maladie	(6, 13)	-11,5	0,1	-1,6	0,3	-1,7
Limitations	(7, 12)	-12,1	-0,3	-2,7	0,9	-2,4
Retard des soins	(5, 10)	8,2	-0,2	0,2	-0,3	0,5
Hospitalisation	(5, 12)	13,4	0,2	1,5	0,3	2,0
Moyenne commune *Y*	(5, 10)	0,3	0,2	0,3	0,3	0,3

Tableau 6 Ratios des variances (ou des EQM) à la variance (ou à l'EQM) de l'estimateur poststratifié sans regroupement de cellules (PS1) Les chiffres pour Hájek sont répétés dans les quatre parties du tableau pour faciliter les comparaisons)

Caractéristiques	Ratios des variances à la variance de l'estimateur poststratifié (PS1)				Ratios des EQM à l'EQM de l'estimateur poststratifié (PS1)			
	Hájek	PS2 (regroupement standard)	PS.WR1 (troncature des poids puis regroupement)	PS.WR2 (correction maximale fixe des poids)	Hájek	PS2 (regroupement standard)	PS.WR1 (troncature des poids puis regroupement)	PS.WR2 (correction maximale fixe des poids)
Regroupement par adjacence, seuil de correction = 2								
Assurance-maladie	0,877	1,014	1,025	0,991	1,500	1,101	1,018	1,006
Limitations	0,821	0,966	1,035	0,977	1,555	1,008	1,017	0,992
Retard des soins	1,099	1,023	1,003	1,000	1,239	1,023	1,000	1,000
Hospitalisation	1,290	1,169	1,000	1,073	1,733	1,244	1,000	1,070
Moyenne commune *Y*	0,755	0,805	0,908	0,818	0,752	0,801	0,904	0,826
Regroupement selon la moyenne proche, seuil de correction = 2								
Assurance-maladie	0,877	1,013	1,014	1,008	1,500	1,006	1,006	1,006
Limitations	0,821	0,999	1,025	0,994	1,555	1,008	1,017	1,008
Retard des soins	1,099	0,997	0,998	1,001	1,239	1,000	1,000	1,000
Hospitalisation	1,290	1,011	1,000	1,008	1,733	1,012	1,000	1,012
Moyenne commune *Y*	0,776	0,935	0,974	0,902	0,781	0,933	0,973	0,906
Regroupement par adjacence, seuil de correction = 1,8								
Assurance-maladie	0,877	0,960	1,044	0,976	1,500	1,179	1,024	1,048
Limitations	0,821	0,939	1,032	0,961	1,555	1,034	1,017	1,000
Retard des soins	1,099	1,051	0,991	1,032	1,239	1,057	0,989	1,034
Hospitalisation	1,290	1,225	1,043	1,201	1,733	1,442	1,023	1,256
Moyenne commune *Y*	0,780	0,815	0,882	0,828	0,779	0,816	0,893	0,829
Regroupement selon la moyenne proche, seuil de correction = 1,8								
Assurance-maladie	0,877	1,010	1,006	1,019	1,500	1,018	1,000	1,024
Limitations	0,821	0,983	1,051	0,975	1,555	1,034	1,034	1,017
Retard des soins	1,099	1,003	0,995	1,001	1,239	1,000	1,000	1,000
Hospitalisation	1,290	1,052	1,001	1,059	1,733	1,035	1,000	1,047
Moyenne commune *Y*	0,771	0,924	0,958	0,876	0,778	0,932	0,959	0,879

Tableau 7
Taux de couverture en pourcentage des intervalles de confiance à 95 % calculés en utilisant la loi *t* avec 25 DDL (Les chiffres pour Hájek et PS1 ne sont pas affectés par le regroupement et sont répétés dans les quatre parties du tableau pour faciliter les comparaisons)

Caractéristiques	Hájek	PS1 (pas de regroupement)	PS2 (regroupement standard)	PS.WR1 (troncature des poids puis regroupement)	PS.WR2 (correction maximale fixe des poids)
Regroupement par adjacence, seuil de correction = 2					
Assurance-maladie	75,9	93,8	90,1	94,4	93,6
Limitations	70,9	94,5	93,1	94,5	93,9
Retard des soins	92,0	94,0	94,5	94,0	94,6
Hospitalisation	82,2	94,5	91,8	94,3	93,9
Moyenne commune *Y*	94,8	93,8	94,6	94,4	94,5
Regroupement selon la moyenne proche, seuil de correction = 2					
Assurance-maladie	75,9	93,8	93,7	94,2	94,0
Limitations	70,9	94,5	93,0	94,3	93,5
Retard des soins	92,0	94,0	93,6	93,9	93,8
Hospitalisation	82,2	94,5	94,3	94,7	94,6
Moyenne commune *Y*	93,7	92,9	92,2	92,5	93,2
Regroupement par adjacence, seuil de correction = 1,8					
Assurance-maladie	75,9	93,8	87,5	94,1	92,4
Limitations	70,9	94,5	92,0	94,2	93,3
Retard des soins	92,0	94,0	94,8	94,5	94,5
Hospitalisation	82,2	94,5	88,4	94,0	91,1
Moyenne commune *Y*	94,8	94,1	94,3	94,8	94,7
Regroupement selon la moyenne proche, seuil de correction = 1,8					
Assurance-maladie	75,9	93,8	92,8	94,3	93,3
Limitations	70,9	94,5	92,3	94,5	93,0
Retard des soins	92,0	94,0	93,8	94,0	94,4
Hospitalisation	82,2	94,5	93,8	94,6	93,8
Moyenne commune *Y*	94,9	94,5	93,6	93,8	94,8

Le tableau 7 donne les couvertures empiriques des IC à 95 % calculées en utilisant les proportions estimées et l'estimateur de variance par linéarisation qui accompagne naturellement chacune d'elles. Une loi *t* avec 25 degrés de liberté est utilisée dans tous les cas. Les taux de couverture pour l'estimateur de Hájek sont fort médiocres, comme prévu, variant de 70,9 % à 92 % pour les quatre premières caractéristiques. Les estimateurs poststratifiés, PS1 et PS.WR1, donnent une couverture variant de 93,8 % à 94,7 %, c'est-à-dire proche du taux nominal de 95 %. En revanche, la couverture de PS2 est un peu moins bonne pour l'assurance-maladie et l'hospitalisation, particulièrement dans le cas (adjacence, $f_{\max} = 1,8$) où les couvertures sont de 87,5 % et 88,4 %. Les taux de couverture sont légèrement plus faibles pour PS.WR2 que pour PS.WR1, mais raisonnablement proches du taux nominal. L'utilisation du regroupement selon la moyenne proche améliore généralement les cas de couverture médiocre observés avec la méthode de regroupement par adjacence. Pour la moyenne commune *Y*, les taux de couverture sont bons, variant de 92,2 % à 94,9 %.

En résumé, les estimateurs à poids restreint, PS.WR1 et PS.WR2, offrent un certain avantage comparativement à l'estimateur par regroupement standard, PS2. Ils sont généralement moins biaisés et retiennent une plus grande part de la correction de la sous-couverture que PS2. Cependant, l'élément le plus crucial dans le contrôle du biais est, avant tout, la façon dont les cellules sont regroupées. Le regroupement d'après la proximité des moyennes ou des taux de couverture des cellules est de loin préférable au regroupement selon un critère d'adjacence ne se fondant ni sur l'un ni sur l'autre de ces paramètres. Nous n'avons observé une amélioration de l'EQM due au regroupement des cellules que si les moyennes de cellule étaient les mêmes. Toutefois, en pratique, l'égalité de ces moyennes est l'exception et non la règle.

5. Conclusion

Les concepteurs d'enquêtes auprès des ménages ou des établissements ont souvent à l'esprit une longue liste de post-strates ou de cellules lorsqu'ils élaborent les systèmes de pondération. Si la taille d'échantillon dans une post-strate est faible ou que l'estimation d'échantillon du chiffre de population dans une post-strate est fort différente d'un dénombrement externe de contrôle, la post-strate peut être regroupée avec une strate adjacente. La justification habituelle du regroupement est qu'il réduit le risque d'obtenir des poids extrêmes ainsi que les variances des estimations.

Toutefois, un mauvais choix de la méthode de regroupement a au moins deux conséquences indésirables, à savoir i) une base de sondage déficiente ou une couverture d'échantillon qui n'est pas entièrement corrigée dans certaines cellules et ii) des estimations calculées selon l'approche standard de regroupement des cellules susceptibles d'être assez biaisées. Le second problème peut donner lieu à des intervalles de confiance dont la couverture est sensiblement plus faible que le taux nominal. Le regroupement de cellules induit un biais lorsque les taux de couverture, les moyennes de cellules ou les deux sont corrélés dans une post-strate regroupée. Le biais peut être positif ou négatif, selon la corrélation.

Les cellules devraient être regroupées en se fondant sur la similarité des taux de couverture, des moyennes de population de cellule, ou des deux afin d'éviter de causer un biais. Cette méthode de regroupement peut différer considérablement des procédures standard consistant à ne regrouper que des cellules « adjacentes », par exemple à combiner des groupes d'âge contigus. Si les cellules adjacentes coïncident avec des cellules dont les taux de couverture ou les moyennes sont semblables, aucun biais n'est produit; cet aspect devrait toutefois être vérifié plutôt que présumé.

Le regroupement fondé sur les moyennes de cellules pose au moins deux problèmes pratiques. L'un est que la théorie nous indique de regrouper les cellules en nous appuyant sur les moyennes de population, mais que dans un échantillon particulier, nous ne disposerons que d'estimations de la population couverte par la base de sondage. Or, la couverture peut être si déficiente que les moyennes des parties couvertes et non couvertes de la population seront sensiblement différentes, même dans les post-strates initiales. Il s'agirait alors d'un cas de « non-couverture non ignorable ». Dans de telles conditions, la poststratification fondée uniquement sur l'ensemble initial de cellules ou de combinaison de ces cellules ne permet pas de corriger le biais de couverture. Un deuxième problème pratique est que les données sur de nombreux items sont recueillies dans la plupart des enquêtes. Le regroupement fondé sur les moyennes de cellules pour une variable pourrait ne pas donner de bons résultats pour d'autres variables. Le cas échéant, le compromis, proposé par Little et Vartivarian (2005) pour la correction de la non-réponse, consistant à regrouper les cellules en se basant sur une moyenne pondérée des moyennes d'un ensemble important de variables devrait être une bonne solution.

Les présents travaux de recherche pouvaient être étendus en vue d'examiner les propriétés de la classe d'estimateurs par calage en ce qui a trait à la correction des erreurs de couverture. La poststratification en représente un cas particulier. Lorsque l'on combine des catégories de

variables auxiliaires qualitatives parce que les tailles d'échantillon sont faibles ou pour d'autres raisons, les mêmes problèmes de biais que ceux illustrés ici peuvent être introduits dans des estimateurs par calage plus généraux. Un moyen de donner une certaine possibilité de s'écarter des valeurs de contrôle tout en retenant d'importantes variables auxiliaires est déjà proposé dans Rao et Singh (1997). L'effet de leur proposition sur le biais de couverture doit être étudié.

Remerciements

Les auteurs remercient le rédacteur associé et les examinateurs de leur examen minutieux et de leurs commentaires constructifs. Le présent article présente les résultats généraux d'une étude entreprise en partie par le personnel du National Center for Health Statistics (NCHS). Les opinions exprimées n'engagent que les auteurs et ne reflètent pas forcément celles du NCHS. Les travaux de R. Valliant sont financés partiellement en vertu des contrats de services professionnels 200-2004-M-09302 et 200-2006-M-17916 conclus entre l'Université du Michigan et le National Center for Health Statistics.

Bibliographie

Bureau of the Census (2002). *Sources and Accuracy of Estimates for Poverty in the United States*: *2002*. Washington DC.

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

Fuller, W. (1966). Estimation employing post strata. *Journal of the American Statistical Association*, 61, 1172-1183.

Gonzalez, J.F., Town, M. et Kim, J. (2005). Mean square error analysis of health estimates from the Behavioral Risk Factor Surveillance System for counties along the United States/Mexico border region. *Proceedings of the Section on Survey Methods Research*. Alexandria VA: American Statistical Association.

Kalton, G., et Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-97.

Kalton, G., et Maligalig, D.S. (1991). A comparison of methods of weighting adjustment for nonresponse. *Proceedings of the U.S. Bureau of the Census Annual Research Conference*, 409-428.

Kim, J.J. (2004). Effect of collapsing rows/columns of weighting matrix on weights. *Proceedings of the Section on Survey Methods Research*, American Statistical Association.

Kim, J.J., Thompson, J.H., Woltman, H.F. et Vajs, S.M. (1982). Empirical results from the 1980 Census Sample Estimation Study. *Proceedings of Section on Survey Research Methods*, American Statistical Association, 170-175.

Kim, J.J., et Tompkins, L. (2007). Comparisons of current and alternative collapsing approaches for improved health estimates. Article présenté à la 11ième *Biennial CDC/ASTDR Symposium on Statistical Methods*, Atlanta, Georgia, 17-18 avril 2007.

Kim, J.J., Tompkins, L., Li, J. et Valliant, R. (2005). A simulation study of cell collapsing in poststratification. *Proceedings of the Section on Survey Methods Research*, American Statistical Association.

Kostanich, D., et Dippo, C. (2000). *Current Population Survey*: *Design and Methodology*. Article technique 63. Washington DC: Department of Commerce.

Lazzeroni, L., et Little, R.J.A. (1998). Random-effects models for smoothing poststratification weights. *Journal of Official Statistics*, 14, 61-78.

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

Little, R.J.A., 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.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9, 1-19.

Lumley, T. (2005). Survey: Analysis of complex survey samples. R package version 3.0-1. University of Washington: Seattle.

R Development Core Team (2005). R: *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org.

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

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

Tompkins, L., et Kim, J.J. (2006). Evaluation of collapsing criteria in sample weighting. Internal NCHS memorandum.

Tremblay, V. (1986). Critères pratiques pour la définition des classes de pondération. *Techniques d'enquête*, 12, 91-103.

Techniques d'enquête, décembre 2007
Vol. 33, N° 2, pp. 171-178
Statistique Canada, N° 12-001-X au catalogue

Un estimateur à base de sondage unique fondé sur la multiplicité pour les sondages à bases multiples

Fulvia Mecatti [1]

Résumé

Les sondages à bases multiples ont été proposés au départ pour favoriser la réduction des coûts dans un contexte d'*optimalité*. Alors que les sondages de populations *spéciales*, *rares* et *difficiles à échantillonner* prennent de l'importance, il arrive souvent, en pratique, que l'on ne dispose pas d'une liste unique des unités de la population comme base de sondage. Récemment, des plans de sondage à bases multiples ont été proposés dans la littérature afin d'accroître la couverture de la population, d'améliorer les taux de réponse et de saisir les différences et les sous-groupes. Diverses approches de l'estimation d'après des bases de sondage multiples ont été présentées, toutes fondées sur la partition virtuelle de l'ensemble de bases de sondage chevauchantes disponibles en domaines disjoints. Par conséquent, la classification correcte des unités d'échantillonnage dans les domaines est requise pour les applications pratiques. Dans le présent article, nous proposons un estimateur pour bases de sondage multiples fondé sur une approche de *multiplicité*. Les estimateurs fondés sur la multiplicité requièrent moins d'information sur l'appartenance d'une unité à un domaine et ne sont donc pas sensibles aux erreurs de classification. En outre, l'estimateur proposé est analytiquement simple, si bien qu'il est facile à appliquer et que sa variance est donnée exactement. Nous présentons aussi les résultats empiriques d'une grande étude par simulation conçue pour comparer l'estimateur fondé sur la multiplicité aux principaux estimateurs concurrents.

Mots clés : Populations difficiles à échantillonner; sondage à base double; erreur de classification; méthode itérative du quotient; estimation de la variance.

1. Introduction

Dans l'échantillonnage en population finie classique, l'une des hypothèses fondamentales est que l'on dispose comme base d'échantillonnage d'une liste unique et complète des unités formant la population cible. Dans certains cas, un ensemble de deux ou plusieurs listes est disponible pour les besoins du sondage. Le cas général de $Q \geq 2$ listes, singulièrement partielles et éventuellement chevauchantes, est appelé *sondage à bases multiples*. Ce type de sondage a été introduit à l'origine (Hartley 1974) comme moyen de réduire les coûts tout en obtenant la même précision que par un sondage à base unique habituel. Dans la pratique contemporaine de l'échantillonnage, les sondages de *populations spéciales*, *rares* et *difficiles à échantillonner* deviennent plus fréquents (Kalton et Anderson 1986; Sudman et Kalton 1986; Sudman, Sirken et Cowan 1988), mais il arrive souvent qu'il n'existe pas de liste unique des unités et que la taille N de la population soit un paramètre inconnu qu'il faut estimer. Les auteurs d'études publiées récemment examinent les sondages à bases multiples dans le but principal d'accroître la couverture de la population, d'améliorer les taux de réponse et de saisir plus exactement les différences et les sous-groupes (Iachan et Dennis 1993; Carlson et Hall 1994; Haines et Pollock 1998; Eurostat 2000). Dans un article récent, Lohr et Rao (2006) mentionnent qu'« à mesure que les populations du Canada, des États-Unis et d'autres pays se diversifient, l'utilisation de bases de sondage différentes pourrait permettre de mieux refléter les divers sous-groupes. […] Nous nous attendons à ce que l'usage de plans de sondage modulaires faisant appel à des bases multiples se généralise dans l'avenir » [traduction]. Une application contemporaine de ce principe pourrait s'observer dans les sondages en ligne, dont on peut améliorer la couverture de la population et réduire le biais dû aux caractéristiques du site Internet servant pour la collecte des données en utilisant simultanément deux ou plusieurs sites indépendants. Puisqu'une même unité peut visiter plus d'un site participant au sondage, les sites se chevauchent pour configurer un cadre à bases multiples.

L'estimation dans le contexte des sondages à bases multiples, telle qu'elle a été élaborée au départ par Hartley (1962, 1974), est fondée sur la partition virtuelle de la population (c'est-à-dire l'union inconnue des Q bases de sondage chevauchantes) en $2^Q - 1$ *domaines* disjoints (c'est-à-dire les intersections mutuellement exclusives de bases de sondage). Donc, le total Y d'une variable étudiée y, considéré comme étant le paramètre à estimer, est exprimé sous forme d'une somme de *totaux de domaine*. Les données d'échantillon provenant des Q bases de sondage sont utilisées pour produire des estimations des totaux de domaine. Enfin, les totaux de domaine estimés sont combinés pour produire l'estimation du total de population Y. Un certain nombre d'estimateurs ont été élaborés selon diverses approches de l'estimation sur bases

1. Fulvia Mecatti, Département de statistique, Université de Milan-Bicocca, Via Bicocca degli Arcimboldi, 8. Ed. U7, 20126 Milan, Italie. Courriel : fulvia.mecatti@unimib.it.

de sondage multiples (voir la section 2). Puisque tous les estimateurs décrits dans la littérature s'appuient sur la partition en domaines susmentionnée, leur application pratique requiert la détermination correcte de l'appartenance de chaque unité échantillonnée aux domaines. Il s'agit là d'une hypothèse forte qui n'est pas toujours vérifiée en pratique, comme le soutiennent par exemple Lohr et Rao (2006). Cette hypothèse sous-entend en effet que chaque unité échantillonnée doit être interrogée en vue de recueillir les réponses au sondage ainsi que les renseignements sur son appartenance à chaque base de sondage utilisée, afin de pouvoir classer correctement les unités dans les domaines. En plus du risque naturel d'erreur de classification, il pourrait exister un risque lié à la *confidentialité* et à la nature *délicate* de l'appartenance des unités à la base de sondage, susceptible à la fois d'accroître le taux de non-réponse et d'affecter la précision de l'estimateur. Il pourrait en être ainsi, par exemple, lorsqu'on étudie des caractéristiques de *nature délicate* (comportements intimes, toxicomanie...) ou que l'on échantillonne des populations *difficiles à joindre* (immigrants illégaux, anciens détenus, patients...). Dans le présent article, nous adoptons une approche différente de l'estimation dans les sondages à bases multiples. Nous proposons le concept de *multiplicité* de l'unité, c'est-à-dire le *nombre* de bases de sondage auxquelles appartient cette unité, pour remplacer les approches existantes fondées sur l'appartenance aux domaines, c'est-à-dire *à quelles* bases de sondage appartient l'unité. Nous présentons un estimateur sans biais, naturellement insensible aux erreurs de classification dans les domaines et s'appliquant à n'importe quel nombre de bases de sondage. L'estimateur fondé sur la multiplicité proposé possède une structure analytique simple, de sorte qu'il est facile à appliquer, tandis que sa variance exacte est donnée sous forme explicite et peut donc être estimée sans difficulté pour toute taille d'échantillon.

À la section 2, nous présentons une discussion générale des principales contributions à l'estimation d'après des bases de sondage multiples dans une perspective unifiée, ainsi que la notation nécessaire. À la section 3, nous proposons un estimateur fondé sur la multiplicité et analysons l'estimation de la variance. À la section 4, nous présentons les résultats d'une grande étude par simulation conçue pour comparer l'estimateur proposé aux principaux estimateurs concurrents.

2. Estimation optimale, pseudo-optimale et d'après une base de sondage unique

Bien que la littérature traite principalement du cas de la base de sondage double ($Q = 2$), un cadre théorique général pour les sondages à bases multiples ($Q \geq 2$) a été proposé récemment par Lohr et Rao (2006). En utilisant leur notation en bases multiples, nous passons brièvement en revue diverses approches d'estimation et présentons les estimateurs disponibles d'une manière unifiée qui met en relief le fait qu'ils dépendent de la caractéristique d'appartenance aux domaines des unités échantillonnées.

Soit $A_1 \cdots A_q \cdots A_Q$ un ensemble de $Q \geq 2$ bases de sondage chevauchantes, dont l'union offre une couverture de population appropriée pour les objectifs du sondage. Posons que les ensembles d'indices K sont les sous-ensembles de l'intervalle de l'indice de base de sondage $q = 1 \cdots Q$. Pour chaque ensemble d'indices $K \subseteq \{1 \cdots q \cdots Q\}$, un domaine est défini comme étant l'ensemble $D_K = (\bigcap_{q \in K} A_q) \bigcap (\bigcap_{q \notin K} A_q^C)$, où ^{C} représente la complémentation.

Soit $\delta_i(K)$ *l'indicateur d'appartenance au domaine*, qui prend la valeur 1 si l'unité i est incluse dans le domaine D_K et 0 autrement. Le total estimé Y sur l'union (inconnue) des Q bases de sondage chevauchantes est alors exprimé comme la somme sur l'ensemble de $2^Q - 1$ domaines disjoints

$$Y = \sum_{i \in \bigcup_q A_q} y_i = \sum_K \sum_{i \in \bigcup_q A_q} \delta_i(K)\, y_i. \qquad (1).$$

Soit s_q un échantillon sélectionné à partir de la base de sondage A_q sous un plan donné, indépendamment pour $q = 1 \cdots Q$. Une expression générale d'un estimateur à bases de sondage multiples fondée sur la classification par domaine est alors

$$\hat{Y} = \sum_K \sum_{q \in K} \sum_{i \in s_q} w_i^{(q)}\, \delta_i(K)\, y_i. \qquad (2)$$

Soulignons que, lorsqu'un estimateur sans biais du total Y est donné, un estimateur de la taille de population N est également donné en remplaçant simplement les valeurs d'échantillon y_i par des 1.

Les estimateurs proposés dans la littérature sont obtenus en fixant les poids $w_i^{(q)}$ dans l'expression (2) conformément à trois grandes approches. Puisque les sondages à bases multiples ont été proposés au départ dans le but de réaliser des économies tout en obtenant une précision égale ou supérieure à celle des sondages à base unique habituels, il a d'abord été recommandé de suivre une approche *optimale* en utilisant les poids optimaux $w_{i,\,\mathrm{opt}}^{(q)}$ dans (2), c'est-à-dire en minimisant la variance de l'estimateur (Hartley 1962, 1974; Lund 1968; Fuller et Burmeister 1972). Les estimateurs optimaux possèdent des propriétés théoriques optimales (Skinner 1991; Lohr et Rao 2000), mais posent des problèmes pratiques dus principalement à leur complexité (des formules explicites, mais compliquées sont données pour les poids optimaux $w_{i,\,\mathrm{opt}}^{(q)}$ pour n'importe quel nombre de bases de sondage dans Lohr et Rao 2006, section 3). En outre, les poids optimaux dépendent des covariances de

population inconnues, si bien qu'ils doivent être estimés d'après les données d'échantillon, ce qui complique les calculs et affecte l'optimalité, puisque la variabilité supplémentaire de l'estimation des covariances produit des erreurs quadratiques moyennes plus grandes (Lohr et Rao 2006, section 7).

Afin d'améliorer l'applicabilité, on a proposé d'adopter une approche de *base de sondage unique* (BU) grâce à l'utilisation de *poids fixes* de sorte qu'il n'y ait pas de biais par rapport au plan. Dans le cas de l'échantillonnage aléatoire simple dans chaque base de sondage, l'estimateur BU sur base unique s'obtient en remplaçant dans (2) les poids $w_i^{(q)}$ par les poids $w_{i,\,\mathrm{SF}}^{(q)} = w^{(K)} = (\sum_{q\in K} f_q)^{-1}$, où $f_q = n_q / N_q$ désigne la fraction d'échantillonnage dans la base de sondage (Bankier 1986; Kalton et Anderson 1986; Skinner 1991; Skinner, Holmes et Holt 1994). Puisque les poids fixes diffèrent habituellement des poids optimaux, l'estimateur BU est généralement moins efficace que l'estimateur optimal (Lohr et Rao 2000). Enfin, une *approche pseudo-optimale* a été proposée (Skinner et Rao 1996; Lohr et Rao 2000) afin d'obtenir un estimateur d'une applicabilité plus générale que les estimateurs optimaux et d'une plus grande efficacité que les estimateurs BU. Un estimateur du pseudo maximum de vraisemblance (PMV) pour les sondages à bases multiples s'obtient en faisant dans (2) la substitution $w_{i,\,\mathrm{PML}}^{(q)} = w^{(K)} = \hat{N}_K / \sum_{q\in K}\sum_{i\in s_q}\delta_i(K) = \hat{N}_K / n_K$, où les tailles estimées de domaine $\hat{N}_K$ sont les solutions d'un système d'équations non linéaires. Bien qu'il soit difficile à mettre en œuvre pour des applications pratiques (une approximation linéaire itérative de $\hat{N}_K$ sous échantillonnage aléatoire simple est donnée dans Lohr et Rao 2006, section 4.1), l'estimateur PMV garde les bonnes propriétés théoriques de l'approche optimale.

Notons que la formule (2) contient l'indicateur d'appartenance au domaine $\delta_i(K)$; donc les estimateurs optimal, pseudo-optimal et BU s'appliquent uniquement si la classification des données d'échantillon dans les $2^Q - 1$ domaines est effectuée correctement.

À la section suivante, nous présentons un estimateur pour bases de sondage multiples qui s'appuie sur une approche à base de sondage unique fondée sur la *multiplicité* ne nécessitant pas de classification par domaine.

3. L'estimateur à base de sondage unique fondé sur la multiplicité

La notion de multiplicité a été introduite pour la première fois dans le contexte de l'échantillonnage par réseau (Casady et Sirken 1980, Sirken 2004). Elle fait aussi partie des outils de la méthode généralisée de partage des poids (Lavallée 2002, 2007), ainsi que de la théorie de l'estimation par échantillonnage de centres (Mecatti 2004), puisque l'échantillonnage de centres et le sondage à bases multiples sont équivalents dans certaines conditions. Dans Lohr et Rao, (2006), la multiplicité du domaine D_K est définie comme étant la cardinalité de l'ensemble d'indices K. Puisque les domaines sont mutuellement exclusifs, la multiplicité est aussi une caractéristique de chaque unité de la population, c'est-à-dire le *nombre* de bases de sondage dans lesquelles chaque unité est incluse parmi les Q bases utilisées pour le sondage.

Soit m_i la multiplicité de l'unité i. Notons que la multiplicité de l'unité peut être déterminée simplement en demandant aux unités échantillonnées *à combien de bases de sondage elles appartiennent*.

Puisque, de toute évidence $\sum_q \sum_{i\in A_q} y_i = \sum_{i\in \cup_q A_q} m_i\, y_i$, il s'ensuit que

$$Y = \sum_{q=1}^{Q}\sum_{i\in A_q} y_i\, m_i^{-1}. \qquad (3)$$

Notons que l'expression (3), qui ne comporte que des sommations sur les bases de sondage, offre un avantage pratique par rapport à l'équation (1). En fait, les domaines fournissent une partition virtuelle (inconnue) de la population, tandis que la sélection de l'échantillon a effectivement lieu dans les Q bases de sondage chevauchantes. Cela nous mène à un estimateur BU fondé sur la multiplicité donné par

$$\hat{Y}_M = \sum_{q=1}^{Q}\sum_{i\in s_q} w_i^{(q)}\, y_i\, m_i^{-1} \qquad (4)$$

où les poids fixes $w_i^{(q)}$ permettent de s'assurer, par exemple, de l'absence de biais par rapport au plan. Dans le cas de l'échantillonnage aléatoire simple de chaque base de sondage, nous avons $w_i^{(q)} = f_q^{-1}$, $\forall i \in s_q$.

Contrairement aux estimateurs optimal, PMV et BU décrits à la section 2, l'estimateur (4) ne contient pas l'indicateur d'appartenance à l'échantillon et il est très facile à mettre en œuvre dans les applications pratiques. De surcroît, il convient de souligner que, pour l'échantillonnage aléatoire simple de chaque base de sondage, dans l'estimateur fondé sur la multiplicité (4), les valeurs échantillonnées sont pondérées par $(f_q\, m_i)^{-1}$, c'est-à-dire un coefficient *particulier* à la base de sondage; inversement, dans l'estimateur BU, les valeurs échantillonnées sont pondérées par $w_i^{(K)} = (\sum_{q\in K} f_q)^{-1}$, c'est-à-dire un coefficient *moyen* calculé sur l'ensemble des bases de sondage intervenant dans chaque domaine. Par conséquent, l'estimateur $\hat{Y}_M$ devrait être plus précis que l'estimateur BU, comme le confirment les résultats des simulations. En outre, étant donné sa structure d'Horvitz-Thompson, il est possible de calculer sa variance exacte sous une forme explicite. Dans le cas de l'échantillonnage aléatoire simple de chaque base de sondage, la variance de l'estimateur est donnée par

$$V(\hat{Y}_M) = \sum_{q=1}^{Q} \frac{N_q - n_q}{n_q(N_q - 1)} \left[N_q \sum_{i \in A_q} y_i^2 m_i^{-2} - \left(\sum_{i \in A_q} y_i m_i^{-1} \right)^2 \right]. \quad (5)$$

Un estimateur sans biais de la variance sous échantillonnage aléatoire simple de chaque base de sondage est alors donné par

$$\hat{v}(\hat{Y}_M) = \sum_{q=1}^{Q} \frac{N_q (N_q - n_q)}{n_q^2 (N_q - 1)} \left[N_q \sum_{i \in s_q} y_i^2 m_i^{-2} - f_q^{-1} \left(\sum_{i \in s_q} y_i m_i^{-1} \right)^2 \right]. \quad (6)$$

Les propriétés de l'estimateur fondé sur la multiplicité pour des tailles finies d'échantillon ont été étudiées empiriquement pour l'échantillonnage aléatoire simple et comparées à celles des principaux estimateurs concurrents par une méthode de simulation.

4. Étude par simulation

Plusieurs résultats de simulation concernant les estimateurs pour base de sondage double ont été publiés (Bankier 1986; Skinner et Rao 1996; Lohr et Rao 2000). Dans le cas général de $Q \geq 2$ bases de sondage, Lohr et Rao (2006) ont étudié en détail les erreurs quadratiques moyennes empiriques d'un groupe de huit estimateurs sous les approches optimale, pseudo-optimale et pour base de sondage unique, dans un cadre à trois bases de sondage sous un plan de sondage à deux degrés. Leurs résultats donnent à penser que les estimateurs optimaux le sont en théorie, mais qu'en pratique, la variabilité supplémentaire due à l'estimation des poids optimaux produit des erreurs quadratiques moyennes plus importantes. Donc, l'estimateur PMV semble être celui qui donne les meilleurs résultats en ce qui concerne l'efficacité relative empirique. En outre, l'étude comportait l'examen d'un cas où environ 10 % des unités échantillonnées avaient été classées incorrectement dans les domaines, ce qui a poussé les auteurs à recommander d'étudier plus en profondeur les effets des erreurs de classification sur les propriétés de l'estimateur.

Dans notre étude, nous comparons les estimateurs pseudo-optimal et pour base unique à l'estimateur fondé sur la multiplicité (4) en poursuivant trois grands objectifs :

i) déterminer les conditions empiriques dans lesquelles l'estimateur fondé sur la multiplicité est plus efficace que l'estimateur BU (section 4.2);
ii) considérer l'ajustement par la méthode itérative du quotient (MIQ) à des tailles connues de base de sondage N_q, tel qu'il a déjà été proposé pour améliorer l'efficacité de l'estimateur BU (section 4.3);
iii) étudier les effets de l'accroissement des taux d'erreur de classification sur les propriétés empiriques des estimateurs PMV et BU (simple et ajusté par la MIQ) comparativement à l'insensibilité naturelle de l'estimateur fondé sur la multiplicité (section 4.4).

4.1 Mise en œuvre

L'étude par simulation a été exécutée dans des conditions artificielles de sondage à trois bases et mise en œuvre comme il suit. N pseudo-valeurs de population y_i sont générées à partir d'une loi gamma. Certains simulations préliminaires ont indiqué que ni l'accroissement de la taille de la population N ni l'utilisation de valeurs différentes pour les paramètres gamma (produisant une forme asymétrique et une forme presque symétrique) ne font varier sensiblement le profil de performance relative des estimateurs considérés. L'étude a été réalisée en fixant $N = 1\,200$ et en générant les valeurs à partir d'une loi gamma dont les paramètres étaient 1,5 et 2. Chaque pseudo-valeur y_i est affectée aléatoirement aux $Q = 3$ bases de sondage conformément à 3 essais de Bernoulli indépendants avec probabilité $\alpha_q = N_q / N$, $q = 1, 2, 3$. Plusieurs scénarios concernant la couverture ainsi que le chevauchement des bases de sondage ont été obtenus en choisissant diverses valeurs de α_q, sous deux contraintes : *a*) $\sum_q \alpha_q \geq 1$ pour s'assurer que les trois bases couvrent l'entièreté de la population et *b*) aucune des trois bases n'est vide. Dans certains cas, le chevauchement souhaité des bases de sondage a été produit en fixant le ratio N_K / N des unités de population incluses dans chaque domaine.

Étant donné un ensemble de fractions d'échantillonnage $f_q = n_q / N_q$, $q = 1, 2, 3$, un échantillon aléatoire simple est sélectionné indépendamment dans chaque base de sondage, itérativement pour 10 000 exécutions de la simulation. Pour un estimateur donné, disons $\hat{Y}$, il est supposé que l'ensemble de valeurs $\{\hat{Y}_p,\ p = 1...10\ 000\}$ correspond à sa distribution de Monte Carlo et l'on calcule la moyenne empirique $E_{mc}(\hat{Y}) = \sum_p \hat{Y}_p / 10\,000$ et l'erreur quadratique moyenne empirique $\mathrm{EQM}_{mc}(\hat{Y}) = \sum_p [\hat{Y}_p - Y]^2 / 10\,000$. L'erreur de Monte Carlo est contrôlée en acceptant uniquement les simulations produisant un biais relatif empirique $\mathrm{BR}_{mc}(\hat{Y}) = 100 \cdot |E_{mc}(\hat{Y}) - Y| / Y$ inférieur à 1,5 % pour les estimateurs que l'on sait être sans biais. En outre, en utilisant la variance exacte de l'estimateur fondé sur la multiplicité donnée par (6), on s'assure que les simulations donneront $|\mathrm{EQM}_{mc}(\hat{Y}_M) - V(\hat{Y}_M)| \leq 0,03$. Plusieurs scénarios ont été étudiés en combinant divers niveaux de couverture des bases de sondage, de chevauchement des bases de sondage et de disproportion de l'échantillonnage, ce qui aboutit à la production de 29 populations simulées. À la figure 1, ces populations sont représentées par des points

dans le plan formé par les deux principaux paramètres de simulation, à savoir la couverture (totale) des bases de sondage sur l'axe horizontal (donnée par $\sum_q \alpha_q$) et la disproportion de l'échantillonnage sur l'axe vertical, c'est-à-dire la dispersion des fractions d'échantillonnage f_q, mesurée par $\sum_q \sum_{q'} |f_q - f_{q'}| / 3^2$. Dans la figure 1, les diverses formes des symboles représentant les populations/points indiquent divers niveaux de chevauchement, c'est-à-dire le taux total d'unités de population classées dans les quatre domaines chevauchants.

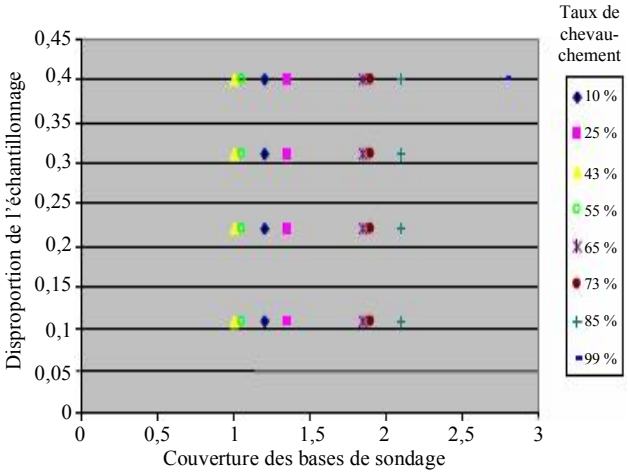


Figure 1 Populations simulées

4.2 Estimation en base de sondage unique fondée sur la multiplicité comparativement à l'approche simple

Comme il est mentionné à la section 3, l'estimateur fondé sur la multiplicité fait intervenir des poids particuliers à la base de sondage, tandis que l'estimateur BU simple est fondé sur des coefficients moyens. Par conséquent, les deux estimateurs coïncident si la fraction d'échantillonnage $f_q = f$ est la même dans chaque base de sondage, c'est-à-dire pour un échantillonnage proportionnel, et ils produisent des estimations différentes en cas d'échantillonnage disproportionnel. Les résultats de simulation fournissent des preuves empiriques que l'estimateur fondé sur la multiplicité est plus exact que l'estimateur BU simple. Ils montrent que l'estimateur $\hat{Y}_M$ est plus efficace dans tous les cas examinés, sauf un cas extrême où les trois bases de sondage sont presque complètes et que le chevauchement total est proche de 100 %. Si l'on omet ce cas particulier, le gain d'efficacité de $\hat{Y}_M$ par rapport à l'estimateur BU, mesuré par un ratio d'efficacité empirique ordinaire (voir le tableau 1), varie de 5 % à 48 %, et n'est jamais inférieur à 26 % dans la moitié des simulations. L'efficacité de l'estimateur fondé sur la multiplicité par rapport à l'estimateur BU simple augmente à mesure que s'accroît la disproportion de l'échantillonnage (voir le tableau 2), tandis qu'elle s'avère essentiellement indépendante de l'accroissement du niveau de couverture et de chevauchement des bases de sondage.

Tableau 1
Ratio d'efficacité empirique de $\hat{Y}_M$ *vs.* l'estimateur BU : statistiques élémentaires sur 28 populations simulées

Moyenne	Max.	Min.	Médiane	75e quantile
0,7425	0,95	0,52	0,74	0,89

Tableau 2
Ratio d'efficacité empirique de $\hat{Y}_M$ *vs.* l'estimateur BU pour une disproportion d'échantillonnage croissante

Disproportion d'échantillonnage	0,11 0,22 0,31 0,40
Ratio d'efficacité empirique moyen pour divers niveaux de couverture/chevauchement des bases de sondage	0,92 0,81 0,68 0,57

4.3 Ajustement par la méthode itérative du quotient (MIQ)

D'aucuns ont proposé de procéder à un ajustement par la MIQ en utilisant les tailles connues des bases de sondage N_q (Bankier 1986) afin d'accroître l'efficacité de l'estimateur BU simple. Les résultats théoriques et empiriques déjà publiés confirment que l'estimateur BU ajusté par la MIQ (BUmiq) peut être considérablement plus efficace que l'estimateur BU simple (Skinner 1991; Lohr et Rao 2000, 2006; Mecatti 2005).

Afin de corriger l'estimateur fondé sur la multiplicité par la MIQ, nous devons émettre l'hypothèse que nous connaissons l'appartenance des unités échantillonnées aux domaines. En utilisant cette information supplémentaire, quoique redondante, nous pouvons réécrire $\hat{Y}_M$ sous la forme

$$\hat{Y}_M = \sum_K \sum_{q \in K} (|K| f_q)^{-1} \sum_{i \in s_q} \delta_i(K) y_i \qquad (7)$$

où $|K|$ indique le nombre de bases de sondage intervenant dans le domaine D_K et est égal à la multiplicité m_i de l'unité pour tout $i \in D_K$. En fixant les poids initiaux à $h_{Kq}^{(0)} = (|K| f_q)^{-1}$, nous obtenons la t^{e} itération de l'estimateur fondé sur la multiplicité corrigé par la MIQ (Mmiq) en introduisant les poids corrigés suivants par substitution dans (7)

$$h_{Kq}^{(t)} = \begin{cases} \dfrac{N_q h_{Kq}^{(t-1)}}{\sum_{q \in K} h_{Kq}^{(t-1)} n_q} & \text{si } q \in K \\ h_{Kq}^{(t-1)} & \text{si } q \notin K \end{cases}$$

où $q = Q$ si t est un multiple de Q et $q = t \bmod (Q)$ autrement, pour $t = 1, 2, \cdots$ jusqu'à la convergence.

Nous avons étudié différents niveaux de couverture des bases de sondage, combinés à différents ensembles de fractions d'échantillonnage donnant lieu à une disproportion croissante d'échantillonnage, dans les simulations.

Les résultats empiriques montrent que Mmiq est plus efficace que BUmiq dans 38 % des cas étudiés et qu'il est aussi ou moins efficace dans les autres cas. Les gains d'efficacité varient de 3 % à 74 % et ont lieu pour les faibles niveaux de couverture des bases de sondage. Lorsque la couverture de la base de sondage augmente (et donc également le chevauchement), l'estimateur Mmiq est supérieur à l'estimateur BUmiq en cas de forte disproportion d'échantillonnage seulement. Dans les autres conditions, c'est-à-dire un accroissement de la couverture et du chevauchement des bases de sondage, combiné à une disproportion faible à moyenne d'échantillonnage, Mmiq peut être sensiblement moins efficace que BUmiq (voir le tableau 3 pour dix cas indicatifs) et fortement biaisé. Les résultats empiriques laissent donc entendre que l'ajustement par la MIQ a de meilleurs effets sous une approche de base de sondage unique simple que sous une approche de multiplicité, bien qu'il existe des conditions dans lesquelles cette dernière demeure supérieure. Il faudra poursuivre les travaux de recherche à cet égard. En particulier, puisque la MIQ est en fait un cas particulier de calage (Deville et Särndal 1992; Deville, Särndal et Sautory 1993), les résultats pourraient être améliorés en appliquant la méthode plus générale de calage à l'estimateur $\hat{Y}_M$. Le calage de l'estimateur fondé sur la multiplicité, considéré comme un cas particulier de la méthode généralisée de partage des poids, est décrit dans Lavallée (2002, 2007).

Tableau 3
Efficacité de Mmiq *vs.* BUmiq : dix exécutions de simulation indicatrices

Couverture de la base de sondage $\alpha_q = N_q/N$			Fraction d'échantillonnage $f_q = n_q/N_q$			Ratio d'efficacité empirique Mmiq vs. BUmiq
0,60	0,60	0,60	0,01	0,95	0,15	0,26
0,35	0,35	0,35	0,80	0,20	0,50	0,54
0,85	0,85	0,85	0,01	0,95	0,15	0,71
0,35	0,40	0,50	0,70	0,80	0,60	0,96
0,85	0,85	0,85	0,80	0,20	0,50	1,01
0,60	0,60	0,60	0,70	0,80	0,60	1,09
0,80	0,50	0,35	0,01	0,95	0,15	1,22
0,35	0,40	0,50	0,01	0,95	0,15	1,63
0,70	0,05	0,95	0,70	0,80	0,60	2,09
0,70	0,05	0,95	0,80	0,20	0,50	5,79

4.4 Erreur de classification

Le but de la dernière partie de l'étude par simulation est d'étudier la sensibilité des estimateurs pseudo-optimal (PMV) et à base de sondage unique (simple et ajusté par la MIQ) à des niveaux croissants d'erreur de classification des unités échantillonnées dans les domaines, comparativement à l'insensibilité structurelle de l'estimateur fondé sur la multiplicité proposé. Pour un taux donné d'erreur de classification, le nombre souhaité d'unités échantillonnées qui doivent être classées incorrectement est tiré du domaine dont la taille est la plus grande et assigné aléatoirement aux autres domaines, indépendamment pour chaque base de sondage.

Les tableaux 4 et 5 donnent les statistiques élémentaires résumant les résultats des simulations dans le cas d'une classification exacte et dans celui d'une légère erreur de classification égale à 1 % des unités échantillonnées. Notons que, pour la classification exacte, tous les estimateurs paraissent être sans biais (ou presque sans biais). En ce qui concerne l'efficacité, selon d'autres résultats de simulation (Lohr et Rao 2006), les estimateurs BUmiq et PMV ont des propriétés comparables. Comme prévu, pour la classification exacte, ils sont plus efficaces que $\hat{Y}_M$ dans tous les cas étudiés (sauf deux cas isolés), à cause de la quantité différente d'information utilisée dans le processus d'estimation. Cependant, les estimateurs BU (simple et ajusté par la MIQ) et PMV ont tendance à devenir biaisés et moins efficaces que $\hat{Y}_M$ en présence d'un faible taux d'erreur de classification.

Tableau 4
Biais relatif pour une erreur de classification de 1 % : statistiques élémentaires sur 29 populations simulées

BR_{mc} (absolu) pour 1 % d'unités échantillonnées mal classées	Moyenne	Min.	Max.	Médiane	75e quantile
$\hat{Y}_M$	0	0	0	0	0
BU	2,5880	0,83	7,02	2,65	2,13
BUmiq	1,7632	0,73	4	1,97	1,65
PMV	2,7352	0,23	4,67	3,46	2,87

Tableau 5
Ratio d'efficacité empirique de $\hat{Y}_M$ *vs.* les estimateurs BU et PMV : statistiques élémentaires sur les 29 populations simulées pour la classification exacte et pour une légère erreur de classification

Ratio d'efficacité empirique	Moyenne	Min.	Max.	Médiane	75e quantile
Classification exacte					
BUmiq	1,43	0,69	3,21	1,51	1,28
PMV	1,41	0,72	3,30	1,47	1,25
Erreur de classification de 1 %					
BU	0,39	0,13	0,71	0,54	0,34
BUmiq	0,78	0,13	1,98	0,95	0,74
PMV	0,77	0,14	1,94	0,98	0,70

Enfin, nous nous concentrons sur le cas de l'efficacité maximale des estimateurs BU, BUmiq et PMV

comparativement à $\hat{Y}_M$ pour la classification exacte, c'est-à-dire le cas d'un chevauchement/d'une couverture élevé des bases de sondage et d'une faible disproportion d'échantillonnage. Dans ces conditions, nous avons étudié des taux croissants d'erreur de classification des unités échantillonnées dans les domaines (variant de 0 à 50 %). Les tableaux 6 et 7 montrent, respectivement, le biais relatif et le ratio d'efficacité de $\hat{Y}_M$ par rapport aux estimateurs BU, BUmiq et PMV, pour des niveaux croissants d'erreur de classification. Il convient de souligner que, si les effets négatifs de l'erreur de classification sont rapides et importants pour tous les estimateurs étudiés, l'estimateur PMV est celui qui est le moins affecté.

En conclusion, outre le fait qu'il est simple, l'estimateur fondé sur la multiplicité proposé est recommandé lorsque le risque (même léger) d'erreur de classification des unités échantillonnées dans les domaines est une réelle possibilité.

Tableau 6
Biais relatif (absolu) pour un taux croissant d'erreur de classification

Erreur de classification en %	$\hat{Y}_M$	BU	BUmiq	PMV
0	0	0	0	≈ 0
1 %	0	2.57	1.38	4.3
5 %	0	13.57	7.15	2.75
10 %	0	17.80	14.14	4.56
20 %	0	25	25	6
50 %	0	144	68	39

Tableau 7
Ratio d'efficacité empirique de $\hat{Y}_M$ *vs.* les estimateurs BU, BUmiq et PMV pour un taux croissant d'erreur de classification

Erreur de classification en %	$\hat{Y}_M$ *vs.* BU	$\hat{Y}_M$ *vs.* BUmiq	$\hat{Y}_M$ *vs.* PMV
0	0,640	3,210	3,300
1 %	0,260	1,040	1,100
5 %	0,020	0,060	0,370
10 %	0,010	0,020	0,150
20 %	0,004	0,004	0,080
50 %	≈ 0	0,001	0,006

Remerciements

Les présents travaux ont été financés en partie par une subvention du ministère italien des Universités et de la Recherche. L'auteure tient à remercier Jon N.K. Rao pour sa discussion fructueuse et ses conseils utiles. Elle remercie également le rédacteur en chef, deux rédacteurs adjoints et deux examinateurs anonymes de leurs commentaires et suggestions constructifs.

Bibliographie

Bankier, M.D. (1986). Estimators based in several stratified samples with applications to multiple frame surveys. *Journal of American Statistical Association*, 81, 1074-1079.

Casady, R.J., et Sirken, M.G. (1980). A multiplicity estimator for multiple frame sampling. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 601-605.

Carlson, B.L., et Hall, J.W. (1994). Weighting sample data when multiple sample frames are used. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 882-887.

Deville, J.C. et Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of American Statistical Association*, 87, 376-382.

Deville, J.-C., Särndal, C.-E. et Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of American Statistical Association*, 88, 1013-1020.

Eurostat (2000). *Push and Pull Factors of International Migration.* Country Report-Italy, 3/2000/E/n.5, Bruxelles: European Communities Printing Office.

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

Haines, D.E., et Pollock, K.H. (1998). Combinaison de bases multiples pour estimer la taille et les chiffres de la population. *Techniques d'enquête*, 24, 81-91.

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

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

Iachan, R., et Dennis, M.L. (1993). A multiple frame approach to sampling the homeless and transient population. *Journal of Official Statistics*, 9, 747-764.

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

Lavallée, P. (2002). *Le sondage indirect ou la méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles, Editions Ellipses.

Lavallée, P. (2007). *Indirect Sampling*. New York : Springer.

Lohr, S.L., et Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of American Statistical Association*, 95, 271-280.

Lohr, S., et Rao, J.N.K. (2006). Multiple frame surveys: Point estimation and inference. *Journal of American Statistical Association*, 101, 1019-1030.

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

Mecatti, F. (2004). Échantillonnage de centres : stratégie d'enquête auprès des populations difficiles à échantillonner. *Recueil du Symposium de Statistique Canada.*

Mecatti, F. (2005). Estimation d'après une base de sondage unique dans une enquête à base de sondage multiple. *Recueil du Symposium de Statistique Canada.*

Sirken, M.G. (2004). Enquêtes auprès de populations rares ou difficiles à rejoindre au moyen d'échantillonnage par réseau : revue historique. *Recueil du Symposium de Statistique Canada, Discours principal.*

Skinner, C.J. (1991). On the efficiency of raking ratio estimator for multiple frame surveys. *Journal of American Statistical Association*, 86, 779-784.

Skinner, C.J., Holmes, D.J. et Holt, D. (1994). Multiple frame sampling for multivariate stratification. *Revue Internationale de Statistique*, 62, 333-347.

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

Sudman, S., et Kalton, G. (1986). New developments in the sampling of special populations. *Annual Review of Sociology*, 12, 401-429.

Sudman, S., Sirken, M.G. et Cowan, C.D. (1988). Sampling rare and elusive populations. *Science*, 2, 991-996.

Techniques d'enquête, décembre 2007
Vol. 33, N° 2, pp. 179-186
Statistique Canada, N° 12-001-X au catalogue

Estimation de la variance pour un ratio en présence de données imputées

David Haziza [1]

Résumé

Dans le présent article, nous étudions le problème de l'estimation de la variance pour un ratio de deux totaux quand l'imputation hot deck aléatoire marginale est utilisée pour remplacer les données manquantes. Nous considérons deux approches d'inférence. Dans la première, l'établissement de la validité d'un modèle d'imputation est nécessaire. Dans la seconde, la validité d'un modèle d'imputation n'est pas nécessaire, mais il faut estimer les probabilités de réponse, auquel cas il est nécessaire d'établir la validité d'un modèle de non-réponse. Nous obtenons les estimateurs de la variance sous deux cadres distincts, à savoir le cadre à deux phases habituel et le cadre inversé.

Mots clés : Modèle d'imputation; modèle de non-réponse; imputation hot deck aléatoire marginale; cadre inversé; cadre à deux phases; estimation de la variance.

1. Introduction

L'estimation de la variance en présence de données imputées pour des paramètres univariés simples, tels que des totaux ou des moyennes de population, a fait l'objet de nombreuses études ces dernières années; voir, par exemple, Särndal (1992), Deville et Särndal (1994), Rao et Shao (1992), Rao (1996), et Shao et Steel (1999). Il est fréquent, en pratique, de devoir estimer le ratio de deux totaux de population, $R = Y / X$, où $(Y, X) = \sum_{i \in U} (y_i, x_i)$, y et x sont deux variables d'intérêt pour lesquelles des données peuvent manquer et U est la population finie (de taille N) étudiée. Alors que l'estimation de la variance dans le cas d'un ratio en présence de données imputées est un problème qui se pose souvent en pratique (surtout dans le cas des enquêtes auprès des entreprises), autant que nous sachions, la question n'a pas été étudiée à fond dans la littérature. Dans le présent article, nous considérons le cas de l'imputation hot deck aléatoire marginale (HDAM) effectuée dans le même ensemble de classes d'imputation pour les deux variables y et x. Autrement dit, pour corriger la non-réponse, on procède séparément à une imputation hot deck aléatoire pour chaque variable dans le même ensemble de classes d'imputation. Cette situation se présente fréquemment en pratique. Pour simplifier, nous considérons le cas d'une seule classe d'imputation. Les extensions aux classes d'imputation multiples sont relativement simples pour la plupart des calculs présentés ici.

Nous obtenons dans le présent article des estimateurs de la variance qui tiennent compte de l'échantillonnage, de la non-réponse et de l'imputation. Deux cadres distincts d'estimation de la variance ont été étudiés dans la littérature : i) le cadre à deux phases habituel (par exemple, Särndal (1992)) et ii) le cadre inversé (par exemple, Shao et Steel (1999)). Dans le cadre à deux phases, la non-réponse est considérée comme une deuxième phase de sélection. En d'autres termes, un échantillon aléatoire est sélectionné parmi la population selon le plan d'échantillonnage donné. Puis, étant donné l'échantillon sélectionné, l'ensemble de répondants est généré selon le mécanisme de non-réponse. Dans le cadre inversé, l'ordre de l'échantillonnage et de la réponse est inversé. Autrement dit, la population est d'abord répartie aléatoirement en une population de répondants et une population de non-répondants, selon le mécanisme de non-réponse. Puis, un échantillon aléatoire est sélectionné à partir de la population (contenant les répondants et les non-répondants), selon le plan d'échantillonnage. Comme nous le verrons à la section 4, le cadre inversé facilite le calcul des estimations de la variance, mais contrairement au cadre à deux phases, nécessite l'hypothèse supplémentaire que le mécanisme de non-réponse ne dépend pas de l'échantillon sélectionné. Cette hypothèse est satisfaite dans de nombreuses situations observées en pratique. Pour chaque cadre, l'inférence peut être fondée sur un modèle d'imputation (MI) ou sur un modèle de non-réponse (MN). L'approche MI requiert la validité d'un modèle d'imputation et l'approche MN requiert la validité d'un modèle de non-réponse.

À la section 2, nous introduisons la notation, les hypothèses et l'estimateur imputé d'un ratio dans le cas de l'imputation HDAM pondérée. Les approches MI et MN sont présentées aux sections 2.1 et 2.2. À la section 2.3, nous discutons du biais de l'estimateur imputé. À la section 3, nous obtenons les estimateurs de la variance sous le cadre à deux phases et l'approche MI en utilisant la méthode proposée par Särndal (1992). Nous montrons que, sous l'imputation HDAM, l'estimateur de la variance naïf (qui traite les valeurs imputées comme des valeurs observées) surestime généralement la variance d'échantillonnage quand y et x sont positivement corrélées. À la section 4, nous

1. David Haziza, Département de mathématiques et de statistique, Université de Montréal, Montréal, (Québec), H3C 3J7, Canada. Courriel : David.Haziza@umontreal.ca.

obtenons les estimateurs de la variance sous le cadre inversé et l'approche MI ainsi que MN, en utilisant la méthode proposée par Shao et Steel (1999). Enfin, nous présentons nos conclusions à la section 5.

2. Notation et hypothèses

Notre but est d'estimer R. Nous sélectionnons un échantillon aléatoire, s, de taille n, selon un plan d'échantillonnage donné $p(s)$. Un estimateur basé sur des données complètes est donné par

$$\hat{R} = \frac{\hat{Y}_{\mathrm{HT}}}{\hat{X}_{\mathrm{HT}}}, \tag{2.1}$$

où $(\hat{Y}_{\mathrm{HT}}, \hat{X}_{\mathrm{HT}}) = \sum_{i\in s} w_i(y_i, x_i)$ désignent les estimateurs Horvitz-Thompson de Y et X, respectivement et $w_i = 1/\pi_i$ représente le poids d'échantillonnage de l'unité i, où π_i est la probabilité d'inclusion dans l'échantillon. Dans (2.1), l'estimateur $\hat{R}$ est asymptotiquement p - sans biais pour R, c'est-à-dire $E_p(\hat{R}) \approx R$, où l'indice p désigne l'espérance et la variance par rapport au plan d'échantillonnage $p(s)$. Puisque $\hat{R}$ est une fonction non linéaire des totaux estimés, sa variance exacte sous le plan, $V_p(\hat{R})$, ne peut pas être calculée facilement. Pour contourner ce problème, on recourt souvent à la linéarisation de Taylor pour approximer la variance exacte. Un estimateur asymptotiquement p - sans biais de la variance approximative de $\hat{R}$ est donné par

$$\hat{V}_{\mathrm{SAM}} = \sum_{i\in s}\sum_{j\in s}\Delta_{ij}\, e_i\, e_j, \tag{2.2}$$

où $e_i = 1/\hat{X}_{\mathrm{HT}}(y_i - \hat{R}x_i)$, $\Delta_{ij} = (\pi_{ij} - \pi_i\,\pi_j)/\pi_{ij}\,\pi_i\,\pi_j$ et π_{ij} est la probabilité de sélection conjointe des unités i et j. Notons que $\pi_{ii} = \pi_i$. Dans le cas de l'échantillonnage simple sans remise, l'estimateur de la variance (2.2) se réduit à

$$\hat{V}_{\mathrm{SAM}} = \left(1 - \frac{n}{N}\right)\frac{1}{n\bar{x}^2}\left[s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}\right], \tag{2.3}$$

où

$$s_y^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y})^2,\ s_x^2 = \frac{1}{n-1}\sum_{i\in s}(x_i - \bar{x})^2$$

et

$$s_{xy} = \frac{1}{n-1}\sum_{i\in s}(x_i - \bar{x})(y_i - \bar{y})$$

avec

$$(\bar{y},\ \bar{x}) = \frac{1}{n}\sum_{i\in s}(y_i,\ x_i).$$

Nous examinons maintenant le cas où des données pourraient manquer pour les variables x et y. Soit a_i l'indicateur de réponse de l'unité i, tel que $a_i = 1$ si l'unité i répond à la variable y et $a_i = 0$, autrement. De même, soit b_i l'indicateur de réponse de l'unité i, tel que $b_i = 1$ si l'unité i répond à la variable x et $b_i = 0$, autrement. Soit $s_r^{(y)}$ l'ensemble de répondants pour la variable y de taille r_y et $s_r^{(x)}$, l'ensemble de répondants pour la variable x de taille r_x. En outre, soit r_{xy} le nombre de répondants aux deux variables y et x. Enfin, soit y_i^* et x_i^*, les valeurs imputées pour remplacer les valeurs manquantes y_i et x_i, respectivement. Un estimateur imputé de R est donné par

$$\hat{R}_I = \frac{\sum_{i\in s} w_i\,\tilde{y}_i}{\sum_{i\in s} w_i\,\tilde{x}_i}, \tag{2.4}$$

où $\tilde{y}_i = a_i\,y_i + (1 - a_i)\,y_i^*$ et $\tilde{x}_i = b_i\,x_i + (1 - b_i)\,x_i^*$. Sous imputation HDAM pondérée, pour remplacer la valeur manquante y_i^*, on sélectionne un donneur j aléatoirement avec remise à partir de $s_r^{(y)}$ de sorte que

$$P(y_i^* = y_j) = \frac{w_j}{\sum_{l\in s} w_l\,a_l}.$$

De même, pour remplacer la valeur manquante x_i^*, on sélectionne un donneur k aléatoirement avec remise à partir de $s_r^{(x)}$, de sorte que

$$P(x_i^* = x_k) = \frac{w_k}{\sum_{l\in s} w_l\,b_l}.$$

Notons que, si y_i ainsi que x_i manquent, j n'est généralement pas égal à k sous l'imputation HDAM.

L'imputation hot deck aléatoire dans les classes est très répandue en pratique, parce que i) elle préserve la variabilité des données originales et que ii) elle produit des valeurs plausibles. Le deuxième point est particulièrement important dans le cas de variables d'intérêt catégoriques. Cependant, l'imputation hot deck aléatoire dans les classes souffre d'une composante de variance supplémentaire due à l'utilisation d'un mécanisme d'imputation aléatoire. La principale raison motivant l'utilisation de l'imputation HDAM pondérée est qu'elle mène à un estimateur asymptotiquement sans biais sous l'approche NM (voir la section 2.1), contrairement à l'imputation HDAM non pondérée.

Soit $E_I(.\mid s, s_r^{(y)}, s_r^{(x)})$, $V_I(.\mid s, s_r^{(y)}, s_r^{(x)})$ et $\mathrm{Cov}_I(.,.\mid s, s_r^{(y)}, s_r^{(x)})$, les opérateurs d'espérance, de variance et de covariance conditionnelles par rapport au mécanisme d'imputation aléatoire (ici, l'imputation HDAM pondérée). Par un développement en série de Taylor du premier ordre, on peut montrer que

$$E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) \approx \frac{\bar{y}_r}{\bar{x}_r} \equiv \hat{R}_r, \qquad (2.5)$$

où

$$\bar{y}_r = \frac{\sum_{i \in s} w_i a_i y_i}{\sum_{i \in s} w_i a_i}$$

et

$$\bar{x}_r = \frac{\sum_{i \in s} w_i b_i y_i}{\sum_{i \in s} w_i b_i}$$

désignent les moyennes pondérées des répondants pour les variables y et x, respectivement. L'approximation (2.5) sera valide si la taille d'échantillon dans les classes est suffisamment grande, ce que nous supposons être le cas. Maintenant, soient

$$s_{yr}^2 = \frac{1}{\sum_{i \in s} w_i a_i} \sum_{i \in s} w_i a_i (y_i - \bar{y}_r)^2$$

et

$$s_{xr}^2 = \frac{1}{\sum_{i \in s} w_i b_i} \sum_{i \in s} w_i b_i (x_i - \bar{x}_r)^2$$

la variabilité des valeurs de y et des valeurs de x dans les ensembles de répondants $s_r^{(y)}$ et $s_r^{(x)}$, respectivement. En notant que, sous l'imputation HDAM pondérée,

$$V_I(y_i^*) = s_{yr}^2, \; V_I(x_i^*) = s_{xr}^2$$

et

$$\mathrm{Cov}_I(y_i^*, x_i^*) = 0,$$

nous pouvons approximer $V_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)})$ par

$$V_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)})$$

$$\approx \frac{1}{\bar{x}_r^2} \left[\sum_{i \in s} w_i^2 (1 - a_i) s_{yr}^2 + \hat{R}_r^2 \sum_{i \in s} w_i^2 (1 - b_i) s_{xr}^2 \right]. \quad (2.6)$$

Les expressions (2.5) et (2.6) seront utilisées aux sections suivantes lors de la discussion du biais et de la variance de l'estimateur imputé $\hat{R}_I$. Comme nous le verrons aux sections 3 et 4, la variance conditionnelle (2.6) est une mesure de la variabilité due au mécanisme d'imputation.

Nous allons maintenant décrire deux approches d'inférence qui seront utilisées pour obtenir les estimateurs de la variance aux sections 3 et 4, à savoir l'approche du modèle de non-réponse (MN) et l'approche du modèle d'imputation (MI).

2.1 Approche du modèle de non-réponse

Dans l'approche MN, l'inférence est faite par rapport à la distribution conjointe induite par le plan d'échantillonnage et le modèle de non-réponse. Ce dernier est un ensemble d'hypothèses au sujet de la loi inconnue des indicateurs de réponse $\mathbf{R}_s = \{(a_i, b_i); i \in s\}$. Cette loi inconnue est souvent appelée mécanisme de non-réponse. Soit $p_{yi} = P(a_i = 1 \mid s, \mathbf{Z}_s)$, la probabilité de réponse de l'unité i pour la variable y, où $\mathbf{Z}_s = \{\mathbf{z}_i; i \in s\}$ et $\mathbf{z}_i$ est un vecteur de variables auxiliaires disponibles pour toutes les unités de l'échantillon utilisées pour former les classes d'imputation. De même, soit $p_{xi} = P(b_i = 1 \mid s, \mathbf{Z}_s)$ la probabilité de réponse de l'unité i pour la variable x. Nous supposons que les unités répondent indépendamment les unes des autres; c'est-à-dire que $p_{yij} = P(a_i = 1, a_j = 1 \mid s, \mathbf{Z}_s) = p_{yi} p_{yj}$ pour $i \neq j$ et $p_{xij} = P(b_i = 1, b_j = 1 \mid s, \mathbf{Z}_s) = p_{xi} p_{xj}$ pour $i \neq j$. Cependant, nous ne supposons pas que, pour une unité donnée i, la réponse pour la variable y est indépendante de celle pour la variable x. Autrement dit, si nous posons que $p_{xyi} = P(a_i = 1, b_i = 1 \mid s, \mathbf{Z}_s)$, nous avons alors $p_{xyi} \neq p_{xi} p_{yi}$, en général. Dans une classe d'imputation, nous supposons que le mécanisme de réponse est uniforme, de sorte que $p_{yi} = p_y$, $p_{xi} = p_x$ et $p_{xyi} = p_{xy}$.

Nous supposons aussi que, après conditionnement sur s et $\mathbf{Z}_s$, le mécanisme de non-réponse est indépendant de toutes les autres variables qui interviennent dans l'estimateur imputé (2.4), ainsi que des probabilités de sélection conjointe. Autrement dit, la distribution de $\mathbf{R}_s$ ne dépend pas de $\mathbf{Y}_s = \{y_i; i \in s\}$, $\mathbf{W}_s = \{w_i; i \in s\}$ et $\mathbf{\Pi}_s = \{\pi_{ij}; i \in s, j \in s\}$, après conditionnement sur s et $\mathbf{Z}_s$. Par conséquent, sauf pour les indicateurs de réponse a_i et b_i, nous supposons que toutes les variables qui interviennent dans l'estimateur imputé (2.4) ainsi que les probabilités de sélection conjointe sont traitées comme étant fixes lorsque l'on prend en considération les espérances et les variances par rapport au modèle de non-réponse. À partir d'ici, nous utilisons l'indice q pour désigner l'espérance et la variance par rapport au mécanisme de non-réponse.

2.2 Approche du modèle d'imputation

Dans l'approche MI, l'inférence est faite par rapport à la distribution conjointe induite par le modèle d'imputation, le plan d'échantillonnage et le modèle de non-réponse. Le modèle d'imputation est un ensemble d'hypothèses au sujet de la loi inconnue de $(\mathbf{Y}_U, \mathbf{X}_U) = \{(y_i, x_i); i \in U\}$. Dans une classe d'imputation, sous imputation HDAM, le modèle d'imputation, m, est donné par

$$m: \begin{cases} y_i = \mu_y + \varepsilon_i \\ \\ x_i = \mu_x + \eta_i \end{cases} \tag{2.7}$$

où ε_i est un terme d'erreur aléatoire tel que $E_m(\varepsilon_i) = 0$, $E_m(\varepsilon_i \varepsilon_j) = 0$, pour $i \neq j$, $V_m(\varepsilon_i) = \sigma_\varepsilon^2$ et η_i est un terme d'erreur aléatoire tel que $E_m(\eta_i) = 0$, $E_m(\eta_i \eta_j) = 0$, pour $i \neq j$, $V_m(\varepsilon_i) = \sigma_\eta^2$. En outre, nous supposons que $E_m(\varepsilon_i \eta_i) = \sigma_{\varepsilon\eta}$. Ici, $E_m(.)$, $V_m(.)$ et $\mathrm{Cov}_m(.)$ désignent respectivement les opérateurs d'espérance, de variance et de covariance par rapport au modèle m. La notation tient compte implicitement du fait que les espérances ou les variances par rapport au modèle m sont conditionnelles à $\mathbf{Z}_U = \{\mathbf{z}_i;\ i \in U\}$. Dans cette approche, nous supposons que la distribution des erreurs $(\boldsymbol{\varepsilon}_U, \boldsymbol{\eta}_U) = \{(\varepsilon_i, \eta_i);\ i \in U\}$ ne dépend pas de s, $s_r^{(y)}$, $s_r^{(x)}$, $\mathbf{W}_U = \{w_i;\ i \in U\}$ ni $\boldsymbol{\Pi}_U = \{\pi_{ij};\ i \in U,\ j \in U\}$, après conditionnement sur $\mathbf{Z}_U$. Par conséquent, sauf les variables d'intérêt y et x, toutes les variables qui interviennent dans l'estimateur imputé (2.4) sont traitées comme étant fixes lorsqu'on considère les espérances et les variances par rapport au modèle d'imputation.

2.3 Biais de l'estimateur imputé

Pour étudier le biais de l'estimateur imputé (2.4), nous utilisons la décomposition standard de l'erreur totale de $\hat{R}_I$:

$$\hat{R}_I - R = \left[\hat{R} - R\right] + \left[E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R}\right] + \left[\hat{R}_I - E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)})\right]. \tag{2.8}$$

Le premier terme $\hat{R} - R$ du deuxième membre de (2.8) est appelé erreur d'échantillonnage, le deuxième terme $E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R}$ est appelé erreur de non-réponse, tandis que le troisième terme $\hat{R}_I - E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)})$ est appelé erreur d'imputation.

À l'aide d'un développement en série de Taylor du premier ordre, il est facile de montrer que, sous l'approche MN, l'estimateur imputé (2.4) est asymptotiquement pqI - sans biais, c'est-à-dire que $E_{pqI}(\hat{R}_I - R) \approx 0$. De plus, sous l'approche MI et le modèle (2.7), nous pouvons montrer que l'estimateur imputé (2.4) est asymptotiquement $mpqI$ - sans biais, c'est-à-dire que $E_{mpqI}(\hat{R}_I - R) \approx 0$. Donc, l'estimateur imputé est robuste en ce sens qu'il est valide sous l'approche MN ou sous l'approche MI. Notons que, pour que le biais asymptotique soit nul sous les deux approches, nous imposons que la taille d'échantillon soit suffisamment grande dans chaque classe d'imputation. Dans ce qui suit, nous supposons donc que le biais de $\hat{R}_I$ est négligeable.

3. Estimation de la variance : cadre à deux phases

À la présente section, nous obtenons des estimateurs de la variance sous le cadre à deux phases et l'approche MI, suivant la méthode proposée par Särndal (1992), ainsi que par Deville et Särndal (1994). En utilisant la décomposition (2.8), la variance totale de $\hat{R}_I$ peut être approximée par

$$V_{mpqI}(\hat{R}_I - R) \approx E_{mpqI}(\hat{R}_I - R)^2 = V_{\mathrm{SAM}} + V_{\mathrm{NR}} + \tilde{V}_I + 2V_{\mathrm{MIX}}, \tag{3.1}$$

où $V_{\mathrm{SAM}} = E_m V_p(\hat{R}) = E_m(V_{\mathrm{SAM}})$ est la variance d'échantillonnage de l'estimateur sous données complètes $\hat{R}$, $V_{\mathrm{NR}} = E_{pq} V_m(E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$ est la variance de non-réponse de l'estimateur imputé $\hat{R}_I$, $\tilde{V}_I = E_{mpq} V_I(\hat{R}_I - E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) \mid s, s_r^{(y)}, s_r^{(x)})$ est la variance due à l'imputation de l'estimateur imputé $\hat{R}_I$, et $V_{\mathrm{MIX}} = E_{pqm}[(E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R})(\hat{R} - R) \mid s, s_r^{(y)}, s_r^{(x)}]$ une composante mixte. Notons que l'expression (3.1) ne contient qu'un seul terme de produit croisé, $2V_{\mathrm{MIX}}$, parce que tous les autres sont asymptotiquement nuls.

3.1 Estimation de la variance d'échantillonnage V_{SAM}

Soit $\hat{V}_{\mathrm{ORD}}$ l'estimateur de variance naïf de $\hat{R}_I$, c'est-à-dire l'estimateur de variance obtenu en traitant les valeurs imputées comme s'il s'agissait de valeurs observées. Nous parvenons à cet estimateur en remplaçant e_i par $\tilde{e}_i = 1/\hat{X}_I(\tilde{y}_i - \hat{R}_I \tilde{x}_i)$ dans (2.2), ce qui donne

$$\hat{V}_{\mathrm{ORD}} = \sum_{i \in s} \sum_{j \in s} \Delta_{ij} \tilde{e}_i \tilde{e}_j. \tag{3.2}$$

Comme nous le montrons maintenant dans le cas de l'échantillonnage simple sans remise, $\hat{V}_{\mathrm{ORD}}$ surestime V_{SAM} sous l'imputation HDAM quand $\sigma_{\varepsilon\eta} > 0$ (comme cela est habituellement le cas en pratique). Après certains calculs algébriques, nous obtenons

$$E_{ml}(\hat{V}_{\mathrm{ORD}} - \hat{V}_{\mathrm{SAM}} \mid s, s_r^{(y)}, s_r^{(x)}) \approx \frac{2}{\mu_x^2}\left(\frac{\mu_y}{\mu_x}\right)\left(1 - \frac{n}{N}\right)\left(1 - \frac{r_{xy}}{n}\right)\frac{\sigma_{\varepsilon\eta}}{n}. \tag{3.3}$$

L'expression (3.3) montre que $\hat{V}_{\mathrm{ORD}}$ est $mpqI$ - biaisé pour $\hat{V}_{\mathrm{SAM}}$, à moins que $\sigma_{\varepsilon\eta} = 0$, $r_{xy} = n$ (ce qui est le cas des données complètes) ou que $n = N$ (ce qui est le cas d'un recensement). Le fait que $\hat{V}_{\mathrm{ORD}}$ ne soit pas un estimateur valide de V_{SAM} s'explique facilement en notant que, si l'imputation HDAM préserve la variabilité, c'est-à-dire s_x^2 et s_y^2, correspondant aux variables x et y, elle ne

préserve pas la covariance, s_{xy}, dans (2.3). En effet, l'imputation a tendance à sous-estimer la relation entre les variables qui sont positivement corrélées. Donc, $\hat{V}_{\text{ORD}}$ surestime V_{SAM}, à cause de la présence du signe moins devant s_{xy} dans (2.3). Pour surmonter cette difficulté, Särndal (1992) a proposé d'estimer $V_{\text{DIF}} = E_{mI}(\hat{V}_{\text{SAM}} - \hat{V}_{\text{ORD}} \mid s, s_r^{(y)}, s_r^{(x)})$ au moyen d'un estimateur mI - sans biais, $\hat{V}_{\text{DIF}}$; c'est-à-dire $E_{mI}(\hat{V}_{\text{DIF}} \mid s, s_r^{(y)}, s_r^{(x)}) = V_{\text{DIF}}$. Toutefois, le calcul de cette composante pour un plan arbitraire comporte des opérations algébriques fastidieuses dans le cas d'un ratio. Par conséquent, nous proposons une solution de rechange qui ne nécessite aucun calcul, mais comprend la construction d'un nouvel ensemble de valeurs imputées. Elle peut être décrite comme suit : chaque fois que $a_i = 0$ et (ou) $b_i = 0$, on choisit un donneur j aléatoirement avec remise à partir de l'ensemble de répondants aux deux variables y *et* x (c'est-à-dire l'ensemble d'unités échantillonnées pour lesquelles $a_i = 1$ et $b_i = 1$) avec la probabilité $w_j / \sum_{l \in s} w_l a_l b_l$ et on impute le vecteur (x_j, y_j). Autrement dit, chaque fois qu'une réponse manque pour une variable, la valeur observée est écartée et dite manquante; les valeurs manquantes sont ensuite remplacées par les valeurs d'un donneur sélectionné au hasard parmi l'ensemble de répondants aux deux variables x et y (souvent appelé ensemble de donneurs communs). De même, lorsque les valeurs des deux variables manquent, le vecteur (x_j, y_j) d'un donneur j est imputé. Puis, on applique l'estimateur de variance standard (2.2) valide sous réponse complète en utilisant ces valeurs imputées. Soit $\hat{V}_{\text{ORD}}^{*}$ l'estimateur de variance résultant. Notons que ce nouvel ensemble de valeurs imputées est utilisé pour obtenir un estimateur valide de la variance d'échantillonnage, mais ne l'est pas pour estimer le paramètre d'intérêt R. On peut montrer que $\hat{V}_{\text{ORD}}^{*}$ est un estimateur de V_{SAM} asymptotiquement $mpqI$ - sans biais. En pratique, on pourrait, par exemple, créer un fichier d'estimation de la variance contenant le nouvel ensemble de valeurs imputées et utiliser l'un des systèmes standard d'estimation de la variance (employés dans le cas de données complètes) pour obtenir l'estimation de la variance d'échantillonnage.

3.2 Estimation de la variance de non-réponse V_{NR}

Un estimateur $\hat{V}_{\text{NR}}$ de $V_{\text{NR}} = E_{pq} V_m(E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$ peut s'obtenir en estimant simplement $V_m(E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$. Par un développement en série de Taylor de premier ordre, nous arrivons à

$$
\begin{aligned}
V_{\text{NR}} \approx\ & \frac{1}{\mu_x^2}\left\{\left[\frac{\sum_{i\in s} w_i^2 a_i}{\hat{N}_a^2} + \frac{\sum_{i\in s} w_i^2}{\hat{N}^2} - 2\frac{\sum_{i\in s} w_i^2 a_i}{\hat{N}\hat{N}_a}\right]\sigma_\varepsilon^2\right. \\
& + \left[\frac{\sum_{i\in s} w_i^2 b_i}{\hat{N}_b^2} + \frac{\sum_{i\in s} w_i^2}{\hat{N}^2} - 2\frac{\sum_{i\in s} w_i^2 b_i}{\hat{N}\hat{N}_b}\right]\left(\frac{\mu_y}{\mu_x}\right)^2\sigma_\eta^2 \\
& \left. -2\left[\frac{\sum_{i\in s} w_i^2 a_i b_i}{\hat{N}_a\hat{N}_b} + \frac{\sum_{i\in s} w_i^2}{\hat{N}^2} - \frac{\sum_{i\in s} w_i^2 a_i}{\hat{N}\hat{N}_a} - \frac{\sum_{i\in s} w_i^2 b_i}{\hat{N}\hat{N}_b}\right]\left(\frac{\mu_y}{\mu_x}\right)\sigma_{\varepsilon\eta}\right\}, \quad (3.4)
\end{aligned}
$$

où $(\hat{N}, \hat{N}_a, \hat{N}_b) = \sum_{i\in s} w_i (1, a_i, b_i)$. Maintenant, posons que $s_{Ix}^2 = 1/\hat{N}\sum_{i\in s} w_i (\tilde{x}_i - \bar{x}_I)^2$ et $s_{Iy}^2 = 1/\hat{N}\sum_{i\in s} w_i (\tilde{y}_i - \bar{y}_I)^2$ avec $(\bar{x}_I, \bar{y}_I) = 1/\hat{N}\sum_{i\in s} w_i (\tilde{x}_i, \tilde{y}_i)$. Notons que s_{Ix}^2 et s_{Iy}^2 désignent respectivement la variabilité d'échantillon des valeurs de x et des valeurs de y après l'imputation. Nous pouvons montrer que s_{Ix}^2 et s_{Iy}^2 sont, respectivement, asymptotiquement mI - sans biais pour les variances sous le modèle σ_η^2 et σ_ε^2. En outre, posons que $s_{xyr} = 1/\hat{N}_{ab}\sum_{i\in s} w_i a_i b_i (x_i - \bar{x}_{rr})(y_i - \bar{y}_{rr})$, où $\hat{N}_{ab} = \sum_{i\in s} w_i a_i b_i$ et $(\bar{x}_{rr}, \bar{y}_{rr}) = \hat{N}_{ab}^{-1}\sum_{i\in s} w_i a_i b_i (x_i, y_i)$. Notons que s_{xyr} est m - sans biais pour la covariance sous le modèle $\sigma_{\varepsilon\eta}$. Il s'ensuit que $\hat{V}_{\text{NR}}$ s'obtient par estimation des quantités inconnues dans (3.4), ce qui mène à

$$
\begin{aligned}
\hat{V}_{\text{NR}} =\ & \frac{1}{\bar{x}_I^2}\left\{\left[\frac{\sum_{i\in s} w_i^2 a_i}{\hat{N}_a^2} + \frac{\sum_{i\in s} w_i^2}{\hat{N}^2} - 2\frac{\sum_{i\in s} w_i^2 a_i}{\hat{N}\hat{N}_a}\right]s_{Iy}^2\right. \\
& + \left[\frac{\sum_{i\in s} w_i^2 b_i}{\hat{N}_b^2} + \frac{\sum_{i\in s} w_i^2}{\hat{N}^2} - 2\frac{\sum_{i\in s} w_i^2 b_i}{\hat{N}\hat{N}_b}\right]\hat{R}_I^2 s_{Ix}^2 \\
& \left. -2\left[\frac{\sum_{i\in s} w_i^2 a_i b_i}{\hat{N}_a\hat{N}_b} + \frac{\sum_{i\in s} w_i^2}{\hat{N}^2} - \frac{\sum_{i\in s} w_i^2 a_i}{\hat{N}\hat{N}_a} - \frac{\sum_{i\in s} w_i^2 b_i}{\hat{N}\hat{N}_b}\right]\hat{R}_I s_{xyr}\right\}. \quad (3.5)
\end{aligned}
$$

L'estimateur (3.5) est asymptotiquement $mpqI$ - sans biais pour V_{NR}. Dans le cas particulier de l'échantillonnage aléatoire simple sans remise, l'expression (3.5) se réduit à

$$\hat{V}_{\mathrm{NR}} = \frac{1}{\bar{x}_I^2}\left[\left(1-\frac{r_y}{n}\right)\frac{s_{Iy}^2}{r_y} + \hat{R}_I^2\left(1-\frac{r_x}{n}\right)\frac{s_{Ix}^2}{r_x} - 2\hat{R}_I\left(1-\frac{r_x r_y}{n r_{xy}}\right)\left(\frac{r_{xy}}{r_x}\right)\left(\frac{r_{xy}}{r_y}\right)\frac{s_{xyr}}{r_{xy}}\right].$$

3.3 Estimation de la variance d'imputation $\tilde{V}_I$

Un estimateur $\hat{V}_I$ de $\tilde{V}_I = E_{mpq} V_I(\hat{R}_I - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$ peut s'obtenir en estimant simplement $V_I(\hat{R}_I - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$ donnée par (2.6). Un estimateur asymptotiquement I - sans biais de $V_I(\hat{R}_I - \hat{R} \mid s, s_r^{(y)}, s_r^{(x)})$ est alors donné par

$$\hat{V}_I = \frac{1}{\bar{x}_I^2}\left[\sum_{i\in s} w_i^2 (1-a_i) s_{Iy}^2 + \hat{R}_I^2 \sum_{i\in s} w_i^2 (1-b_i) s_{Ix}^2\right]. \quad (3.6)$$

Il en découle que $\hat{V}_I$ dans (3.6) est asymptotiquement $mpqI$ - sans biais pour $\tilde{V}_I$. Dans le cas particulier de l'échantillonnage aléatoire simple sans remise, l'expression (3.6) se réduit à

$$\hat{V}_I = \frac{N^2}{\bar{x}_I^2}\left[\left(1-\frac{r_y}{n}\right)\frac{s_{Iy}^2}{n} + \hat{R}_I^2\left(1-\frac{r_x}{n}\right)\frac{s_{Ix}^2}{n}\right].$$

3.4 Estimation de la composante mixte V_{MIX}

Enfin, nous obtenons un estimateur $\hat{V}_{\mathrm{MIX}}$ de V_{MIX} en estimant

$$E_m[(E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R})(\hat{R} - R) \mid s, s_r^{(y)}, s_r^{(x)}].$$

Grâce à un développement en série de Taylor de premier ordre, nous obtenons

$$E_m[(E_I(\hat{R}_I \mid s, s_r^{(y)}, s_r^{(x)}) - \hat{R})(\hat{R} - R) \mid s, s_r^{(y)}, s_r^{(x)}]$$

$$\approx \frac{1}{\mu_x^2}\left\{\left[\frac{\sum_{i\in s} w_i^2 a_i}{\hat{N}\hat{N}_a} - \frac{\sum_{i\in s} w_i^2}{\hat{N}^2}\right]\sigma_\varepsilon^2 + \left[\frac{\sum_{i\in s} w_i^2 b_i}{\hat{N}\hat{N}_b} - \frac{\sum_{i\in s} w_i^2}{\hat{N}^2}\right]\left(\frac{\mu_y}{\mu_x}\right)^2\sigma_\eta^2 - 2\left[\frac{\sum_{i\in s} w_i^2 a_i}{\hat{N}\hat{N}_a} + \frac{\sum_{i\in s} w_i^2 b_i}{\hat{N}\hat{N}_b} - 2\frac{\sum_{i\in s} w_i^2}{\hat{N}^2}\right]\left(\frac{\mu_y}{\mu_x}\right)\sigma_{\varepsilon\eta}\right\}. \quad (3.7)$$

Un estimateur de (3.7) est donc donné par

$$\hat{V}_{\mathrm{MIX}} = \frac{1}{\bar{x}_I^2}\left\{\left[\frac{\sum_{i\in s} w_i^2 a_i}{\hat{N}\hat{N}_a} - \frac{\sum_{i\in s} w_i^2}{\hat{N}^2}\right]s_{Iy}^2 + \left[\frac{\sum_{i\in s} w_i^2 b_i}{\hat{N}\hat{N}_b} - \frac{\sum_{i\in s} w_i^2}{\hat{N}^2}\right]\hat{R}_r^2 s_{Ix}^2 - 2\left[\frac{\sum_{i\in s} w_i^2 a_i}{\hat{N}\hat{N}_a} + \frac{\sum_{i\in s} w_i^2 b_i}{\hat{N}\hat{N}_b} - 2\frac{\sum_{i\in s} w_i^2}{\hat{N}^2}\right]\hat{R}_I s_{xyr}\right\}. \quad (3.8)$$

L'estimateur (3.8) est asymptotiquement $mpqI$ - sans biais pour V_{MIX}. Dans le cas de l'échantillonnage aléatoire simple sans remise, la composante $\hat{V}_{\mathrm{MIX}}$ est nulle. De façon plus générale, la composante $\hat{V}_{\mathrm{MIX}}$ est nulle pour tout plan de sondage autopondéré à une étape (c'est-à-dire un plan d'échantillonnage pour lequel tous les poids d'échantillonnage sont égaux). Dans le cas de plans d'échantillonnage avec probabilités inégales, il est important d'inclure la composante $\hat{V}_{\mathrm{MIX}}$, parce que sa contribution (positive ou négative) à la variance globale peut être importante (Brick, Kalton et Kim (2004)).

Enfin, un estimateur asymptotiquement $mpqI$ - sans biais de la variance totale $V_{\mathrm{TOT}} = V_{mpqI}(\hat{R}_I - R)$ est donc donné par

$$\hat{V}_{\mathrm{TOT}}^{(\mathrm{TP})} = \hat{V}_{\mathrm{ORD}}^{*} + \hat{V}_{\mathrm{NR}} + \hat{V}_I + 2\hat{V}_{\mathrm{MIX}}.$$

4. Estimation de la variance : cadre inversé

Nous allons maintenant obtenir des estimateurs de la variance sous le cadre inversé et l'approche MN ainsi que MI, selon la méthode proposée par Shao et Steel (1999). Rappelons que, sous ce cadre, nous devons formuler l'hypothèse supplémentaire que les probabilités de réponse ne dépendent pas de l'échantillon s. Sous l'approche MN, la variance totale de $\hat{R}_I$ peut être approximée par

$$V(\hat{R}_I - R) \approx E_q V_p E_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}) + E_{pq} V_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}) + V_q E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}), \quad (4.1)$$

où $\mathbf{a} = (a_1, \ldots, a_N)'$ et $\mathbf{b} = (b_1, \ldots, b_N)'$ désignent les vecteurs des indicateurs de réponse aux variables y et x, respectivement.

Sous l'approche MI, la variance totale de $\hat{R}_I$ peut être approximée par

$$V(\hat{R}_I - R) \approx E_{mq} V_p E_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}) + E_{mpq} V_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}) + E_q V_m E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}). \quad (4.2)$$

Sous l'approche MN ainsi que sous l'approche MI, nous pouvons obtenir un estimateur du premier terme du deuxième membre de (4.1) et de (4.2) en trouvant un estimateur asymptotiquement pI - sans biais de $V_p E_I(\hat{R}_I|\mathbf{a}, \mathbf{b})$. En outre, nous pouvons estimer le deuxième terme du deuxième membre de (4.1) et de (4.2) par $\hat{V}_I$ donné par (3.6). Sous l'approche MN, nous pouvons obtenir un estimateur du dernier terme du deuxième membre de (4.1) en estimant $V_q E_{pI}(\hat{R}_I \mid \mathbf{a}, \mathbf{b})$, tandis que sous l'approche MI, nous pouvons obtenir un estimateur du dernier terme du deuxième membre de (4.2) en estimant $V_m E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$. Par conséquent, les estimateurs des deux premiers termes de (4.1) et de (4.2) sont identiques et donc valides, quelle que soit l'approche (MN ou MI) utilisée pour l'inférence. Seul le troisième terme du deuxième membre de (4.1) et de (4.2) dépendra de l'approche choisie. Dans le cas de l'approche MI, la spécification et la validation du modèle d'imputation sont essentielles à l'obtention de l'absence de biais de la troisième composante, tandis que dans l'approche MN, l'absence de biais de la troisième composante dépend de la spécification correcte du modèle de non-réponse.

4.1 Estimation de $V_p E_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$

Partant d'un développement en série de Taylor du premier ordre et de l'expression (2.5), un estimateur de $V_p E_I(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$, dénoté par $\hat{V}_1$, est donné par

$$\hat{V}_1 = \sum_{i\in s}\sum_{j\in s} \Delta_{ij}\, \xi_i\, \xi_j, \qquad (4.3)$$

où

$$\xi_i = \frac{1}{\bar{x}_r}\left[\frac{1}{\hat{N}_a} a_i (y_i - \bar{y}_r) - \left(\frac{\bar{y}_r}{\bar{x}_r}\right)\frac{1}{\hat{N}_b} b_i (x_i - \bar{x}_r)\right].$$

Autrement dit, l'estimateur $\hat{V}_1$ est obtenu à partir de l'estimateur de variance sous données complètes (2.2) en remplaçant e_i par ξ_i. Dans le cas de l'échantillonnage aléatoire simple sans remise, l'estimateur (4.3) se réduit à

$$\hat{V}_1 = \left(1 - \frac{n}{N}\right)\frac{1}{\bar{x}_r^2}\left[\frac{s_{yr}^2}{r_y} + \hat{R}_r^2\frac{s_{xr}^2}{r_x} - 2\hat{R}_r\left(\frac{r_{xy}}{r_x}\right)\left(\frac{r_{xy}}{r_y}\right)\frac{s_{xyr}}{r_{xy}}\right]. \qquad (4.4)$$

4.2 Estimation de $V_q E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ sous l'approche MN

Pour commencer, notons

$$E_{pI}(\hat{R}_I) \approx \frac{Y_a N_b}{N_a X_b},$$

où $(Y_a, N_a) = \sum_{i\in U} a_i (y_i, 1)$ et $(X_b, N_b) = \sum_{i\in U} b_i (x_i, 1)$.

Par un développement en série de Taylor de premier ordre, nous pouvons montrer que $V_q E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ peut être approximé par

$$V_q E_{pI}(\hat{R}_I - R|\mathbf{a},\mathbf{b}) \approx \frac{1}{N\bar{X}^2}\left[\left(\frac{1-p_y}{p_y}\right)S_y^2 + R^2\left(\frac{1-p_x}{p_x}\right)S_x^2 - 2R\left(\frac{p_{xy} - p_x p_y}{p_x p_y}\right)S_{xy}\right], \qquad (4.5)$$

où

$$S_y^2 = \frac{1}{N-1}\sum_{i\in U}(y_i - \bar{Y})^2,$$

$$S_x^2 = \frac{1}{N-1}\sum_{i\in U}(x_i - \bar{X})^2$$

et

$$S_{xy} = \frac{1}{N-1}\sum_{i\in U}(x_i - \bar{X})(y_i - \bar{Y})$$

avec

$$(\bar{Y}, \bar{X}) = \frac{1}{N}\sum_{i\in U}(y_i, x_i).$$

Nous obtenons un estimateur de $V_q E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ en estimant les quantités inconnues dans (4.5), ce qui donne

$$\hat{V}_2^{(\mathrm{NM})} = \frac{1}{\hat{N}\bar{x}_I^2}\left[\left(\frac{1-\hat{p}_y}{\hat{p}_y}\right)s_{Iy}^2 + \hat{R}_I^2\left(\frac{1-\hat{p}_x}{\hat{p}_x}\right)s_{Ix}^2 - 2\hat{R}_I\left(\frac{\hat{p}_{xy} - \hat{p}_x\hat{p}_y}{\hat{p}_x\hat{p}_y}\right)s_{xyr}\right]$$

$$= \frac{1}{\bar{x}_I^2}\left[\left(\frac{1}{\hat{N}_a} - \frac{1}{\hat{N}}\right)s_{Iy}^2 + \hat{R}_I^2\left(\frac{1}{\hat{N}_b} - \frac{1}{\hat{N}}\right)s_{Ix}^2 - 2\hat{R}_I\left(\frac{\hat{N}_{ab}}{\hat{N}_b\hat{N}_b} - \frac{1}{N}\right)s_{xyr}\right], \qquad (4.6)$$

où $\hat{p}_y = \frac{\hat{N}_a}{\hat{N}}$, $\hat{p}_x = \frac{\hat{N}_b}{\hat{N}}$ et $\hat{p}_{xy} = \frac{\hat{N}_{ab}}{\hat{N}}$.

L'estimateur (4.6) est asymptotiquement pqI - sans biais pour la variance approximative (4.5), en notant que s_{Iy}^2, s_{Ix}^2 et s_{xyr} sont asymptotiquement pqI - sans biais pour S_y^2, S_x^2 et S_{xy}, respectivement. Dans le cas de l'échantillonnage aléatoire simple sans remise, l'estimateur (4.6) se réduit à

$$\hat{V}_2^{(\mathrm{NM})} = \frac{1}{\bar{x}_I^2}\left[\left(\frac{n}{N} - \frac{r_y}{N}\right)\frac{s_{Iy}^2}{r_y} + \hat{R}_I^2\left(\frac{n}{N} - \frac{r_x}{N}\right)\frac{s_{Ix}^2}{r_x} - 2\hat{R}_I\left(\frac{n}{N} - \frac{r_x r_y}{N r_{xy}}\right)\left(\frac{r_{xy}}{r_x}\right)\left(\frac{r_{xy}}{r_y}\right)\frac{s_{xyr}}{r_{xy}}\right]. \qquad (4.7)$$

4.3 Estimation de $E_q V_m E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ sous l'approche MI

Par un développement en série de Taylor de premier ordre, nous pouvons montrer que $E_q V_m E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ peut être approximé par

$$E_q V_m E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b}) \approx \frac{1}{\mu_x^2}\left[\left(\frac{1}{E_q(N_a)} - \frac{1}{N}\right)\sigma_\varepsilon^2 + \left(\frac{\mu_y}{\mu_x}\right)^2\left(\frac{1}{E_q(N_b)} - \frac{1}{N}\right)\sigma_\eta^2 - 2\left(\frac{\mu_y}{\mu_x}\right)\left(E_q\left(\frac{N_{ab}}{N_b N_b}\right) - \frac{1}{N}\right)\sigma_{\varepsilon\eta}\right], \quad (4.8)$$

où $N_{ab} = \sum_{i \in U} a_i b_i$. Nous obtenons un estimateur de $E_q V_m E_{pI}(\hat{R}_I - R \mid \mathbf{a}, \mathbf{b})$ en estimant les quantités inconnues dans (4.8), ce qui mène à

$$\hat{V}_2^{(\mathrm{IM})} = \frac{1}{\bar{x}_I^2}\left[\left(\frac{1}{\hat{N}_a} - \frac{1}{\hat{N}}\right)s_{Iy}^2 + \hat{R}_I^2\left(\frac{1}{\hat{N}_b} - \frac{1}{\hat{N}}\right)s_{Ix}^2 - 2\hat{R}_I\left(\frac{\hat{N}_{ab}}{\hat{N}_b \hat{N}_b} - \frac{1}{N}\right)s_{xyr}\right]. \quad (4.9)$$

L'estimateur (4.9) est asymptotiquement *mpqI* - sans biais pour la variance approximative (4.8). Il est intéressant de souligner que, sous l'imputation HDAM pondérée, l'estimateur $\hat{V}_2^{(\mathrm{NM})}$ dans (4.6) obtenu sous l'approche MN est identique à l'estimateur $\hat{V}_2^{(\mathrm{IM})}$ dans (4.9) obtenu sous l'approche MI. Toutefois, il pourrait ne pas en être ainsi sous une méthode d'imputation différente. En outre, la composante $\hat{V}_2$ est négligeable par rapport à $\hat{V}_1$ quand la fraction d'échantillonnage n/N est négligeable, où $\hat{V}_2$ représente $\hat{V}_2^{(\mathrm{NM})}$ ou $\hat{V}_2^{(\mathrm{IM})}$. Dans ce cas, la composante $\hat{V}_2$ peut être omise des calculs.

Enfin, un estimateur de la variance totale sous le cadre inversé est donné par

$$\hat{V}_{\mathrm{TOT}}^{(\mathrm{RE})} = \hat{V}_1 + \hat{V}_I + \hat{V}_2.$$

Sous le cadre inversé, les approches MN et MI mènent toutes deux au même estimateur de la variance totale. Donc, l'estimateur de variance $\hat{V}_{\mathrm{TOT}}^{(\mathrm{RE})}$ est robuste en ce sens qu'il est valide sous l'approche MN ou MI.

5. Résumé et conclusion

Dans le présent article, nous avons obtenu des estimateurs de variance pour l'estimateur imputé d'un ratio sous deux cadres différents. Le cadre inversé facilite la détermination des expressions de la variance (comparativement au cadre à deux phases habituel), particulièrement si la fraction d'échantillonnage est faible, auquel cas nous pouvons omettre la composante $\hat{V}_2$. Toutefois, contrairement au cadre à deux phases, il nécessite l'hypothèse supplémentaire voulant que les probabilités de réponse ne dépendent pas de l'échantillon réalisé s. De plus, le cadre à deux phases s'appuie sur une décomposition naturelle de l'erreur totale qui aboutit à une décomposition naturelle de la variance totale. Par conséquent, cette dernière peut être exprimée comme la somme de la variance d'échantillonnage, de la variance de non-réponse et de la variance d'imputation, ce qui permet au statisticien d'enquête de se faire une idée de l'importance relative de chaque composante. Sous l'approche inverse, il n'existe aucune interprétation simple des composantes de la variance (sauf la variance d'imputation).

Nous avons considéré le cas de l'imputation HDAM pondérée dans les classes. Une autre version de l'imputation hot deck aléatoire pondérée, que nous appelons imputation hot deck aléatoire conjointe (HDAC) pondérée est identique à l'imputation HDAM pondérée, excepté que, lorsque les données manquent pour les deux variables, un donneur j est sélectionné au hasard parmi l'ensemble de donneurs communs (c'est-à-dire l'ensemble de répondants aux deux variables y et x) avec la probabilité $w_j / \sum_{l \in s} w_l a_l b_l$ et le vecteur (x_j, y_j) est imputé. Cette version de la méthode aide à préserver les relations entre les variables étudiées, contrairement à l'imputation indépendante de chaque variable. Les résultats de l'imputation HDAC peuvent être obtenus en utilisant des techniques comparables à celles présentées ici. Enfin, les résultats exposés dans le présent article peuvent être généralisés facilement à l'imputation par la régression tant déterministe qu'aléatoire dans les classes d'imputation.

Remerciements

L'auteur remercie un Éditeur Associé de ses suggestions et commentaires constructifs qui ont permis d'améliorer la qualité de l'article. Les travaux de recherché de David Haziza ont été financés par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada.

Bibliographie

Brick, M.J., 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.

Deville, J.-C., et Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.

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.

Särndal, C.-E. (1992). Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation. *Techniques d'enquête*, 18, 257-268.

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

Techniques d'enquête, décembre 2007
Vol. 33, N° 2, pp. 187-193
Statistique Canada, N° 12-001-X au catalogue

Bootstrap efficace pour les enquêtes-entreprises

James Chipperfield et John Preston [1]

Résumé

L'Australian Bureau of Statistics vient de développer un système généralisé d'estimation pour traiter les données de ses enquêtes-entreprises annuelles et infra-annuelles de grande portée. Les plans de sondage de ces enquêtes comportent d'un grand nombre de strates, un échantillonnage aléatoire simple dans les strates, des fractions d'échantillonnage non négligeables, ainsi qu'un chevauchement d'échantillons pour des périodes consécutives et peuvent faire l'objet de modifications de la base de sondage. Un défi important consistait à choisir la méthode d'estimation de la variance répondant le mieux aux critères suivants : être valide pour une grande gamme d'estimateurs (par exemple, ratio et régression généralisée), nécessiter un temps de calcul limité, être facilement adaptable à divers plans de sondage et estimateurs, et avoir de bonnes propriétés théoriques en ce qui concerne le biais et la variance. Le présent article décrit le bootstrap rééchelonné sans remise (BRSR) qui a été mis en œuvre par l'ABS et montre qu'il est sensiblement plus efficace que le bootstrap rééchelonné avec remise (BRAR) de Rao et Wu (1988). Les principaux avantages du bootstrap comparativement à d'autres estimateurs de variance par rééchantillonnage sont son efficacité (c'est-à-dire son exactitude par unité d'espace de mémorisation) et la simplicité relative avec laquelle il peut être spécifié dans un système. Le présent article décrit l'estimateur de variance du bootstrap BRSR pour les estimations ponctuelles et les estimations des changements qui peut être exprimé comme une fonction des moyennes de population finie. Les résultats des simulations entreprises dans le cadre du processus d'évaluation montrent que le BRSR est plus efficace que le BRAR, particulièrement dans les situations où la taille des échantillons dans les strates peut être aussi petite que 5.

Mots clés : Variance; bootstrap; échantillonnage stratifié.

1. Introduction

En 2000, l'Australian Bureau of Statistics (ABS) a commencé par obtenir auprès de l'Australian Taxation Office (ATO) un registre des entreprises contenant des données fiscales. Les éléments de données incluaient le chiffre d'affaires, le chiffre des ventes et d'autres postes de dépenses. En 2001, l'ABS a utilisé ce registre comme base de sondage pour certaines enquêtes afin d'accroître l'efficacité de ses plans d'échantillonnage. Les données du registre sont mises à jour au moins annuellement pour chaque entreprise. Afin de maximiser l'utilisation de ces éléments de données administratives à l'étape de l'estimation, l'ABS a développé un système généralisé d'estimation appelé ABSEST capable de prendre en charge l'estimation par la régression généralisée (GREG) et l'estimation de la variance. Depuis juillet 2005, ABSEST a été utilisé systématiquement pour l'enquête mensuelle sur le commerce de détail menée par l'ABS.

Un système généralisé d'estimation est un outil fort précieux pour les organismes statistiques, car il permet de gérer diverses exigences quant aux produits d'une enquête en maintenant une grande rigueur statistique à un coût acceptable. L'ABS a investi d'importantes ressources dans son système généralisé d'estimation pour les enquêtes-entreprises. Avant 1998, ce système permettait de produire des estimations de type Horvitz-Thompson, par le ratio et à deux phases, ainsi que des estimations de la variance fondées sur des approximations par développement en série de Taylor (ST). En 1999, la méthode de développement en série de Taylor a été remplacée par celle du jackknife. Les réactions concernant la conception et la convivialité de l'application informatique ont révélé que les changements apportés au système généralisé d'estimation rendaient sa maintenance et son développement de plus en plus difficiles et que le temps de traitement pouvait être indésirablement long. Ces caractéristiques essentielles ont joué un rôle important dans le choix de la méthode d'estimation de la variance pour ABSEST.

Les données de sortie de base des enquêtes-entreprises de l'ABS sont des estimations ponctuelles, des estimations de changements entre deux points dans le temps, et des estimations de taux. Les enquêtes-entreprises sont réalisées selon un plan d'échantillonnage avec probabilités égales dans les strates comportant un très grand nombre de strates (des centaines), à une ou à deux phases et, pour les enquêtes comportant un échantillonnage à plus d'une occasion, un chevauchement des échantillons pouvant varier de 0 à 100 %. La taille d'échantillon des enquêtes-entreprises va de moins de 1 000 à 15 000; au niveau des strates, elle peut être aussi faible que trois ou atteindre plusieurs centaines.

À la section 2, nous présentons l'estimateur GREG. À la section 3, nous discutons de divers estimateurs de la variance pour l'estimateur GREG et justifions le choix de l'estimateur de variance par le bootstrap pour ABSEST. À la section 4, nous décrivons les estimateurs de variance par

1. James Chipperfield, Australian Bureau of Statistics. Courriel : james.chipperfield@abs.gov.au; John Preston, Australian Bureau of Statistics. Courriel : john.preston@abs.gov.au.

le bootstrap rééchelonné sans remise (BRSR) [en anglais, without replacement scaled bootstrap] et par le bootstrap rééchelonné avec remise (BRAR) [en anglais, with replacement scaled bootstrap] de Rao et Wu (1988) pour les estimations ponctuelles sous des plans de sondage à une phase. À la section 5, nous décrivons le BRSR pour les estimations de changements. À la section 6, nous évaluons les propriétés de biais et de variance du BRSR et du BRAR au moyen d'une étude par simulation. À la section 7, nous présentons certaines conclusions.

2. Estimateur par la régression généralisée (GREG)

À la présente section, nous décrivons brièvement l'estimateur GREG qui est implémenté dans ABSEST. Considérons une population finie U divisée en H strates $U = \{U_1, U_2, ..., U_H\}$, où U_h est constituée de N_h unités. Le total de population finie étudié est $Y = \sum_h Y_h$, où $Y_h = \sum_{i \in U_h} y_{hi}$ et $h = 1, ..., H$. Dans la strate h, l'échantillon s_h de n_h unités est sélectionné à partir de U_h par échantillonnage aléatoire simple sans remise (EASSR). L'ensemble complet d'échantillons est désigné par $s = \{s_1, s_2, ..., s_H\}$.

Considérons le cas où un vecteur de variables auxiliaires $\boldsymbol{x}_i = (x_{1i}, ..., x_{ki}, ..., x_{Ki})^T$ de dimension K est disponible pour $i \in s$ et le vecteur correspondant de totaux de population $\boldsymbol{X} = \sum_{i \in U} \boldsymbol{x}_i$ est connu. L'estimateur GREG (Särndal, Swensson et Wretman 1992, page 227) est donné par $\hat{Y}_{\text{reg}} = \sum_{i \in s} \tilde{w}_i y_i = \sum_{i \in s} w_i y_i + (\boldsymbol{X} - \hat{\boldsymbol{X}})^T \hat{\boldsymbol{B}}$, où $\tilde{w}_i = w_i g_i$, $w_i = N_h / n_h$, $\hat{\boldsymbol{B}} = \hat{\boldsymbol{T}}^{-1} \hat{\boldsymbol{t}}$ avec $\hat{\boldsymbol{T}}^{-1}$ représentant l'inverse généralisée de $\hat{\boldsymbol{T}}$, $\hat{\boldsymbol{X}} = \sum_{i \in s} w_i \boldsymbol{x}_i$ et $g_i = (1 + \sigma_i^{-2} \boldsymbol{x}_i^T \hat{\boldsymbol{T}}^{-1} (\boldsymbol{X} - \hat{\boldsymbol{X}})^T)$, $\hat{\boldsymbol{T}} = \sum_{i \in s} w_i \boldsymbol{x}_i \boldsymbol{x}_i^T \sigma_i^{-2}$, $\hat{t} = \sum_{i \in s} w_i \boldsymbol{x}_i y_i \sigma_i^{-2}$, σ_i^2 est une constante motivée par le modèle de superpopulation $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i$ telle que les résidus ε_i sont indépendants identiquement distribués de moyenne 0 et de variance σ_i^2, et $E(\hat{\boldsymbol{B}}) = \boldsymbol{\beta}$. Il est bien connu que $\hat{Y}_{\text{reg}}$ est sans biais jusqu'à l'ordre $O(n^{-1})$. Les poids $\tilde{w}_i$ sont mémorisés pour permettre le calcul rapide des estimations. En pratique, des bornes sont imposées aux poids, $\tilde{w}_i$. Si les poids, $\tilde{w}_i$, donnés par l'équation susmentionnée se situent en dehors de ces bornes, ils sont calculés par itération (voir la méthode 5 de Singh et Mohl 1996).

L'expression de $\hat{Y}_{\text{reg}}$ peut être adaptée à une gamme d'estimations, y compris des estimations de domaines et des estimations à plusieurs phases (voir Estevao, Hidiroglou et Särndal 1995). Par exemple, si $\boldsymbol{x}_i = 1$, $\hat{Y}_{\text{reg}}$ devient l'estimateur d'Horvitz-Thompson donné par $\hat{Y} = \sum_h w_h \sum_{i \in s_h} y_{hi}$ avec une variance estimée $\widehat{\text{Var}}(\hat{Y}) = \sum_h N_h^2 n_h^{-1} (1 - f_h) \hat{s}_h^2$, où $\hat{s}_h^2 = (n_h - 1)^{-1} \sum_{i \in s_h} (y_{hi} - \hat{\bar{y}}_h)^2$, $\hat{\bar{y}}_h = \sum_{i \in s_h} y_{hi} / n_h$ et $f_h = n_h / N_h$.

3. Comparaison de divers estimateurs de variance

La méthode d'estimation de la variance adoptée pour ABSEST devait avoir, dans les études par simulation, de bonnes propriétés de biais et de variance comparativement aux options décrites dans la littérature. Afin de simplifier la maintenance et le développement du système, il fallait que les spécifications du système d'estimation de la variance soient génériques, afin que tous les calculs soient en grande partie indépendants de l'estimateur. (ABSEST ne doit être capable de prendre en charge que l'EASSR dans les strates et les plans d'échantillonnage à un degré.) Enfin, la réduction au minimum des coûts de calcul était une autre considération importante.

En premier lieu, nous avons également envisagé les méthodes du bootstrap, du jackknife et des répliques équilibrées répétées (BRR pour Balanced Repeated Replication) (Shao et Tu 1995, Rao et Wu 1988). Considérons l'estimation de la variance d'une fonction $\hat{\theta} = \theta(\hat{\boldsymbol{Y}})$, où $\hat{\boldsymbol{Y}}$ est un vecteur de dimension P d'estimations $\hat{\boldsymbol{Y}} = \sum_{i \in s} w_i y_i$, $\boldsymbol{y}_i$ est le vecteur réponse provenant de l'unité i ayant les éléments y_{pi}, et θ est une fonction lisse. L'estimation de la variance par une méthode de rééchantillonnage comporte les étapes suivantes :

i) sous-échantillonner indépendamment l'ensemble s un nombre total R de fois;

ii) pour chacun des R sous-échantillons, calculer $w_i^* = b_i^* w_i$, où b_i^* dépend du nombre de fois que l'unité i est sélectionnée dans le sous-échantillon;

iii) calculer $\hat{\theta}^* = \hat{\theta}(\hat{\boldsymbol{Y}}^*)$, où $\hat{\boldsymbol{Y}}^* = \sum_{i \in s} w_i^* \boldsymbol{y}_i$;

iv) estimer la variance de θ par $\hat{\text{Var}}_{\text{rep}}(\hat{\theta}) = (R - 1)^{-1} \sum_{r=1}^{R} (\hat{\theta}^{*(r)} - \hat{\theta})^2$, où $\hat{\theta}^{*(r)}$ est l'estimation de θ basée sur le r^{e} échantillon répété. Nota : l'expression pour les poids de rééchantillonnage, $w_i^* = b_i^* w_i$, comprend le jackknife, le bootstrap et les répliques répétées équilibrées comme cas particuliers.

Comme $\hat{Y}_{\text{reg}}$ peut être exprimé par une fonction $\hat{\theta}$, nous pouvons calculer la variance de $\hat{Y}_{\text{reg}}^*$ suivant les étapes susmentionnées où iii) et iv) deviennent, respectivement : iii) calculer $\hat{Y}_{\text{reg}}^* = \sum_{i \in s} \tilde{w}_i^* y_i$, où $\tilde{w}_i^* = w_i^* g_i^*$ et g_i^* a la même forme que g_i, mais est calculé en utilisant les poids w_i^* au lieu des poids w_i pour $i \in s$; iv) estimer la variance par $\widehat{\text{Var}}_{\text{rep}}(\hat{Y}_{\text{reg}}) = (R - 1)^{-1} \sum_{r=1}^{R} (\hat{Y}_{\text{reg}}^{*(r)} - \hat{Y}_{\text{reg}})^2$.

La caractéristique intéressante de ces méthodes de rééchantillonnage est que seules la sélection des échantillons répétés et la valeur b_i^* sont nécessaires pour calculer des estimations sans biais de la variance pour de nombreux plans d'échantillonnage utilisés couramment et pour des estimateurs ayant de bonnes approximations par développement en série de Taylor au premier ordre. En outre, si l'on

mémorise les poids de rééchantillonnage, $\tilde{w}_i^*$, l'estimation de la variance de $\hat{Y}_{\text{reg}}^*$ ne requiert que de simples calculs qui peuvent se faire rapidement; cette approche de mémorisation des poids de rééchantillonnage a été appliquée avec succès dans le système généralisé d'estimation de l'ABS pour les enquêtes-ménages. Une fois que les poids de rééchantillonnage sont disponibles, l'estimation de la variance pour diverses analyses, comme la régression linéaire, ne demande que des calculs simples qui prennent peu de temps et n'exige pas que l'analyste connaisse le plan d'échantillonnage. Nous allons maintenant examiner les mérites relatifs de certains estimateurs de variance par rééchantillonnage en vue de leur implémentation dans ABSEST.

Le jackknife avec suppression d'une unité forme des échantillons répétés, s^*, en éliminant une unité à la fois. Cela implique que $R = n$. Pour les enquêtes à grande échelle, cette exigence de capacité de mémorisation est excessive. Le jackknife avec suppression d'un groupe, même s'il réduit R en laissant tomber un groupe d'unités à la fois dans une strate, comporterait encore au moins $R = 2H$ répliques, un minimum de deux groupes par strate étant nécessaire pour calculer la variance. Malgré de bons résultats dans une étude empirique où $n_h = 2$ (voir Shao et Tu 1995, page 251), le jackknife a été rejeté étant donné qu'il nécessite une capacité de mémorisation excessive.

Pour les plans stratifiés, la méthode des répliques répétées équilibrées (BRR) rééchelonnées requiert environ $R = H$ poids de rééchantillonnage. Premièrement, les échantillons répétés sont formés par subdivision aléatoire de l'échantillon de strate s_h en deux groupes, suivie de l'affectation de l'un de ces groupes à s_h^* pour chaque $h = 1, ..., H$. L'affectation des groupes aux répliques, définie par une matrice de Hadamard, est exécutée de façon à éliminer la covariance entre strates dans les échantillons répétés. La méthode des répliques répétées équilibrées groupées (GBRR) (voir Shao et Tu 1995) permet de réduire arbitrairement R au prix de l'introduction d'une covariance entre strates dans les échantillons répétés. Preston et Chipperfield (2002) ont montré dans une évaluation empirique, dans le cas d'une enquête-entreprises typique de l'ABS, que les estimations BRR (et GBRR) sont considérablement plus instables que les estimations bootstrap.

Dans leur résumé de la littérature, Kovar, Rao et Wu (1988) constatent que le bootstrap rééchelonné a tendance à produire un biais plus important que le jackknife ou le développement en série de Taylor (ST) lorsqu'on estime la variance des estimations GREG. Comme l'évaluation relative de ces méthodes variait selon le modèle de simulation sous-jacent et la taille de l'échantillon de strate, il était important de procéder à une évaluation fondée sur un modèle et un plan d'échantillonnage typiques des enquêtes-entreprises de l'ABS. À la section 6, nous montrons que ces propriétés sont acceptables. Contrairement aux autres méthodes de rééchantillonnage, il est possible, pour le bootstrap, de choisir arbitrairement la valeur de R, donc, de satisfaire les contraintes de capacité de mémorisation et de calcul. En outre, la sélection des échantillons répétés bootstrap est plus facile à spécifier dans un système informatique que celle des échantillons répétés BRR.

Nous avons considéré les mérites relatifs de plusieurs autres estimateurs de la variance en vue de leur implémentation dans ABSEST. La méthode de développement en série de Taylor (ST) ne convenait pas, parce que l'expression de sa variance pour les paramètres à estimer complexes comporte de nombreux termes particuliers à ces paramètres, ce qui rend difficile son adaptation à un système généralisé. Ce problème a été étudié par Nordberg (2000), qui a décrit une méthode de calcul de la variance qui génère automatiquement des développements en série de Taylor. La méthode a été implémentée dans un système informatique appelé CLAN. Ce dernier peut traiter toute fonction des moyennes sous échantillonnage avec probabilités proportionnelles à la taille (PPT) et sous échantillonnage en grappes. Une limite de CLAN est qu'il ne produit pas de poids de rééchantillonnage permettant d'exécuter une analyse complexe, telle que l'analyse par la régression, que ce soit dans un organisme statistique ou à l'externe; il nécessite la connaissance du plan d'échantillonnage et serait assez difficile à spécifier et à maintenir dans un système généralisé (comparativement au bootstrap). Pour la même raison, d'autres estimateurs de variance linéarisés (décrits dans Estevao, Hidiroglou et Särndal 1995, et évalués dans Yung et Rao 1996) ont été rejetés, malgré de bonnes propriétés théoriques, de bons résultats empiriques et la rapidité des calculs.

Compte tenu des éléments susmentionnés, la méthode d'estimation de la variance privilégiée pour ABSEST était le bootstrap. À la section suivante, nous décrivons le BRSR et le BRAR, dont le premier seulement est implémenté dans ABSEST.

4. Bootstrap rééchelonné sans remise (BRSR) pour les estimations ponctuelles

4.1 Méthode

Pour les estimations GREG ponctuelles, l'estimateur de la variance par le bootstrap rééchelonné sans remise (BRSR) comprend R fois la répétition des étapes suivantes :

a) former l'ensemble s^r en sélectionnant m_h unités par EASSR à partir de s_h indépendamment dans chaque strate $h = 1, ..., H$, où $m_h = [n_h / 2]$ et l'opérateur [.] arrondit son argument au nombre entier le plus proche;

b) calculer $w^*_{hi} = w_{hi}(1 - \gamma_h + \gamma_h\, n_h / m_h\, \delta^*_{hi})$ pour $i \in s_h$, où $\gamma_h = \sqrt{(1 - f_h) m_h / (n_h - m_h)}$, δ^*_{hi} est égal à 1 si $i \in s_h^r$ et à 0 autrement;

c) calculer $\tilde{w}^*_{hi} = w^*_{hi}\, g^*_{hi}$ pour $i \in s$; et

d) calculer la r^{e} estimation bootstrap de Y, $\hat{Y}^*_{\text{reg}} = \sum_{i \in s} \tilde{w}^*_{hi}\, y_i$. La justification de $m_h = [n_h / 2]$ est présentée à la section 4.2. L'estimateur de variance bootstrap est donné par l'approximation de Monte Carlo, $\widehat{\text{Var}}_B(\hat{Y}_{\text{reg}}) = (R-1)^{-1} \sum_{r=1}^{R} (\hat{Y}^{*(r)}_{\text{reg}} - \hat{Y}_{\text{reg}})^2$. La méthode BRAR est identique à la méthode BRSR, excepté que les échantillons répétés sont sélectionnés par EASAR et que le facteur d'échelle devient $\gamma_h = \sqrt{(1-f_h) m_h / (n_h - 1)}$, où m_h est souvent fixé à $n_h - 1$ dans la littérature. Preston et Chipperfield (2002) ont constaté que l'erreur de rééchantillonnage, c'est-à-dire l'erreur due à l'échantillonnage répété et au conditionnement à l'ensemble d'échantillons, est significativement plus faible pour le BRSR que pour le BRAR.

Il est facile de voir que les estimateurs BRSR et BRAR sont des estimateurs sans biais de $\text{Var}(\hat{\theta})$. La variance approximative par développement en série de Taylor est donnée par $\widehat{\text{Var}}(\hat{\theta}) = \nabla' \hat{\theta}\, \hat{V}(\hat{\boldsymbol{Y}}) \nabla \hat{\theta}$, où $\hat{V}(\hat{\boldsymbol{Y}})$ est la matrice de dimensions $P \times P$ contenant les éléments

$$\widehat{\text{Cov}}(\hat{Y}_p, \hat{Y}_{p'}) = \frac{N^2(1-f)}{n}\, \hat{s}_{p,\,p'}$$

où

$$\hat{s}_{p,\,p'} = \frac{1}{n-1} \sum_{i \in s} (y_{pi} - \hat{\bar{y}}_p)(y_{p'i} - \hat{\bar{y}}_{p'});$$

$$\hat{\bar{y}}_p = \frac{1}{n} \sum_{i \in s} y_{pi};$$

$$\hat{Y}_p = \sum_{i \in s} w_i\, y_{pi}$$

pour $p,\ p' = 1, \ldots, P$, et $\nabla' = (\partial / \partial Y_1, \ldots, \partial / \partial Y_P)|_{\hat{\boldsymbol{Y}}}$. Il est aisé de constater que

$$E_*(\widehat{\text{Var}}(\hat{\theta}^*)) = \nabla \hat{\theta}' E_*(\hat{V}(\hat{\boldsymbol{Y}}^*)) \nabla \hat{\theta} \;= \nabla' \hat{\theta}\, \hat{V}(\hat{\boldsymbol{Y}}) \nabla \hat{\theta},$$

en notant que

$$E_*[\widehat{\text{Cov}}(\hat{Y}^*_p,\ \hat{Y}^*_{p'})] = \widehat{\text{Cov}}(\hat{Y}_p,\ \hat{Y}_{p'})$$

où E_* désigne l'espérance sous rééchantillonnage. Notons que les constantes d'échelle appliquées à w_{hi} pour calculer les poids de rééchantillonnage sont choisies de façon à obtenir le facteur de correction de population finie correct. Il s'ensuit que l'approximation de Monte Carlo de la variance, $\widehat{\text{Var}}_B(\hat{\theta}) = (R-1)^{-1} \sum_{r=1}^{R} (\hat{\theta}^{*(r)} - \hat{\theta})^2$, est sans biais pour $\text{Var}(\hat{\theta})$.

4.2 Note sur l'efficacité relative des échantillonnages BRAR et BRSR

Afin de simplifier la notation, posons que $\hat{v}_{\text{boot}} = \widehat{\text{Var}}_B(\hat{\theta})$. La variance de l'estimateur bootstrap de la variance peut s'écrire

$$\text{Var}(\hat{v}_{\text{boot}}) = \text{Var}_s(E_*[\hat{v}_{\text{boot}} \mid s]) + E_s[\text{Var}_*(\hat{v}_{\text{boot}} \mid s)],$$

où s désigne l'espérance par rapport au plan d'échantillonnage. Si $\hat{v}_{\text{boot}}$ est sans biais (c'est-à-dire $E_*[\hat{v}_{\text{boot}} \mid s] = \text{Var}(\hat{\theta})$), alors $\text{Var}(\hat{v}_{\text{boot}})$ ne dépend pas de la façon de sélectionner les échantillons répétés. Le terme $\text{Var}_*(\hat{v}_{\text{boot}} \mid s)$ est l'erreur de rééchantillonnage sachant l'échantillon et est inversement proportionnel à R. La valeur de R est choisie suffisamment grande pour que $\text{Var}_s(E_*[\hat{v}_{\text{boot}} \mid s])$ soit faible comparativement à $\hat{v}_{\text{boot}}$, la variance d'échantillon estimée. L'efficacité des deux estimateurs bootstrap peut être comparée en se basant sur la grandeur de $\text{Var}_s(E_*[\hat{v}_{\text{boot}} \mid s])$ quand les deux estimateurs ont la même valeur de R. Nous allons maintenant résumer les résultats empiriques basés sur des données réelles qui montrent que le BRSR peut être significativement plus efficace que le BRAR. Les gains d'efficacité se traduisent par une réduction du temps de calcul et(ou) l'obtention d'estimations plus exactes de la variance.

Preston et Chipperfield (2002) ont comparé l'efficacité du BRSR avec $m_h = [n_h / 2]$ et du BRAR avec $m_h = n_h - 1$ (voir Rao et Wu 1984) dans le cas de l'enquête trimestrielle sur l'activité économique en Australie de mars 2000. La taille de l'échantillon de cette enquête au niveau de la strate varie de quatre à plusieurs centaines. Les résultats (tirés de Preston et Chipperfield 2002, tableau 1) montrent qu'au niveau national, $\text{Var}_s(E_*[\hat{v}_{\text{boot}} \mid s])$ est 54 % plus faible pour l'échantillonnage BRSR que pour l'échantillonnage BRAR quand $R = 100$ (Voir Preston et Chipperfield 2002, pour plus d'estimations empiriques de $\text{Var}_s(E_*[\hat{v}_{\text{boot}} \mid s])$ pour le BRAR et le BRSR). Autrement dit, le BRSR nécessite environ deux fois moins de répliques que le BRAR pour atteindre la même erreur de rééchantillonange, ce qui représente un gain d'efficacité important. Un autre avantage du BRSR par rapport au BRAR est que le temps machine pour la sélection des échantillons répétés est considérablement plus court.

Dans le cas des études empiriques, le choix de $m_h = [n_h / 2]$ pour le BRSR minimise $\text{Var}_s(E_*[\hat{v}_{\text{boot}} \mid s])$. Nous soupçonnons qu'à mesure que n augmente, l'écart entre le BRSR et le BRAR tendra à devenir nul. D'autres travaux sont nécessaires en vue d'établir ces propriétés.

5. Variance du changement entre des estimations à une seule phase

Un important produit attendu de nombreuses enquêtes-entreprises est l'estimation du changement entre deux points dans le temps. Représentons la population finie au temps t par $U^{(t)} = \{U_1^{(t)}, U_2^{(t)}, ..., U_H^{(t)}\}$, où $U_h^{(t)}$ est la population de la strate h au temps t qui est constituée de $N_h^{(t)}$ unités. Le total de population au temps t est $Y^{(t)} = \sum_h \sum_{i \in U_h^t} y_{hi}^{(t)}$. La présente section porte sur l'estimation de la variance de $\Delta^{(t)} = \hat{Y}^{(t)} - \hat{Y}^{(t-1)}$, c'est-à-dire la différence entre deux périodes. Les termes correspondant à n_h, f_h et s_h^2 au temps t sont dénotés par $n_h^{(t)}$, $f_h^{(t)}$ et $s_h^{(t)2}$, respectivement. En cas d'échantillonnage à deux occasions, définissons N_c, n_{hc}, $n_{\overline{c}h}^{(1)}$, et $n_{\overline{c}h}^{(2)}$ comme étant le nombre d'unités dans les ensembles suivants : $U_h^{(1)} \cap U_h^{(2)}$, $s_h^{(c)} = s_h^{(1)} \cap s_h^{(2)}$, $s_{h\overline{c}}^{(1)} = s_h^{(1)} - s_h^{(c)}$, et $s_{h\overline{c}}^{(2)} = s_h^{(2)} - s_h^{(c)}$, respectivement. Dans les enquêtes-entreprises de l'ABS, l'échantillon au temps 1 de taille $n_h^{(1)}$ est un EASSR tiré de $U_h^{(1)}$. L'échantillon au temps 2 est l'union des deux échantillons suivants : un EASSR de n_{hc} unités tirées de $s_h^{(c)}$ et un EASSR de $n_{\overline{c}h}^{(2)}$ unités tirées de $U_h^{(2)} - (U_h^{(1)} \cap U_h^{(2)})$. L'échantillon au temps 2 est, en fait, un EASSR tiré de $U_h^{(2)}$. À l'ABS, la taille de l'échantillon chevauchant, n_{hc}, est contrôlée par la méthode du nombre aléatoire permanent (voir Brewer, Gross et Lee 1999).

L'estimateur de $\mathrm{Var}(\hat{\Delta})$ peut être exprimé par

$$\mathrm{Var}(\hat{\Delta}) = \mathrm{Var}(\hat{Y}^{(1)}) + \mathrm{Var}(\hat{Y}^{(2)}) - 2\mathrm{Cov}(\hat{Y}_1^{(1)}, \hat{Y}_2^{(2)}).$$

Considérons l'estimateur d'Horvitz-Thompson $\hat{\Delta} = \hat{Y}^{(2)} - \hat{Y}^{(1)}$, où $t = 1, 2$ et $\hat{Y}^t$ est défini de manière analogue à $\hat{Y}$. Tam (1985) montre que, si $U_h^{(1)} = U_h^{(2)}$, un estimateur sans biais de $\mathrm{Var}(\hat{\Delta})$ sous le plan d'échantillonnage susmentionné est donné par

$$\widehat{\mathrm{Var}}(\hat{\Delta}) = \widehat{\mathrm{Var}}(\hat{Y}^{(1)}) + \widehat{\mathrm{Var}}(\hat{Y}^{(2)}) - 2\widehat{\mathrm{Cov}}(\hat{Y}^{(1)}, \hat{Y}^{(2)}),$$

où

$$\widehat{\mathrm{Var}}(\hat{Y}^{(t)}) = \sum_h N_h^2 (1 - f_t) s_h^{(t)2} / n_h^{(t)},$$

$$\widehat{\mathrm{Cov}}(Y^{(1)}, Y^{(2)}) = \sum_h N_h^2 (1 - f_{12,h})\, s_h^{(12)} n_{hc} / (n_h^{(1)} n_h^{(2)}),$$

$$s_h^{(12)} = (n_{hc} - 1)^{-1} \sum_{i \in s_h^{(c)}} (y_{1i} - \hat{\bar{y}}_1)(y_{2i} - \hat{\bar{y}}_2),$$

$$\hat{\bar{y}}_t = n_{hc}^{-1} \sum_{i \in s_h^{(c)}} y_{ti}$$

pour $t = 1, 2$ et $f_{12,h} = n_h^{(1)} n_h^{(2)} / n_{hc} N_h$.

Si $U_h^{(1)} \neq U_h^{(2)}$, une forme plus générale de l'estimateur de Tam est donnée par $\widehat{\mathrm{Var}}(\hat{\Delta})$, excepté que

$$\widehat{\mathrm{Var}}(\hat{Y}^{(t)}) = \sum_h N_h^{(t)2} (1 - f_t) s_h^{(t)2} / n_h^{(t)},$$

$$\widehat{\mathrm{Cov}}(\hat{Y}^{(1)}, \hat{Y}^{(1)}) = \sum_h N_h^{(1)} N_h^{(2)} / (n_h^{(1)} n_h^{(2)})\, n_c (1 - f_{12,h}) s_h^{(12)}$$

et

$$f_{12,h} = \frac{n_h^{(1)} n_h^{(2)} N_{hc}}{n_{hc} N_h^{(1)} N_h^{(2)}}.$$

Dans la suite de la présente section, nous supposons que $\widehat{\mathrm{Var}}(\hat{\Delta})$ est sans biais pour $\mathrm{Var}(\hat{\Delta})$ quand $U_h^{(1)} \neq U_h^{(2)}$. (Il convient de souligner que $\widehat{\mathrm{Vâr}}(\hat{\Delta})$ peut prendre des valeurs négatives quand $U_h^{(1)} \neq U_h^{(2)}$. Nordberg (2000) donne un estimateur sans biais de $\mathrm{Var}(\hat{\Delta})$ pour l'estimateur par la régression quand $U_h^{(1)} \neq U_h^{(2)}$, mais il n'existe aucun moyen évident de l'utiliser avec la méthode bootstrap décrite dans le présent article.)

L'estimation de la variance de $\hat{\Delta}_{\mathrm{reg}} = \hat{Y}_{\mathrm{reg}}^{(1)} - \hat{Y}_{\mathrm{reg}}^{(2)}$, le changement entre les estimations GREG aux temps 1 et 2, en utilisant le BRSR requiert de répéter R fois les étapes suivantes :

a) former l'ensemble s^* en sélectionnant indépendamment $m_{ch} = [n_{ch}/2]$, $m_{\overline{c}h}^{(1)} = [n_{\overline{c}h}^{(1)}/2]$ et $m_{\overline{c}h}^{(2)} = [n_{\overline{c}h}^{(2)}/2]$ unités par EASSR à partir des ensembles s_{hc}, $s_{h\overline{c}}^{(1)}$, et $s_{h\overline{c}}^{(2)}$, respectivement;

b) pour $i \in s_h^{(1)}$, calculer les poids de rééchantillonnage

$$w_{hi}^{*(1)} = N / n_h^{(1)} \left[1 - \gamma_{\mathrm{ch}} \frac{n_{ch}}{n_h^{(1)}} - \gamma_{1\overline{c}\mathrm{h}} \frac{n_{h\overline{c}}^{(1)}}{n_h^{(1)}} + \gamma_{\mathrm{ch}} \frac{n_{ch}}{m_{ch}} \delta_{hi}^{r(t)} \right]$$

pour $i \in s_{hc}$,

$$w_{hi}^{*(1)} = \left[1 - \gamma_{\mathrm{ch}} \frac{n_{ch}}{n_{1h}} - \gamma_{1\overline{c}\mathrm{h}} \frac{n_{1\overline{c}h}}{n_{1h}} + \gamma_{1\overline{c}\mathrm{h}} \frac{n_{1\overline{c}h}}{m_{1\overline{c}h}} \delta_{hi}^{r(1)} \right]$$

pour $i \in s_{h\overline{c}}^{(1)}$, où

$$\gamma_{1\overline{c}\mathrm{h}} = \sqrt{\frac{[n_{1h}(1 - f_h) - n_{ch}(1 - f_{12,h})]\, m_{1\overline{c}h}}{\{n_{1\overline{c}h}(n_{1\overline{c}h} - m_{1\overline{c}h})\}}},$$

$$\gamma_{\mathrm{ch}} = \sqrt{(1 - f_{12,h})\, m_{ch} / (n_{ch} - m_{ch})}$$

et $\delta_{hi}^{*(t)}$ égale 1 si l'unité i est sélectionnée dans le groupe échantillonné au temps t et 0 autrement;

c) calculer les poids définis de façon analogue pour $i \in s_h^{(2)}$;

d) calculer $\tilde{w}_{hi}^{*(t)} = w_{hi}^{*(t)} g_i^{(t)*}$ pour $i \in s_h^{(1)}, s_h^{(2)}$, où $g_i^{(t)*}$ a la même forme que g_i, mais est calculé en utilisant les poids $w_{hi}^{*(t)}$ au lieu de $w_{hi}^{(t)}$;

e) calculer $\hat{\Delta}^*_{\text{reg}} = \hat{Y}^{*(2)}_{\text{reg}} - \hat{Y}^{*(1)}_{\text{reg}}$, où $\hat{Y}^{*(t)}_{\text{reg}} = \sum_{i \in s^{(t)}} \tilde{w}^{*(t)}_{hi} y_i$. L'estimateur BRSR de la variance est donné par

$$\widehat{\text{Var}}_B(\hat{\Delta}_{\text{reg}}) = (R-1)^{-1} \sum_{r=1}^{R} (\hat{\Delta}^*_{\text{reg}} - \hat{\Delta}_{\text{reg}})^2,$$

où

$$\hat{\Delta}_{\text{reg}} = \hat{Y}^{(1)}_{\text{reg}} - \hat{Y}^{(2)}_{\text{reg}}$$

et

$$\hat{Y}^{(t)}_{\text{reg}} = \sum_{i \in s^{(t)}} \tilde{w}^{(t)}_{hi} y_i.$$

La preuve que $\widehat{\text{Var}}_B(\hat{\Delta}_{\text{reg}})$ est sans biais est simple et semblable à la preuve que $\text{Var}_B(\hat{\theta})$ est sans biais (voir la section 4).

L'approche décrite ci-dessus requiert un ensemble distinct de poids de rééchantillonnage pour les estimations de la variance des changements et des niveaux. Roberts, Kovačević, Mantel et Phillips (2001) considèrent des estimateurs bootstrap approximatifs de la variance des changements utilisant uniquement les poids de rééchantillonnage pour les niveaux, ce qui réduit les coûts de calcul et simplifie la méthode et sa mise en œuvre dans un système informatique.

6. Étude par simulation

À la présente section, nous résumons une étude par simulation réalisée pour les estimations ponctuelles et les estimations des changements afin de mesurer empiriquement le biais et la variabilité du BRSR et du BRAR sous un échantillonnage répété quand $R = 100$. Nous avons généré une population aux points 1 et 2 dans le temps à partir des modèles suivants, $y^{(1)}_i = (0,75x_{1i} + 0,25x_{2i})\, W(0;\ 2,5;\ 1)$ et $y^{(2)}_i = 1,5y^{(1)}_i\ W(0;\ 5;\ 1)$, où les variables auxiliaires sont données par $x_{1i} = 0,25x_{2i} + 0,75\,[100L(0;\ 1;\ 1)]$ et $x_{1i} = 100L(0;\ 1;\ 1)$ où $W(\mu;\ \gamma;\ \alpha)$ et $L(\mu;\ \gamma;\ \alpha)$ sont les lois de Weibull et log-normale dont les paramètres d'emplacement, de forme et de dispersion sont donnés par μ, γ et α. Ces lois reflètent les longues queues des distributions typiques des données d'enquête économique. Les tailles de population aux temps 1 et 2 étaient de 3 000, avec 2 500 unités communes aux deux points dans le temps. Chaque unité de population, i, a été assignée à l'une des cinq strates aux deux points dans le temps en utilisant z_i, où $z_i = x_{1i}\ W(0;\ 2,5;\ 1)$ et les bornes des strates étaient $z_i =$ 50, 100, 150, 250. Nous avons ainsi obtenu des tailles de population de strate variant de 400 à 1 000.

Nous avons tiré, en tout, 3 000 EASSR stratifiés simulés à partir de la population aux temps 1 et 2, où $n^{(t)}_h = 12$, $n^{(1)}_{\bar{c}h} = n^{(2)}_{\bar{c}h} = 4$ et $n_{ch} = 8$ pour tout h et $t = 1, 2$. Pour le BRSR, les tailles des échantillons répétés sont données aux sections 4 et 5. Pour le BRAR, les tailles des échantillons répétés étaient $m_{ch} = [n_{ch} - 1]$, $m^{(1)}_{\bar{c}h} = [n^{(1)}_{\bar{c}h} - 1]$ et $m^{(2)}_{\bar{c}h} = [n^{(2)}_{\bar{c}h} - 1]$ pour les changements, et $m^{(t)}_h = n^{(t)}_h - 1$ pour les niveaux. Pour les changements, l'estimateur BRAR était de la même forme que l'estimateur BRSR, mais avait un facteur d'échelle légèrement différent et tirait les échantillons répétés avec remise.

À partir de chacun des 3 000 échantillons simulés, nous avons calculé $\hat{Y}^j_{\text{reg}}$, où $\hat{Y}^j_{\text{reg}}$ est donné par $\hat{Y}_{\text{reg}}$ avec $\boldsymbol{x}_i = (x_{1i},\ x_{2i})$, $\sigma_i = 1$ et $j = 1, 2, \ldots, 3\,000$. L'erreur-type réelle de $\hat{Y}_{\text{reg}}$ est calculée par

$$S = \sqrt{\frac{1}{3\,000} \sum_{j=1}^{3\,000} (\hat{Y}^j_{\text{reg}} - Y)^2}.$$

L'erreur-type de $\hat{Y}_{\text{reg}}$ estimée par le bootstrap à partir du j^{e} échantillon est

$$\hat{S}^j = \sqrt{\frac{1}{100} \sum_{r=1}^{100} (\hat{Y}^{j*(r)}_{\text{reg}} - Y)^2},$$

où $\hat{Y}^{j*(r)}_{\text{reg}}$ est défini de façon analogue à $\hat{Y}^{*(r)}_{\text{reg}}$. Le biais relatif (BR) de l'erreur-type estimée par le bootstrap est

$$\text{BR}(\hat{S}) = \frac{1}{3\,000S} \sum_{j=1}^{3\,000} (\hat{S}^j - S).$$

La racine carrée relative de l'erreur quadratique moyenne (RREQM) de l'erreur-type estimée par le bootstrap est

$$\text{RREQM}(\hat{S}) = \frac{1}{S} \sqrt{\frac{1}{3\,000} \sum_{j=1}^{3\,000} (\hat{S}^j - S)^2}.$$

Des définitions semblables de la RREQM et du biais sont utilisées pour l'estimation de la variance des changements. Nous comparons aussi les probabilités de couverture à 95 %, c'est-à-dire le pourcentage des intervalles de confiance à 95 % contenant le total de population réel du BRSR et du BRAR pour les niveaux et les changements.

Les résultats présentés dans le tableau 1 montrent que le BR et la RREQM du BRSR et du BRAR sont tous deux acceptablement faibles. Le biais des estimations ponctuelles au temps 1 produit par le BRAR est légèrement plus grand que celui obtenu pour le BRSR, ce qui se traduit par des probabilités de couverture un peu moins bonnes.

Tableau 1
Estimation bootstrap de l'erreur-type des changements et des estimations ponctuelles

	Temps 1			Changement		
Méthode	BR	RREQM	C 95 %	BR	RREQM	C 95 %
BRSR	0,7	17,3	94,7	2,1	20,7	95,3
BRAR	-3,1	15,8	93,7	-1,3	19,6	94,6

7. Sommaire

D'après les résultats des simulations, nous avons considéré le BRSR et le BRAR comme étant tous deux fiablement exacts sous échantillonnage répété. Conditionnellement à l'échantillon, le BRSR s'est avéré significativement plus efficace (jusqu'à 50 %) que le BRAR pour l'échantillonnage stratifié quand il arrive que la taille d'échantillon de strate soit faible. Par conséquent, le BRSR est la méthode qui a été implémentée dans ABSEST.

Bibliographie

Brewer, K.R.W., Gross, W.F. et Lee, G.F. (1999). PRN sampling: The Australian experience. *Proceedings of the International Associations of Survey Statisticians*, Helsinki.

Estevao, V., Hidiroglou, M.A. et Särndal, C.-E. (1995). Methodological principles for a generalised estimation system at Statistics Canada, *Journal of Official Statistics*, 11, 2, 181-204.

Kovar, J., Rao, J.N.K. et Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 25-44.

Nordberg, L. (2000). On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16, 4, 363-378.

Preston, J., et Chipperfield, J.O. (2002). Using a generalised estimation methodology for ABS business surveys. Methodology Advisory Committee (disponible au www.abs.gov.au).

Rao, J.N.K., et Wu, C.F.J. (1984). Bootstrap inference for sample surveys. *Proceeding Section on Survey Methods Research*, *Journal of the American Statistical Association*, 106-112.

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, mars, 83, 401, 231-241.

Roberts, G., Kovačević, M., Mantel, H. et Phillips, O. (2001). Cross sectional inference based on longitudinal surveys: Some experience with Statistics Canada Surveys. Présenté au 2001 FCSM Research Conference, Arlington, VA, 14-16 novembre, Recueil disponible au www.fcsm.gov.

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

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

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

Tam, S.M. (1985). On covariance in finite population sampling. *The Statistician*, 34, 429-433.

Yung, W., et Rao, J.N.K. (1996). Linéarisation des estimateurs de variance jackknife dans un échantillonnage stratifié à degrés multiples. *Techniques d'enquête*, 22, 23-31.

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2007
Vol. 33, N° 2, pp. 195-207
Statistique Canada, N° 12-001-X au catalogue

Estimation bayésienne pour de petites régions en cas de différence entre les strates du plan d'échantillonnage et les domaines d'étude

Jacob J. Oleson, Chong Z. He, Dongchu Sun et Steven L. Sheriff [1]

Résumé

Il s'agit d'obtenir des estimations fiables pour des domaines d'étude où les tailles d'échantillon peuvent être des plus modestes et pour lesquels la strate du plan d'échantillonnage ne coïncide pas avec le domaine. On ignore les tailles de population autant pour le domaine d'étude que pour la strate du plan d'échantillonnage. Dans le calcul des estimations paramétriques des domaines d'étude, le choix d'une taille d'échantillon aléatoire s'impose souvent. Nous proposons une nouvelle famille de modèles mixtes linéaires généralisés (MMLG) à effets aléatoires corrélés lorsqu'il y a plus d'un paramètre inconnu. Le modèle que nous proposons estimera tant la taille de population que le paramètre d'intérêt. Pour ce cadre, nous donnons des formules générales pour les distributions conditionnelles intégrales qu'exigent des simulations de Monte Carlo à chaîne de Markov (MCCM). Nous présentons aussi des équations de prévision et d'estimation bayésiennes pour les domaines d'étude. Nous nous servons enfin de l'enquête de 1998 sur la chasse aux dindons dans le Missouri, laquelle stratifie des échantillons en fonction du lieu de résidence du chasseur, et nous voulons obtenir des estimations au niveau du domaine, c'est-à-dire du comté où le chasseur de dindons s'adonne effectivement à cette activité.

Mots clés : Méthode hiérarchisée bayésienne; méthode de Monte Carlo à chaîne de Markov; prévision bayésienne; taille d'échantillon aléatoire; corrélation spatiale; stratification.

1. Introduction

On a souvent affaire à de petites tailles d'échantillon lorsqu'on analyse des données d'enquête. C'est que, dans bien des cas, on se trouve à étudier des sous-populations comme des groupes sociodémographiques. Il se peut aussi que nous considérions des régions dans l'espace ou des périodes dans le temps comme sous-populations ou domaines d'étude. En raison de la faible taille d'échantillon, les estimateurs directs d'enquête pourraient se révéler très peu fiables. On a qualifié d'estimations régionales (ER) les estimations relatives à des régions à petite taille d'échantillon. Rao (2003) passe efficacement en revue un grand nombre de techniques ER. On peut aussi trouver un certain nombre d'études récentes sur cette estimation de petites régions dans Rao (2005) et Jiang et Lahiri (2006).

On doit disposer de bons modèles pour produire des statistiques fiables sur de petites régions. En modélisation, il existe différentes méthodes, dont celle des meilleures prévisions empiriques (voir Prasad et Rao 1990; Jiang, Lahiri et Wan 2002; Das, Jiang et Rao 2004; Jiang et Lahiri 2006) et les méthodes bayésiennes (voir Malec, Sedransk, Moriarity et LeClere 1997; Ghosh, Natarajan, Stroud et Carlin 1998; He et Sun 2000). Pour un bon examen des estimations régionales bayésiennes, nous renvoyons le lecteur à Rao (2003), qui traite d'une application bien concrète du cadre méthodologique de Bayes. Une étape primordiale dans l'application d'un modèle bayésien est le choix de distributions a priori. Il faut en outre s'attacher avec soin à la qualité de la distribution a postériori et à la robustesse des distributions a priori. Là où une simulation MCCM comme la formule d'échantillonnage de Gibbs entre dans les calculs, on doit contrôler la convergence de la chaîne de Gibbs. Pour plus de détails, voir Carlin et Louis (2000).

Regardons un plan d'échantillonnage à grappes stratifiées où chaque grappe est choisie par échantillonnage aléatoire simple sans remise. Dans notre application, les grappes sont d'une taille inégale et les valeurs de taille sont inconnues au moment de concevoir l'enquête. Considérons maintenant un problème d'estimation de domaines où ceux-ci débordent les limites des grappes nominales. On ignore donc la taille de population des domaines et la taille d'échantillon est aléatoire. Dans notre application, les tailles d'échantillon réalisées pour les domaines sont modestes et, en soi, les techniques d'estimation types en fonction du plan d'échantillonnage (voir Cochran 1977, et Lohr 1999) sont peu sûres. Nous proposerons comme moyen de contourner ce problème un modèle hiérarchisé intégral de Bayes.

Commençons par obtenir des estimations de taux de succès avec les tailles de population dans de petites régions pour les personnes composant chacune des strates nominales. On obtient ces estimations en tirant de l'information des petites régions voisines, et ce, par une structure spatiale imbriquée dans le modèle bayésien. Ainsi, les estimations obtenues sont bien plus stables que celles de méthodes directes d'enquête. Nous établissons ensuite une moyenne pondérée des taux de succès des strates nominales pour les estimations finales de petites régions. Si un comté constitue

1. Jacob J. Oleson, Département de biostatistique C22-GH, Université de l'Iowa, Iowa City, Iowa, 52242-1009, États-Unis; Chong Z. He et Dongchu Sun, Département de statistique, 146, Middlebush Hall, Université du Missouri à Columbia, Missouri, 65211-6100, États-Unis; Steven L. Sheriff, Resource Science Center, Département de la conservation du Missouri, 1110, avenue South College, Columbia, Missouri, 65201.

une de ces petites régions, nous calculons un taux de succès par comté pour chacune des strates nominales et en prenons la moyenne pour une estimation unique de comté. Pour pouvoir combiner les strates du plan d'échantillonnage, nous devrions connaître les tailles individuelles d'effectif des strates. Si ces valeurs sont inconnues, leur estimation est possible par le modèle que nous proposons. Notre propos est donc d'appliquer les méthodes bayésiennes à l'estimation des taux de succès de la chasse aux dindons au niveau des comtés du Missouri. Nous proposons une nouvelle famille de modèles mixtes linéaires généralisés (MMLG) à effets aléatoires corrélés lorsque plusieurs paramètres sont inconnus. Nous appellerons notre modèle MMLG à deux variables. On se sert souvent d'un MMLG à effets aléatoires possibles corrélés lorsqu'il n'y a qu'un seul paramètre inconnu (Sun, Speckman et Tsutakawa 2000). Dans le modèle que nous proposons, nous estimons à la fois la taille de population et le paramètre d'intérêt et, de ce fait, nous faisons progresser la recherche sur les estimations régionales.

Le scénario que nous venons de décrire se présente dans l'enquête de 1998 sur la chasse aux dindons au Missouri (« Missouri Turkey Hunting Survey » ou MTHS). Il s'agit d'une enquête postale après la saison printanière qui a renseigné les services de conservation du Missouri sur le nombre de dindons tués par les chasseurs de jour en jour en saison de chasse et sur les comtés où cette récolte a eu lieu. On a aussi relevé dans cette enquête le nombre total de visites de comtés par les chasseurs à chaque journée de chasse. On a ensuite calculé les taux de succès de la chasse à l'aide de ces données.

Le cas MTHS est présenté en détail à la section 2. À la section 3, on trouvera un résumé du cadre méthodologique proposé et les formules générales applicables. Le but était d'estimer les taux de succès, mais ces méthodes sont généralisables. Nous offrons aussi des formules générales d'estimation et de prévision pour de petites régions, ainsi que les distributions conditionnelles intégrales pour des simulations MCCM. À la section 4 enfin, nous livrons nos derniers commentaires.

2. Enquête de 1998 sur la chasse aux dindons au Missouri

2.1 Contexte de la MTHS de 1998

Les services de conservation du Missouri ont commencé en 1986 relever à intervalles biennaux les tendances de la chasse aux dindons à l'aide de la MTHS. Dans cette enquête, on a demandé aux chasseurs dans quels comtés ils chassaient, quels étaient leurs jours de chasse et si celle-ci donnait des résultats ou non. Le tout a commencé par un échantillonnage aléatoire simple de tous les titulaires de permis de chasse printanière aux dindons. He et Sun (1998) se sont reportés aux données de l'enquête de 1996 pour estimer les taux de réussite de cette chasse dans les 114 comtés du Missouri en se servant d'un modèle bêta-binomial de Bayes. Dans He et Sun (2000), on estime les taux de succès par comté et par semaine de saison de chasse. Pendant la saison printanière de 1996, on n'avait la permission de récolter qu'un dindon par semaine. Ces auteurs ont utilisé un MMLG et estimé les seuls taux de succès. Oleson et He (2004) ont étendu l'application de ce modèle à l'estimation des taux pour chaque jour de la saison de chasse. Ils ont pu voir un taux significatif d'autocorrélation entre les jours de la saison et entre les comtés du Missouri lorsqu'ils ont estimé ces taux de succès.

En 1998, on a modifié le plan d'échantillonnage MTHS. La base de sondage demeurait la liste de tous les chasseurs de dindons ayant la permission de chasser au Missouri. On y trouve entre autres des indications sur le comté de résidence de chaque chasseur. Dans les échantillons aléatoires simples exploités par le passé, les régions populeuses de Kansas City et de St. Louis étaient d'un trop grand poids. C'est pourquoi les comtés proches de ces régions métropolitaines se sont vu attribuer un échantillon important et les comtés plus éloignés (plus au sud, par exemple) ont eu droit à un échantillon insuffisant. Idéalement, nous aimerions stratifier selon les comtés où les chasseurs se sont livrés à cette activité afin de pouvoir constituer des échantillons représentatifs à ce niveau des domaines d'étude. Il est exclu de savoir où les intéressés ont chassé avant leur renvoi des questionnaires, d'où l'impossibilité d'une telle stratification.

Une autre possibilité est de stratifier selon les comtés de résidence des chasseurs, puisque ceux-ci ont tendance à chasser près de l'endroit où ils demeurent (c'est-à-dire en des lieux qui leur sont familiers). Cela pose un problème d'estimation des paramètres d'intérêt, c'est-à-dire des taux de succès de la chasse par comté. Nous voudrions des estimations selon les lieux de chasse, mais le plan d'échantillonnage fait intervenir les lieux de résidence des chasseurs. Le nouveau plan de sondage de la MTHS prévoit un échantillonnage aléatoire simple à grappes stratifiées de taille inégale. Dans ce cas, une grappe représente un chasseur titulaire de permis et ses éléments sont les expéditions de chasse de chaque chasseur. Le lieu de résidence du chasseur sert de facteur de stratification. Voici les strates nominales en question : 1) non-résidents du Missouri; 2) résidents du nord du Missouri; 3) résidents du sud du Missouri; 4) résidents de la région métropolitaine de St. Louis; 5) résidents de la région métropolitaine de Kansas City. La figure 1 indique la délimitation des quatre strates de résidents du Missouri en fonction des trois premiers chiffres du code postal américain. On a procédé par répartition proportionnelle pour établir le nombre de chasseurs échantillonnés

dans chaque strate. Comme on peut le voir au tableau 1, on comptait 110 691 titulaires de permis au total. Nous avons réparti un échantillon de 8 000 chasseurs au prorata entre les cinq strates. Au tableau 1, la colonne « % de l'échantillon » correspond à la partie de tout l'échantillon qui vient d'une strate (échantillon / échantillon total). Quant à la colonne « % échantillonné de la strate », elle correspond à la proportion de chasseurs échantillonnés dans une strate (échantillon / nombre total de permis). Le nombre de grappes dans la population et dans l'échantillon pour chaque strate est une valeur fixe connue dans les deux cas. Les expéditions de chasse de chaque chasseur sont les éléments d'une grappe. Dans la suite de cet exposé, la taille de population est le nombre d'expéditions de tous les chasseurs de la base de sondage, ce que nous appellerons la « pression de chasse ». Dans le cas de la MTHS, la taille de population est inconnue, mais n'a rien d'aléatoire. La taille d'échantillon est considérée comme aléatoire au mieux, car nous ignorons combien d'expéditions fera chaque chasseur. À noter qu'un chasseur n'a pas à chasser dans sa strate du plan d'échantillonnage. Il peut chasser dans bien des comtés au cours de la saison. Pour chaque domaine d'étude (comté), la taille de la population est certes inconnue, mais non aléatoire. La taille d'échantillon est aléatoire, répétons-le, dans la mesure où nous ignorons dans quel comté un chasseur va se livrer à cette activité. De petites tailles d'échantillon sont toujours possibles. Ainsi, les comtés de Dunklin, Pemiscot et New Madrid dans le sud-est du Missouri n'ont aucune expédition déclarée pour la saison de trois semaines en 1998 et le comté de Lawrence dans le sud-ouest n'en a pas pour la première journée et cinq seulement pour le reste de la première semaine.

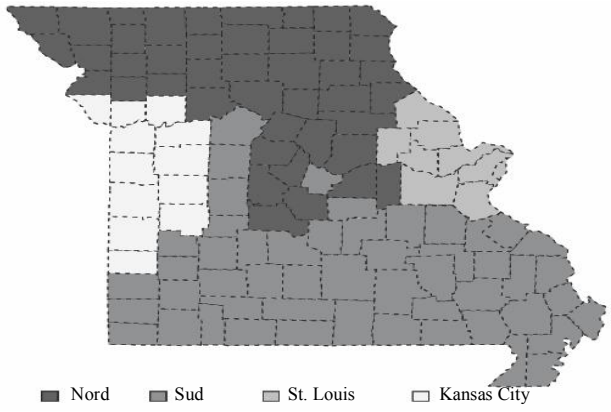


Figure 1 Délimitations du plan d'échantillonnage pour les quatre domaines de chasse des résidents du Missouri

La MTHS est une enquête postale après la saison printanière où les questionnaires sont envoyés à la résidence inscrite du chasseur. On y pose des questions pour chaque jour de la saison de chasse et on demande aux chasseurs s'ils se sont ou non livrés à cette activité, dans quel comté ils ont chassé et si leur chasse a eu du succès ou non. Au total, 41 % des résidents du Missouri ont répondu au premier questionnaire expédié par la poste. Ceux qui ne l'avaient pas fait dans les deux mois ont reçu à nouveau le questionnaire par la poste et, dans ce cas, le taux de réponse s'est établi à 26 %. Le questionnaire a été expédié une troisième fois à ceux qui n'avaient toujours pas répondu au terme des deux premiers mois et, dans ce dernier cas, le taux de réponse s'est établi à 20 %. Dans l'ensemble, le taux de réponse est de 66,5 %. On peut voir au tableau 1 que les non-résidents du Missouri présentaient le taux de réponse le plus élevé avec une valeur de près de 75 %. Suivaient les résidents des deux régions métropolitaines de St. Louis et de Kansas City avec des taux respectifs de 69 % et 66 %. Les taux étaient les plus faibles (63 % environ) pour les régions rurales du nord et du sud de cet État.

Tableau 1 Tailles d'échantillon et taux de réponse pour les 5 strates de la MTHS de 1998

Strate nominale	Nombre total de permis	Échantillon d'envoi	Échantillon de réponse	Taux de réponse (%)	Pourcentage de l'échantillon	Pourcentage échantillonné de la strate
Non-résidents	19 798	1 600	1 180	73,75	22,2	5,96
Nord du Missouri	21 756	1 532	975	63,64	18,3	4,48
Sud du Missouri	34 375	2 421	1 509	62,33	28,4	4,39
St. Louis	19 959	1 405	969	68,97	18,2	4,85
Kansas City	14 803	1 042	688	66,03	12,9	4,65
Total	110 691	8 000	5 321	66,51	100	4,81

En 1998, la saison printanière de chasse aux dindons au Missouri a été de 21 jours consécutifs à compter du lundi 19 avril. La première semaine, le chasseur pouvait récolter un dindon « à barbe ». S'il avait du succès cette première semaine, il avait la permission de récolter un autre dindon « à barbe » les deux dernières semaines. En cas d'insuccès la première semaine, il pouvait récolter deux dindons « à barbe » les deux dernières semaines, mais à raison d'un seul volatile par jour. Durant trois semaines, les biologistes de l'espèce des dindons voulaient être renseignés sur le succès des chasseurs le premier jour en fonction de l'effet de la journée d'ouverture où la « pression de la chasse » est normalement supérieure à celle de tout autre jour de cette même saison. Les six jours restants de la première semaine forment une deuxième période d'intérêt appelée semaine 1. Les semaines 2 et 3 sont modélisées chacune, ce qui donne un total de quatre périodes d'intérêt.

La petite région qui nous intéresse est la combinaison comté-période pour les chasseurs de dindons de chaque strate. Dans la gestion de la population de dindons, on privilégie le comté comme moindre unité géographique

d'intérêt, parce que les chasseurs sont capables d'indiquer où ils ont chassé lorsqu'ils rendent compte du succès de leur chasse. Les intéressés ont besoin d'un permis de chasse et, par conséquent, nous savons combien de chasseurs compte le Missouri au total et à quelles strates ceux-ci appartiennent. Comme les données sont étalées sur les 114 comtés de cet État, on se retrouve avec une information fort clairsemée pour chaque comté. Plutôt que de regarder les taux de succès de jour en jour, nous le faisons pour quatre périodes que nous avons délimitées. Dans le calcul des taux de succès par comté, nous nous reportons cependant au nombre d'expéditions de chasse dans chaque comté. Nous ignorons combien d'expéditions fera chaque chasseur ou dans quel comté il les fera. Nous connaissons donc le nombre de chasseurs dans chaque strate, mais le nombre d'expéditions est inconnu. De plus, nous connaissons le nombre de chasseurs échantillonnés dans chaque strate, mais ignorons le nombre d'expéditions dans un comté en particulier; cette valeur est aléatoire, puisque le nombre d'expéditions variera selon l'échantillon de chasseurs. Il y a deux paramètres principaux à estimer et une variable aléatoire à prévoir. Le premier paramètre correspond au nombre total d'expéditions de tous les chasseurs, soit la « pression de chasse ». Cet aspect importe aux gestionnaires de la faune qui se préoccupent de la qualité de l'expérience de chasse. Trop de chasseurs dans une région ont tendance à se mêler de la chasse d'autrui, d'où une perte de qualité de cette activité. L'autre paramètre à estimer correspond au taux de succès de la chasse ou à la proportion de dindons récoltés par expédition. S'il n'y a pas assez de dindons dans un comté, les gestionnaires de la faune fermeront ce dernier à la chasse un certain nombre d'années pour que la population de dindons se rétablisse. La variable aléatoire à prévoir est celle du nombre total de dindons récoltés. Au Missouri, on a l'obligation de déclarer tout dindon récolté. En 1998, les chasseurs étaient tenus de se rendre à un poste de contrôle où on comptait les dindons récoltés. On a jugé coûteux de maintenir ces postes. Le projet MTHS visait à prévoir le nombre total de dindons récoltés au niveau des comtés et à comparer cette valeur à la valeur de dénombrement des dindons récoltés dans les postes de contrôle.

Nous allons présenter un modèle d'estimation du taux de succès et de la pression de la chasse et prévoir simultanément la récolte totale de dindons. Le modèle tient compte de tailles d'échantillon aléatoires absentes des modèles de He et Sun (1998), He et Sun (2000) et Oleson et He (2004).

2.2 Modèle

Soit p_{ijk} désignant le taux de succès, y_{ijk} le nombre de succès, n_{ijk} le nombre d'expéditions dans l'échantillon et N_{ijk} le nombre total d'expéditions dans le comté i, au moment j, pour les chasseurs de la strate k. Nous modélisons la récolte de dindons à l'aide de distributions binomiales indépendantes

$$(y_{ijk} \mid n_{ijk},\ p_{ijk}) \overset{\text{ind}}{\sim} \text{Binomial}(n_{ijk},\ p_{ijk}), \qquad (1)$$

où $i = 1, ..., I$ est le comté (domaine d'étude), $j = 1, ..., J$ la période et $k = 1, ..., K$ la strate nominale ou de plan d'échantillonnage. Dans ce cas, $I = 114$, $J = 4$, $K = 5$ pour la MTHS. Dans les analyses a priori de cette enquête, on a posé l'existence d'une taille d'échantillon fixe n_{ijk}, qui, à notre avis, serait considérée au mieux comme aléatoire. Nous ignorons le nombre d'expéditions dans chaque comté tant que les questionnaires n'ont pas été renvoyés. Nous ignorons également le nombre total d'expéditions dans la petite région N_{ijk}, et connaissons seulement le nombre de chasseurs possible. Comme les chasseurs doivent s'abstenir après avoir récolté un second dindon, les comtés aux taux de succès supérieurs auraient moins d'expéditions (ou de jours de chasse) pour le même nombre de volatiles récoltés. S'il y a corrélation, la taille d'échantillon doit être jugée aléatoire. Ajoutons que, dans un modèle hiérarchisé de Bayes, si la distribution d'une taille d'échantillon est indépendante de la distribution de la variable de réponse, les estimations sont identiques pour une taille d'échantillon (fixe) aléatoire ou non, n, (voir Durbin 1969). Le taux de succès estimé est plus lisse pour un n_{ijk} fixe que pour une valeur aléatoire (Woodard, He et Sun 2003). Malec et coll. (1997) ont appliqué les estimations régionales bayésiennes à la National Health Interview Survey (NHIS). Il existe deux grandes différences entre notre modèle et celui de Malec et coll. (1997). D'abord, les tailles de population N_{ijk} sont connues dans le modèle de ces auteurs, mais inconnues dans le nôtre. C'est ce qui explique principalement que nous introduisions un MMLG à deux variables. En second lieu, on modélise dans leur modèle les taux de succès en logits comme fonction linéaire des covariables. Ainsi, les estimations dépendent des valeurs des covariables, mais non de la répartition spatiale. Nous ajouterons l'aspect spatial aux covariables dans le calcul logistique des taux de succès, de sorte que les estimations dépendent à la fois des covariables et de la répartition spatiale. Cela s'avère d'autant plus nécessaire que des covariables importantes peuvent ne pas être disponibles.

Pour tenir compte du caractère aléatoire des tailles d'échantillon, nous modélisons n_{ijk} par distribution de Poisson

$$(n_{ijk} \mid N_{ijk}) \overset{\text{ind}}{\sim} \text{Poisson}(R_k N_{ijk}). \qquad (2)$$

Pour n_{ijk}, la moyenne et la variance de la distribution de Poisson sont une constante multipliée par la taille de population N_{ijk}. Cette constante R_k est le rapport entre le nombre échantillonné et le nombre total de chasseurs de la strate k. On peut la calculer à l'aide du tableau 1.

Pour cette distribution, la taille globale d'échantillon $n_{..k}$ est jugée aléatoire. Nous avons exposé à la section qui précède ce qui milite en faveur d'une telle hypothèse. Si $n_{..k}$ était fixe, la distribution multinomiale irait mieux comme modèle. Les valeurs de vraisemblance de ces deux traitements sont des plus convergentes et donnent des résultats comparables dans l'un et l'autre de ces cadres (voir Agresti 2002, pages 8-9).

Nous modélisons p_{ijk} par sa fonction logistique et N_{ijk} par une transformation logarithmique :

$$\eta_{ijk} = \log\left(\frac{p_{ijk}}{1-p_{ijk}}\right), \qquad \omega_{ijk} = \log(N_{ijk}).$$

Nous nous servons de modèles mixtes linéaires pour les distributions a priori à la fois dans η_{ijk} et ω_{ijk} en posant ce qui suit :

$$\eta_{ijk} = \theta_{1jk} + u_{1ik} + e_{1ijk},$$
$$\omega_{ijk} = \theta_{2jk} + u_{2ik} + e_{2ijk}.$$

Pour $a = 1, 2$, θ_{ajk} désigne l'effet fixe dû au j^{e} moment dans la strate k et u_{aik} représente l'effet aléatoire de comté; les erreurs aléatoires e_{aijk} sont i.i.d. $N(0, \delta_{ak}^{(e)})$.

Pour compléter le modèle hiérarchisé de Bayes, nous devons spécifier les distributions a priori pour $\boldsymbol{\theta}_{ak} = (\theta_{a1k}, ..., \theta_{aJk})'$, $\boldsymbol{u}_{ak} = (u_{a1k}, ..., u_{aIk})'$ et $\delta_{ak}^{(e)}$. He et Sun (2000) et Oleson et He (2004) démontrent l'existence d'un taux significatif de corrélation spatiale entre les comtés du Missouri pour l'estimation des taux de succès. Ces auteurs se servent d'une structure autorégressive conditionnelle (ARC) pour modéliser la dépendance spatiale entre comtés voisins. La codensité de $\mathbf{u}_{ak}$ est donnée par

$$f(\mathbf{u}_{ak}) = (2\pi\delta_{ak}^{(u)})^{-\frac{I}{2}} |\, \boldsymbol{I} - \rho_{ak}\boldsymbol{C}\,|^{\frac{1}{2}} \exp\left\{-\frac{1}{2\delta_{ak}^{(u)}}\, \boldsymbol{u}_{ak}^{t}\,(\boldsymbol{I}-\rho_{ak}\boldsymbol{C})\,\boldsymbol{u}_{ak}\right\}. \quad (3)$$

Dans ce cas, $\delta_{1k}^{(u)}$ et $\delta_{2k}^{(u)} > 0$ sont des composantes de la variance et $\boldsymbol{I} - \rho_{ak}\boldsymbol{C}$ est une matrice symétrique définie à valeurs non négatives (Besag 1974). $\boldsymbol{I}$ est la matrice d'identité $I \times I$ et $\boldsymbol{C}$, une matrice d'adjacence dont la composante $\{c_{ij}\}$ est l'unité si les régions i et j sont en mitoyenneté avec $\{c_{ii}\} = 0$. Nous définissons également ρ_{1k} et ρ_{2k} comme paramètres de corrélation spatiale. Soit $\lambda_1 \le ... \le \lambda_I$ les valeurs caractéristiques de la matrice $\mathbf{C}$. Ces matrices $\boldsymbol{I} - \rho_{ak}\boldsymbol{C}$, sont à valeurs définies positives si $\lambda_1^{-1} \le \rho_{ak} \le \lambda_I^{-1}$ (Clayton et Kaldor 1987). Pour les données du Missouri, $I = 114$ et les valeurs numériques de λ_1 et λ_{114} sont respectivement -2,8931 et 5,6938. Ainsi, il y a densité de $\mathbf{u}_{ak}$ si ρ_{ak} se situe dans l'intervalle (-0,3457, 0,1756).

Pour les autres distributions a priori, nous posons des hypothèses. θ_{ajk} est normal avec une moyenne μ_{ajk} et une variance τ_{ajk}. Soit ρ_{ak} en distribution uniforme sur l'intervalle $(\lambda_1^{-1}, \lambda_I^{-1})$. Enfin, une distribution a priori commune pour les composantes de la variance $\delta_{ak}^{(e)}$ et $\delta_{ak}^{(u)}$ est une distribution gamma inverse (Gelman, Carlin, Stern et Rubin 1995) dont les valeurs de densité sont respectivement proportionnelles à

$$\frac{1}{\delta_{ak}^{(e)\alpha_{ak}^{(e)}+1}} \exp\left(-\frac{\beta_{ak}^{(e)}}{\delta_{ak}^{(e)}}\right) \text{ et } \frac{1}{\delta_{ak}^{(u)\alpha_{ak}^{(u)}+1}} \exp\left(-\frac{\beta_{ak}^{(u)}}{\delta_{ak}^{(u)}}\right),$$

Pour évaluer la distribution a postériori, nous employons des méthodes MCCM comme l'échantillonnage de Gibbs (Gelfand et Smith 1990) et tirons des échantillons de cette distribution. On trouvera un aperçu des méthodes MCCM dans Gelman et coll. (1995), Gilks, Richardson et Spiegelhalter (1996) et Robert et Casella (1999). Les distributions conditionnelles intégrales nécessaires à cette évaluation peuvent être directement déterminées à l'aide des distributions du « Fait 1 » à la section 3. La plupart des $(\boldsymbol{\theta}_{ak}, \mathbf{u}_{ak}, \delta_{ak}^{(e)}, \delta_{ak}^{(u)})$ sont de la forme type et se prêtent d'emblée à l'échantillonnage. Pour d'autres distributions conditionnelles intégrales (η_{ik}, ω_{ik} et ρ_{ak}), on constate des valeurs logconcaves de densité pour la MTHS. On peut se servir de la méthode d'échantillonnage-rejet d'adaptation de Gilks et Wild (1992) pour tirer des échantillons aléatoires de densités logconcaves.

2.3 Estimation et prévision bayésiennes

Nous avons obtenu des estimations de (p_{ijk}, N_{ijk}). Nous désirons grouper les estimations de la strate pour pouvoir estimer p_{ij} et N_{ij}. Nous voulons aussi prévoir le nombre inobservé de dindons récoltés dans le comté i, ce que désigne h_i.

Pour dégager des estimations de (p_{ij}, N_{ij}), soit $(p_{ijk}^{(l)}, N_{ijk}^{(l)})$, $l = 1, ..., L$ le résultat de l'application de la chaîne d'échantillonnage de Gibbs après la période initiale d'itérations. Définissons

$$p_{ij}^{(l)} = \frac{\sum_{k=1}^{K} p_{ijk}^{(l)} N_{ijk}^{(l)}}{\sum_{k=1}^{K} N_{ijk}^{(l)}}, \quad l = 1, ..., L.$$

On peut alors approcher la moyenne et la variance de distribution a postériori de p_{ij} respectivement par

$$\hat{E}(p_{ij}) = \frac{1}{L}\sum_{l=1}^{L} p_{ij}^{(l)} \quad (4)$$

et

$$\hat{V}(p_{ij}) = \frac{1}{L-1}\sum_{l=1}^{L} \{p_{ij}^{(l)}\}^2 - \frac{L}{L-1}\{\hat{E}(p_{ij})\}^2, \quad (5)$$

De même, définissons $N_{ij}^{(l)} = \sum_{k=1}^{K} N_{ijk}^{(l)}$. On peut approcher la moyenne et la variance de distribution a postériori de N_{ij} à l'aide aussi du produit de la simulation MCCM.

Nous nous intéressons maintenant à la prévision bayésienne de la récolte inobservée de dindons par les distributions prévisionnelles a postériori. Soit y_{ijk} le nombre de

dindons récoltés dans le comté i et la période j par les chasseurs échantillonnés de la strate k; dans le cas des chasseurs non échantillonnés, soit y^*_{ijk} le nombre de dindons récoltés dans le comté i et la période j pour la strate k. Ainsi, le nombre de dindons récoltés dans le comté i et la période j par les chasseurs de la strate k est $h_{ijk} = y_{ijk} + y^*_{ijk}$. Dans ce cas, y_{ijk} est une valeur connue et nous avons seulement y^*_{ijk} à trouver. Nous pouvons voir y^*_{ijk} étant donné $(n_{ijk}, N_{ijk}, p_{ijk})$ comme une variable aléatoire binomiale ayant la forme de (1). Ainsi,

$$(y^*_{ijk} \mid n_{ijk}, N_{ijk}, p_{ijk}) \overset{\text{ind}}{\sim} \text{Binomial}(N_{ijk} - n_{ijk},\ p_{ijk}). \quad (6)$$

Soit $(p^{(l)}_{ijk}, N^{(l)}_{ijk})$, $l = 1, ..., L$, le produit de l'application de la chaîne de Gibbs après une période initiale d'itérations. La moyenne prévisionnelle de y^*_{ijk} étant donné $\boldsymbol{d} = \{(y_{ijk}, n_{ijk}) : i = 1, .., I,\ j = 1, ..., J,\ k = 1, ..., K\}$ est alors

$$\hat{E}(y^*_{ijk} \mid \boldsymbol{d}) = \frac{1}{L}\sum_{l=1}^{L} (N^{(l)}_{ijk} - n_{ijk})\ p^{(l)}_{ijk}. \quad (7)$$

Enfin, nous écrivons

$$\hat{h}_{ijk} = y_{ijk} + \hat{E}(y^*_{ijk} \mid \boldsymbol{d}),$$

$$\hat{h}_i = \sum_{j=1}^{J}\sum_{k=1}^{K} \hat{h}_{ijk}. \quad (8)$$

La variabilité de h_i est donnée par

$$\hat{V}(h_i \mid \boldsymbol{d}) = \frac{1}{L-1}\sum_{l=1}^{L}\sum_{j=1}^{J}\sum_{k=1}^{K}(N^{(l)}_{ijk} - n_{ijk})p^{(l)}_{ijk}(1-p^{(l)}_{ijk})$$

$$+ \frac{1}{L-1}\sum_{l=1}^{L}\left\{\sum_{j=1}^{J}\sum_{k=1}^{K}[y_{ijk} + (N^{(l)}_{ijk} - n_{ijk})p^{(l)}_{ijk}] - \hat{h}^{(l)}_i\right\}^2.$$

La preuve est une application simple des espérances conditionnelles. À noter que h_{ijk} est défini comme variable aléatoire. Il n'est pas égal à $N_{ijk}p_{ijk}$, pas plus que y_{ijk} ne l'est à $n_{ijk}p_{ijk}$. On ne devrait donc pas se servir simplement de $\hat{N}_{ijk}\hat{p}_{ijk}$ pour estimer h_{ijk}, bien que la valeur puisse être très proche si N_{ijk} est suffisamment élevé.

2.4 Ajustement du modèle

Nous avons exécuté 15 000 itérations pour la chaîne de Gibbs, dont 5 000 ont été écartées comme séquence initiale. Ainsi, les estimations a postériori sont fondées sur 10 000 échantillons autocorrélés de la distribution a postériori. Pour contrôler la convergence de l'échantillonnage de Gibbs, nous avons recouru à la méthode de diagnostic de Heidelberger et Welch (1983), ainsi qu'à un contrôle graphique des parcours d'échantillon. Nous avons procédé à des diagnostics de convergence et à la détermination des valeurs sommaires a postériori au moyen du logiciel BOA (Smith 2005). Laissons les composantes de la variance $\delta^{(e)}_{ak}$ et $\delta^{(u)}_{ak}$ suivre une distribution a priori non informative avec $\text{IG}(2,1)$, ce qui donne l'unité comme moyenne et l'infini comme variance. Nous donnons des valeurs a priori informatives à $\theta_{1jk} \sim N(-1,5, 16)$ et $\theta_{2jk} \sim N(4,25)$. Pour obtenir une distribution a priori informative, nous avons modélisé les deux étapes suivantes, parce qu'une distribution a priori informative qui serait entièrement arbitraire pourrait mener à des estimations a postériori peu sûres (Wasserman 2000). D'abord, nous avons pris des valeurs a priori non informatives (mais conjuguées) pour dégager des moyennes et des écarts-types a postériori pour chacun des paramètres considérés par des simulations MCCM. Nous avons ensuite posé que les moyennes a priori informatives étaient proches des moyennes a postériori de θ_{1jk} et θ_{2jk} et que les écarts-types a priori informatifs étaient environ décuples des écarts-types a postériori une fois les estimations a postériori obtenues par distribution a priori non informative. Les estimations issues de ces deux manières de procéder étaient très convergentes, mais le modèle à distributions a priori informatives donnait des variances moindres.

Nous avons ajusté un grand nombre de modèles simplifiés à des fins de comparaison et de contrôle de modélisation. Comme moyen de sélection de modèle, nous avons appliqué le critère d'information par l'écart quadratique proposé par Spiegelhalter, Best, Carlin et van der Linde (2002). C'est là une généralisation du critère d'information d'Akaike qui mesure les carrés d'écart pour une exécution du modèle MCCM et qui comporte une mesure pénalisée de l'ajustement où la dimensionnalité du modèle entre en ligne de compte. Des valeurs plus basses du critère d'information par écart quadratique indiquent un meilleur ajustement. Notre modèle avec valeurs a priori informatives présente 13 910,5 comme valeur d'écart quadratique. Pour aucun autre modèle, la valeur d'écart n'était significativement réduite par rapport à ce modèle. Le modèle à valeurs a priori non informatives pour θ_{1jk} et θ_{2jk} se caractérise par une valeur d'écart de 14 700,4. Pour un modèle réduit à paramètres de corrélation commune $\rho_{11} = \rho_{12} = \rho_{13} = \rho_{14} = \rho_{15}$ et $\rho_{21} = \rho_{22} = \rho_{23} = \rho_{24} = \rho_{25}$, la valeur d'écart s'établit à 13 896,9.

Comme autre contrôle de modélisation, nous avons calculé les moyennes globales (dans tout l'État du Missouri) des estimations en prenant à la fois les simples estimations « naïves » du plan d'échantillonnage et les estimations bayésiennes qui suivent au tableau 2. Au niveau de l'État, les tailles d'échantillon sont suffisamment grandes pour que les estimations basées sur le plan d'échantillonnage soient jugées fiables. Les estimations selon le plan correspondent de près aux estimations par modèle. Ainsi, l'estimateur de Bayes donne de bons résultats pour la propriété de cohérence avec les estimations selon le plan (voir You et Rao 2003, et Jiang et Lahiri 2006). Nous notons que les

estimations du premier jour sont un peu moindres à cause du lissage, mais que les estimations de taux de succès demeurent bien plus élevées pour ce premier jour que pour toute autre période d'intérêt.

Tableau 2 Estimation de taux de succès pour la MTHS de 1998

Strate nominale	Période	Dindons tués	Expéditions	Estimations nominales	Estimations bayésiennes
Non-résidents	1er jour	105	424	0,248	0,236
	1re semaine	271	1 759	0,154	0,155
	2e semaine	198	1 481	0,134	0,132
	3e semaine	75	635	0,118	0,120
Nord	1er jour	82	452	0,181	0,171
	1re semaine	211	1 535	0,138	0,137
	2e semaine	161	1 663	0,097	0,097
	3e semaine	99	1 129	0,088	0,090
Sud	1er jour	138	751	0,184	0,178
	1re semaine	224	2 475	0,091	0,092
	2e semaine	186	2 375	0,078	0,069
	3e semaine	90	1 748	0,052	0,053
St. Louis	1er jour	54	346	0,156	0,149
	1re semaine	91	1 280	0,071	0,073
	2e semaine	85	1 257	0,068	0,069
	3e semaine	45	828	0,054	0,059
Kansas City	1er jour	52	262	0,199	0,181
	1re semaine	95	894	0,106	0,108
	2e semaine	92	942	0,098	0,099
	3e semaine	55	611	0,090	0,093
Ensemble des régions	1er jour	431	2 235	0,193	0,181
	1re semaine	892	7 943	0,112	0,111
	2e semaine	722	7 718	0,094	0,088
	3e semaine	364	4 951	0,074	0,075
Total		2 409	22 847	0,105	0,102

2.5 Analyse des données pour la MTHS

On trouvera au tableau 3 l'énumération des moyennes et des écarts-types a postériori des paramètres pour l'ensemble des strates nominales. À noter que les estimations de moyenne et de variance de θ_{12k} et θ_{13k}, ainsi que de θ_{22k} et θ_{23k} pour $k = 1, ..., 5$, sont approximativement égales. Il y a donc lieu de penser que les estimations respectives de taux de succès et de « pression de chasse » sont semblables pour les semaines 1 et 2 de la saison de chasse. Une différence subsiste pour la première journée et la semaine 3 de la saison cependant (Vangilder, Sheriff et Olsen 1990; Kimmel 2001).

Nous signalons aussi les estimations paramétriques en corrélation de ρ_{1k} et ρ_{2k}. Les estimations de ρ_{2k} pour la « pression de chasse » sont toutes proches de leur borne supérieure, d'où l'impression que les comtés sont en corrélation étroite pour l'estimation du nombre d'expéditions de chasse. Pour la plupart, les estimations de ρ_{1k} sont aussi éloignées de plus de deux écarts-types de la valeur nulle.

Tableau 3 Moyennes (écarts-types) a postériori **des paramètres du modèle**

	Non-résidents	Nord	Sud	St. Louis	Kansas City
$\hat{\delta}_{e1}$	0,133 (0,034)	0,117 (0,030)	0,136 (0,036)	0,164 (0,051)	0,143 (0,040)
$\hat{\delta}_{e2}$	0,096 (0,017)	0,044 (0,007)	0,041 (0,006)	0,061 (0,011)	0,060 (0,011)
$\hat{\delta}_{z1}$	0,146 (0,040)	0,148 (0,041)	0,140 (0,038)	0,166 (0,053)	0,186 (0,059)
$\hat{\delta}_{z2}$	0,570 (0,092)	0,511 (0,081)	0,633 (0,094)	0,373 (0,062)	0,356 (0,063)
$\hat{\theta}_{11}$	-1,209 (0,131)	-1,530 (0,130)	-1,244 (0,135)	-1,420 (0,159)	-1,472 (0,160)
$\hat{\theta}_{12}$	-1,613 (0,109)	-1,748 (0,111)	-1,867 (0,130)	-1,983 (0,147)	-1,928 (0,150)
$\hat{\theta}_{13}$	-1,801 (0,113)	-2,081 (0,113)	-2,152 (0,129)	-2,032 (0,148)	-2,007 (0,148)
$\hat{\theta}_{14}$	-1,830 (0,132)	-2,135 (0,123)	-2,354 (0,138)	-2,115 (0,156)	-2,008 (0,153)
$\hat{\theta}_{21}$	3,368 (0,098)	3,333 (0,098)	3,297 (0,087)	3,305 (0,088)	3,308 (0,098)
$\hat{\theta}_{22}$	4,582 (0,097)	4,361 (0,093)	4,371 (0,084)	4,349 (0,081)	4,238 (0,090)
$\hat{\theta}_{23}$	4,404 (0,097)	4,429 (0,093)	4,459 (0,088)	4,293 (0,083)	4,229 (0,093)
$\hat{\theta}_{24}$	3,693 (0,101)	4,093 (0,095)	4,039 (0,086)	3,934 (0,082)	3,891 (0,094)
$\hat{\rho}_1$	0,153 (0,019)	0,103 (0,071)	0,163 (0,011)	0,158 (0,016)	0,125 (0,068)
$\hat{\rho}_2$	0,168 (0,007)	0,172 (0,003)	0,171 (0,004)	0,172 (0,004)	0,172 (0,004)

Les simples estimations « naïves » selon le plan d'échantillonnage pour les taux de succès dans chaque comté figurent à la première ligne de la figure 2. Les estimations selon le plan pour les comtés varient de 0 à 0,55 et, dans certains comtés, il n'y a pas d'estimation, car $n_{ij} = 0$. Par ailleurs, on trouvera à la deuxième ligne de la figure 2 les estimations bayésiennes de ces mêmes taux pour le comté i et la période j. Ces estimations de modèle bayésien des taux de succès donnent une plage de variation bien plus raisonnable qui s'étend de 0,03 à 0,30. Nous ne nous attendons pas à ce qu'un comté présente une estimation nulle de taux de succès. Il faut aussi dire que les dindons sont très difficiles à chasser et qu'un haut taux de succès n'est pas à prévoir (He et Sun 1998; He et Sun 2000; Oleson et He 2004). Des valeurs moindres comme celles que dégage le modèle parlent plus à l'intuition donc. On trouvera à la troisième ligne de la figure 2 les écarts-types des estimations du modèle de taux de succès.

Les estimations tendent à décroître tout au long de la saison de chasse. Ajoutons que les estimations les plus élevées se rattachent à la partie nord de l'État du Missouri. C'est aussi ce qu'on a pu vérifier dans des analyses a priori de la MTHS à l'aide des données de 1996 (He et Sun 2000; Oleson et He 2004). Les estimations les plus hautes ne se présentent cependant pas dans les mêmes comtés. Les plus hauts taux d'estimation à l'aide des données de 1998 sont un peu plus à l'est qu'ils ne l'étaient en 1996, ce que l'on doit à une tendance temporelle, ainsi que le font observer les défenseurs du mouvement de conservation.

Nous produisons également des estimations de la taille de population N_{ijk}. Les estimations de modélisation de N_{ij} figurent à la première ligne de la figure 3. Leurs erreurs-types sont présentées à la deuxième ligne. Les valeurs représentées pour les semaines 1, 2 et 3 sont les moyennes quotidiennes de ces semaines. Il est clair que plus de gens chassaient la première journée que tout autre jour de la saison de chasse.

La première carte de la figure 4 livre les données réelles des postes de contrôle. À l'aide de la formule (8), nous prévoyons par modèle le nombre total de dindons récoltés h_i, qui figure dans la deuxième carte de la première ligne. À considérer ces chiffres, on peut penser que les prévisions de modélisation sont d'une précision acceptable. Nous nous attendons à une légère surprévision comme il ressort de la comparaison des courbes. Un certain nombre de chasseurs pourraient sous-estimer la fréquence de leurs expéditions et surestimer le nombre de volatiles récoltés. Le surdénombrement s'explique peut-être par le nombre de dindons récoltés, mais non déclarés aux postes de contrôle. Une autre explication serait peut-être que les chasseurs qui ont répondu au questionnaire sont ceux qui ont eu le plus de succès et que les non-répondants en ont eu peut-être moins ou n'ont pas chassé du tout. Si on compare les valeurs prévues par modèle aux valeurs effectives de récolte, on a la confirmation que le modèle convient et livre des prévisions exactes. Il faut aussi dire que bien des États ne disposent pas de postes de contrôle et qu'on a la preuve qu'ils auraient la possibilité d'obtenir des estimations appropriées au niveau des domaines même si les échantillons de domaine devaient être modestes.

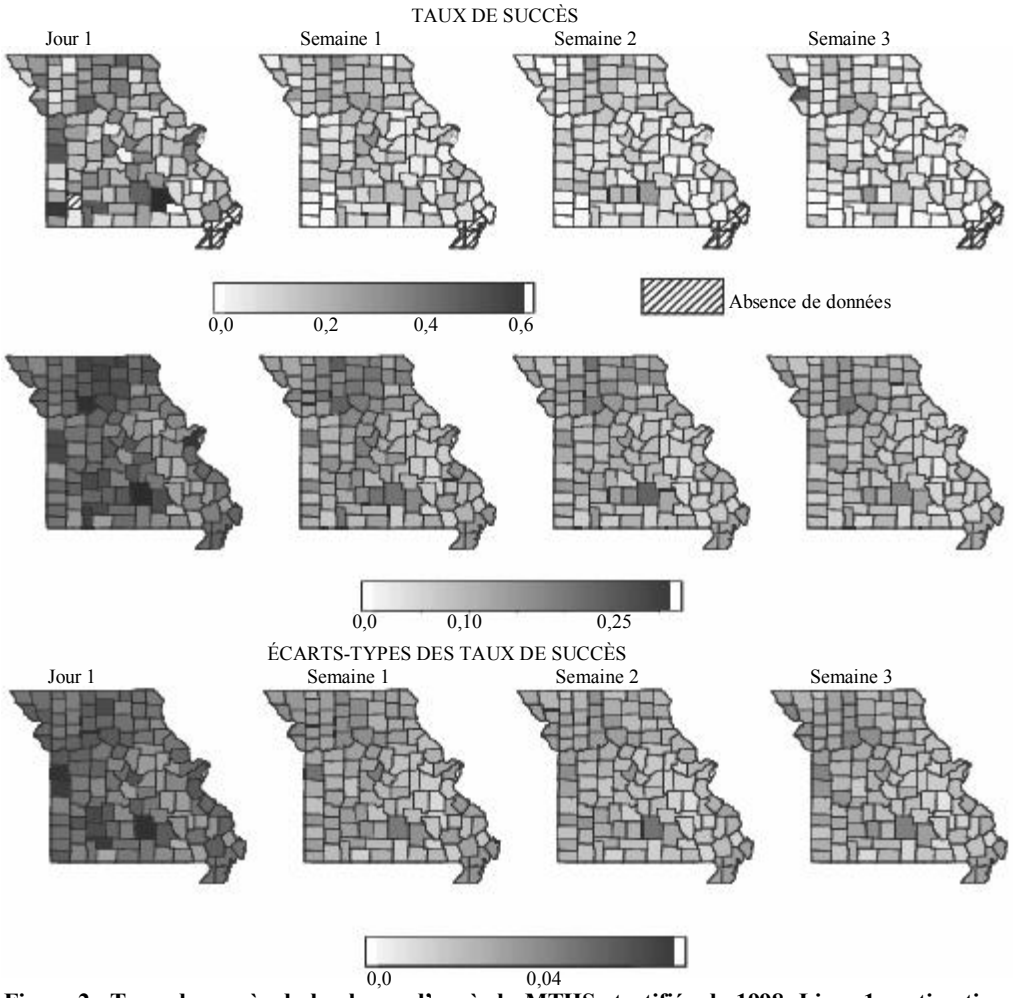


Figure 2 Taux de succès de la chasse d'après la MTHS stratifiée de 1998. Ligne 1 : estimations « nominales » des taux de réussite; ligne 2 : estimations bayésiennes de ces taux; ligne 3 : écarts-types des estimations bayésiennes des taux de succès

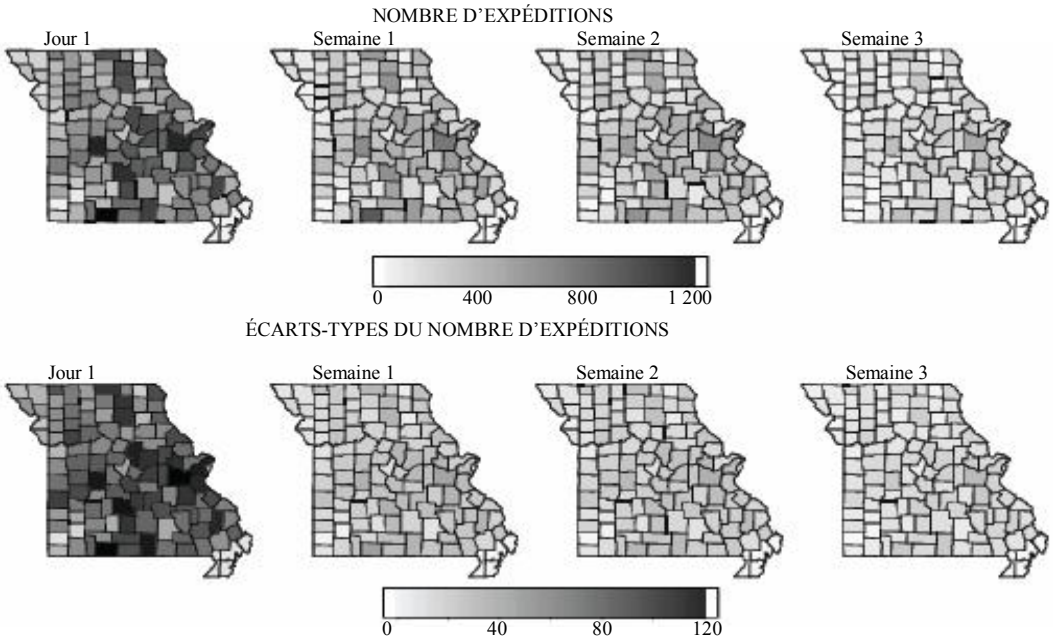


Figure 3 Nombre estimé d'expéditions de chasse d'après la MTHS stratifiée de 1998. Ligne 1 : estimations bayésiennes du nombre d'expéditions; ligne 2 : écarts-types de ces estimations bayésiennes

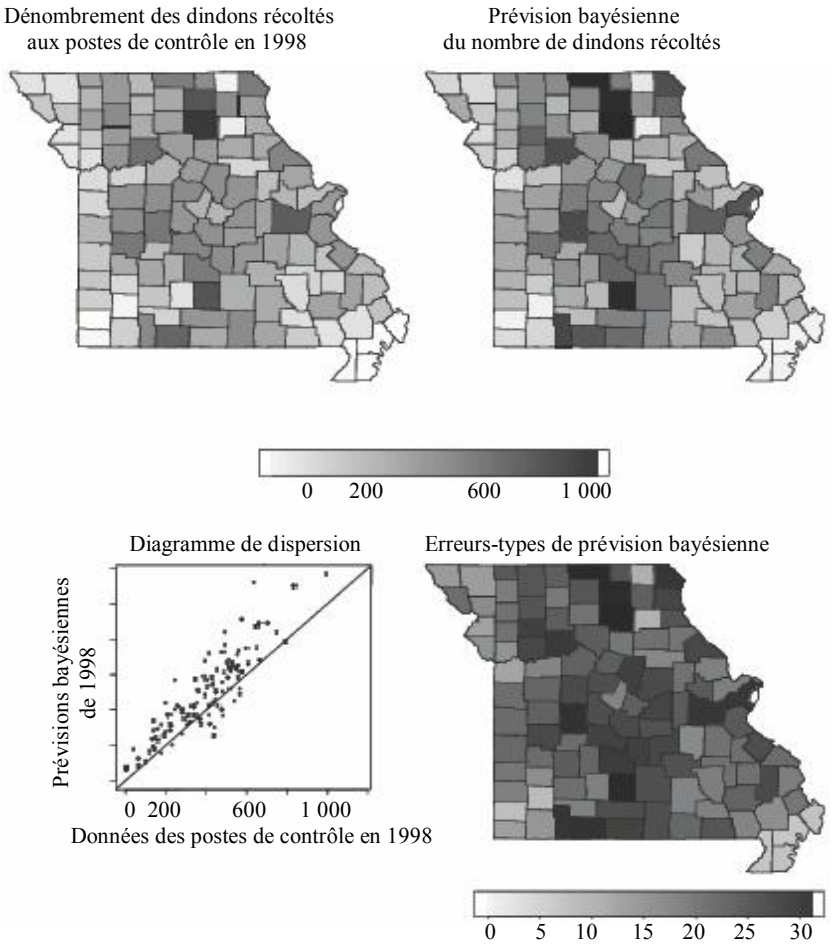


Figure 4 Données des postes de contrôle en 1998, prévisions bayésiennes, diagramme de dispersion de ces postes selon les prévisions et erreurs-types de ces dernières

3. Formules générales pour un MMLG à deux variables

3.1 Modèle général

Nous ne tenons pas compte de la composante temporelle dans le modèle général par souci de simplicité. Soit n_{ik} et Y_{ik} les tailles d'échantillon et les variables d'intérêt pour le domaine d'étude i et la strate nominale k respectivement. Nous posons que $\{Y_{ik}, n_{ik}\}$ étant donné les paramètres inconnus $\{\eta_{ik}, \omega_{ik}, \phi_1, \phi_2\}$, avec $i = 1, ..., I$, et $k = 1, ..., K$ sont indépendants. Nous posons aussi que la fonction de densité conditionnelle de Y_{ik} étant donné n_{ik} appartient à la famille suivante de fonctions de densité de probabilité :

$$g_1(y_{ik} \mid \eta_{ik}, n_{ik}, \phi_1) = \exp[A_1(\phi_1)\{y_{ik}\eta_{ik} - B_1(\eta_{ik}, n_{ik})\} + C_1(y_{ik}, n_{ik}, \phi_1)], \quad (9)$$

où η_{ik} est un paramètre inconnu. On forme souvent l'hypothèse que le paramètre d'échelle ϕ_1 est connu. La fonction de densité de probabilité de n_{ik} appartient à la famille

$$g_2(n_{ik} \mid \omega_{ik}, \phi_2) = \exp[A_2(\phi_2)\{n_{ik}\omega_{ik} - B_2(\omega_{ik})\} + C_2(n_{ik}, \phi_2)], \quad (10)$$

où ω_{ik} est un paramètre inconnu égal à une fonction de taille de population N_{ik} et où ϕ_2 est souvent connu. La codensité de Y_{ik} et n_{ik} (MMLG à deux variables) est alors

$$p(y_{ik}, n_{ik} \mid \eta_{ik}, \omega_{ik}, \phi_1, \phi_2) = g_1(y_{ik} \mid \eta_{ik}, n_{ik}, \phi_1) g_2(n_{ik} \mid \omega_{ik}, \phi_2). \quad (11)$$

La famille de distributions (10) est souvent appelée « modèle linéaire généralisé » et comporte des distributions binomiales, de Poisson, normales et gamma (voir, par exemple, Sun et coll. 2000). La famille de distributions (9) est une généralisation d'un tel modèle linéaire généralisé par intégration d'un paramètre de plus. Quatre cas d'espèce de (9) sont les distributions binomiales, de Poisson, normales et gamma qui font toutes partie de la famille exponentielle.

Le MMLG à deux variables est applicable lorsque les estimations offrent de l'intérêt au niveau des domaines d'étude et que la taille d'échantillon n_{ik} est considérée à la fois comme relevant du hasard et de l'observation. Elle a aussi son utilité lorsque des estimations de N_{ik} sont requises.

Un modèle mixte linéaire peut s'utiliser en distribution a priori de η_{ik} pour rendre compte de sa variabilité. On pourrait toutefois s'intéresser tant à η_{ik} qu'à ω_{ik} ou à une fonction des deux, ω_{ik} étant souvent fonction de la taille de population. Dans ce cas, il faudrait modéliser η_{ik} et ω_{ik} simultanément. Une catégorie générale de MMLG pour η_{ik} et ω_{ik} pourrait être

$$h_1(\boldsymbol{\eta}_k) = \mathbf{X}_1 \boldsymbol{\theta}_{1k} + \mathbf{S}_1 \mathbf{u}_{1k} + \mathbf{e}_{1k}, \quad (12)$$

$$h_2(\boldsymbol{\omega}_k) = \mathbf{X}_2 \boldsymbol{\theta}_{2k} + \mathbf{S}_2 \mathbf{u}_{2k} + \mathbf{e}_{2k}. \quad (13)$$

Avec $a = 1, 2$, $\mathbf{X}_a = \{x_{aik}\}$ et $\mathbf{S}_a = \{s_{aik}\}$ sont des matrices connues de plan d'échantillonnage. Le vecteur $\boldsymbol{\theta}_{ak}$ est le vecteur des effets fixes et $\boldsymbol{u}_{ak}$, celui des effets variables; e_{aik} représente les effets résiduels indépendants et $e_{aik} \sim N(0, \delta_{ak}^{(e)})$. On pose en outre que $\boldsymbol{u}_{ak}$ et e_{aik} sont indépendants l'un de l'autre.

3.2 Distributions a priori supplémentaires

Pour compléter le modèle hiérarchisé de Bayes, nous devons spécifier les distributions a priori pour $(\boldsymbol{\theta}_{ak}, \mathbf{u}_{ak}, \delta_{ak}^{(e)})$, $a = 1, 2$, $k = 1, ..., K$. La distribution a priori commune pour les effets fixes $\boldsymbol{\theta}_{ak}$ est normale avec une variance importante ou une valeur a priori constante. Les effets aléatoires sont en corrélation spatiale dans bien des cas. La densité peut être de la forme ARC où la codensité est donnée par l'équation (3) avec $\mathbf{B}_{ak} = \mathbf{I} - \rho_{ak}\mathbf{C}$. Mentionnons enfin qu'une distribution a priori commune des composantes de la variance $\delta_{ak}^{(e)}$ est une distribution gamma inverse.

Dans l'évaluation de la distribution a postériori, on peut employer des méthodes MCCM comme la formule d'échantillonnage de Gibbs pour tirer des échantillons de cette distribution. Nous présentons les distributions conditionnelles intégrales dans le cas général qui suit.

Fait 1 Soit $(\Omega \mid \cdot)$ les distributions conditionnelles de Ω étant donné tous les autres paramètres et soit $[\Omega \mid \cdot]$ la densité conditionnelle. Voici les densités a postériori conditionnelles de η_{ik} et ω_{ik} :

(i) $[\eta_{ik} \mid \cdot] \propto \exp(A_1(\phi_1)[y_{ik}\eta_{ik} - B_1(\eta_{ik}, n_{ik})] - 1/2\delta_{1k}^{(e)} [h_1(\eta_{ik}) - \mathbf{x}_{1i}^t \boldsymbol{\theta}_{1k} - S_{1i}u_{1ik}]^2) h_1'(\eta_{ik})$ ou, d'une manière équivalente $v_{1ik} = h_1(\eta_{ik})$ ayant la densité conditionnelle

$$[v_{1ik} \mid \cdot] \propto \exp\left\{\begin{array}{l} A_1(\phi_1)\,[y_{ik}\, h_1^{-1}(v_{1ik}) \\ -B_1(h_1^{-1}(v_{1ik}), n_{ik})] \\ -1/2\delta_{1k}^{(e)}\,(v_{1ik} - \mathbf{x}_{1i}^t\boldsymbol{\theta}_{1k} - S_{1i}u_{1ik})^2 \end{array}\right\}.$$

(ii) $[\omega_{ik} \mid \cdot] \propto \exp(A_2(\phi_2)[n_{ik}\omega_{ik} - B_2(\omega_{ik})] - 1/2\delta_{2k}^{(e)} [h_2(\omega_{ik}) - \mathbf{x}_{2i}^t \boldsymbol{\theta}_{2k} - S_{2i}u_{2ik}]^2) h_2'(\omega_{ik})$ ou, d'une manière équivalente, $v_{2ik} = h_2(\omega_{ik})$ ayant la densité conditionnelle

$$[v_{2ik} \mid \cdot] \propto \exp\left\{\begin{array}{l} A_2(\phi_2)[n_{ik}h_2^{-1}(v_{2ik}) \\ -B_2(h_2^{-1}(v_{2ik}))] \\ -1/2\delta_{2k}^{(e)}\,(v_{2ik} - \mathbf{x}_{2i}^t\boldsymbol{\theta}_{2k} - S_{2i}u_{2ik})^2 \end{array}\right\}.$$

Pour $a = 1, 2$, nous obtenons les distributions conditionnelles suivantes :

(iii) $(\boldsymbol{\theta}_{ak}|\cdot)$ sont mutuellement indépendants et présentent les distributions a postériori conditionnelles

$$N_I\left(\begin{array}{l}\left(\frac{1}{\tau_{ak}}\mathbf{I}+\frac{1}{\delta_{ak}^{(e)}}\mathbf{X}_a^t\mathbf{X}_a\right)^{-1}\\ \left(\frac{1}{\delta_{ak}^{(e)}}\mathbf{X}_a^t(\mathbf{v}_{ak}-\mathbf{u}_{ak})+\frac{\mu_{ak}}{\tau_{ak}}\mathbf{1}\right),\\ \left(\frac{1}{\tau_{ak}}\mathbf{I}+\frac{1}{\delta_{ak}^{(e)}}\mathbf{X}_a^t\mathbf{X}_a\right)^{-1}\end{array}\right).$$

(iv) $(\mathbf{u}_{ak}|\cdot)$ sont mutuellement indépendants et présentent les distributions a postériori conditionnelles

$$N_I\left(\begin{array}{l}\left(\frac{1}{\delta_{ak}^{(e)}}\mathbf{S}_a^t\mathbf{S}_a+\frac{1}{\delta_{ak}^{(u)}}\mathbf{B}_{ak}\right)^{-1}\\ \left(\frac{1}{\delta_{ak}^{(e)}}\mathbf{S}_a^t(\mathbf{v}_{ak}-\mathbf{X}_a\boldsymbol{\theta}_{ak}\right),\\ \left(\frac{1}{\delta_{ak}^{(e)}}\mathbf{S}_a^t\mathbf{S}_a+\frac{1}{\delta_{ak}^{(u)}}\mathbf{B}_{ak}\right)^{-1}\end{array}\right).$$

(v) $(\delta_{ak}^{(e)}|\cdot)$ ont les distributions a postériori gamma inverses

$$\text{Inverse Gamma}\left(\begin{array}{l}\alpha_{ak}^{(e)}+I/2,\beta_{ak}^{(e)}\\ +1/2(\mathbf{v}_{ak}-\mathbf{X}_a\boldsymbol{\theta}_{ak}-\mathbf{S}_a\mathbf{u}_{ak})^t\\ (\mathbf{v}_{ak}-\mathbf{X}_a\boldsymbol{\theta}_{ak}-\mathbf{S}_a\mathbf{u}_{ak})\end{array}\right).$$

(vi) $(\delta_{ak}^{(u)}|\cdot)\sim$ Inverse Gamma $(\alpha_{ak}^{(u)}+I/2,\ \beta_{ak}^{(u)}+1/2\,\mathbf{u}_{ak}^t\mathbf{B}_{ak}\mathbf{u}_{ak})$.

vii) $[\rho_{ak}|\cdot]\propto|\mathbf{B}_{ak}|^{1/2}\exp\left\{-\frac{\boldsymbol{u}_{ak}^t\mathbf{B}_{ak}\boldsymbol{u}_{ak}}{2\delta_{ak}^{(u)}}\right\}$.

(viii)Si ϕ_1 a la densité a priori $p(\phi_1)$,

$$[\phi_1|\cdot]\propto p(\phi_1)\exp\left\{\begin{array}{l}A_1(\phi_1)[y_{ik}h_1^{-1}(v_{1ik})\\ -B_1(h_1^{-1}(v_{1ik}),n_{ik})]\\ +C_1(y_{ik},n_{ik},\phi_1)\end{array}\right\}.$$

(ix) Si ϕ_2 a la densité a priori $p(\phi_2)$,

$$[\phi_2|\cdot]\propto p(\phi_2)\exp\left\{\begin{array}{l}A_2(\phi_2)[n_{ik}h_2^{-1}(v_{2ik})\\ -B_2(h_2^{-1}(v_{2ik}))]\\ +C_2(n_{ik},\phi_2)\end{array}\right\}.$$

Lorsqu'on examine les parties (i), (ii) et (vii) du « Fait 1 », les densités conditionnelles de η_{ik},ω_{ik} et ρ_{ak} sont souvent logconcaves.

3.3 Estimations quantitatives des domaines d'étude

Il est souvent intéressant d'estimer certaines quantités relatives aux domaines d'étude. Pour estimer, par exemple, une quantité du domaine i, prenons

$$\psi_i = f_i(\eta_{i1},\omega_{i1};...;\eta_{iK},\omega_{iK}).$$

On peut aisément calculer des estimations bayésiennes de ψ_i en se fondant sur un échantillon aléatoire de la codistribution a postériori. Soit $(\eta_{ik}^{(l)},\omega_{ik}^{(l)})$, $l=1,...,L$ et $k=1,...,K$ le produit de simulation MCCM et la définition

$$\psi_i^{(l)} = f_i(\eta_{i1}^{(l)},\omega_{i1}^{(l)};...;\eta_{iK}^{(l)},\omega_{iK}^{(l)}).$$

Étant donné $\boldsymbol{y}=\{y_{ik}, i=1,...,I, k=1,...,K\}$, avec la moyenne et la variance a postériori de ψ_i, on peut approcher les formes générales de (4) et (5) par

$$\hat{E}(\psi_i|\,\boldsymbol{y}) = \frac{1}{L}\sum_{l=1}^{L}\psi_i^{(l)}$$

et

$$\hat{V}(\psi_i|\,\boldsymbol{y}) = \frac{1}{L-1}\sum_{l=1}^{L}\{\psi_i^{(l)}\}^2 - \frac{L}{L-1}\{\hat{E}(\psi_i|\,\mathbf{y})\}^2,$$

respectivement.

3.4 Prévisions quantitatives des domaines d'étude

Soit N_{ik} la taille de population et Y_{ik}^* la valeur d'intérêt pour les chasseurs non échantillonnés du domaine d'étude i et de la strate nominale k. Nous désirons prévoir les quantités $Y_i^* = \sum_{k=1}^K Y_{ik}^*$, la valeur totale de réponse dans le domaine d'étude i. Pour une valeur donnée de n_{ik}, Y_{ik}^* devrait être de la même famille que Y_{ik}, de sorte que

$$g_3(y_{ik}^*|\,\eta_{ik},n_{ik},N_{ik},\phi_1) = \exp\left[\begin{array}{l}A_1(\phi_1)\{y_{ik}^*\eta_{ik}\\ -B_1(\eta_{ik},N_{ik}-n_{ik})\}\\ +C_1(y_{ik}^*,N_{ik}-n_{ik},\phi_1)\end{array}\right].$$

C'est là simplement (9), n_{ik} étant remplacé par $N_{ik}-n_{ik}$. Pour simplifier la notation, soit $\boldsymbol{\xi}$ les paramètres du modèle et $\boldsymbol{d}$ les données. (Dans ce cas, $\boldsymbol{d}=\{(y_{ik},n_{ik}):\ i=1,...,I, k=1,...,K\}$ et $\boldsymbol{\xi}$ pourrait comprendre les paramètres de modélisation y_{ik} et n_{ik}). Pour la distribution a priori $\pi(\boldsymbol{\xi})$ (propre ou impropre), nous obtenons la densité a postériori $[\boldsymbol{\xi}|\,\boldsymbol{d}]\propto f(\boldsymbol{d}|\boldsymbol{\xi})\pi(\boldsymbol{\xi})$.

Posons qu'une nouvelle observation y_{ik}^* suit $g_3(y_{ik}^*|\,\boldsymbol{d},\boldsymbol{\xi})$, qui pourrait dépendre de $\boldsymbol{d}$. La densité prévisionnelle de y_{ik}^* étant donné $\boldsymbol{d}$ se formule ainsi :

$$[y_{ik}^*|\,\boldsymbol{d}] = \int_{\boldsymbol{\xi}} g_3(y_{ik}^*|\,\boldsymbol{\xi},\boldsymbol{d})[\boldsymbol{\xi}|\,\boldsymbol{d}]\,d\boldsymbol{\xi}.$$

Dans la fonction de perte par erreur quadratique, le meilleur « prédicteur » est alors la moyenne prévisionnelle donnée par

$$E(y_{ik}^* \mid \boldsymbol{d}) = \int_{\xi} \left\{ \int_{y_{ik}^*} y_{ik}^* \, g_3(y_{ik}^* \mid \xi, \boldsymbol{d}) \, dy_{ik}^* \right\} [\xi \mid \boldsymbol{d}] \, d\xi.$$

De même, le meilleur « prédicteur » de $h(y_{i1}^*, ..., y_{iK}^*)$ étant donné $\boldsymbol{d}$ est

$$E\{h(y_{i1}^*, ..., y_{iK}^*) \mid \boldsymbol{d}\} = \int_{\xi} E\{h(y_{i1}^*, ..., y_{iK}^*) \mid \xi, \boldsymbol{d}\} \, [\xi \mid \boldsymbol{d}] \, d\xi, \quad (14)$$

où

$$E\{h(y_{i1}^*, ..., y_{iK}^*) \mid \xi, \boldsymbol{d}\} = \int_{(y_{i1}^*, ..., y_{iK}^*)} h(y_{i1}^*, ..., y_{iK}^*) \prod_{k=1}^{K} [y_{ik}^* \mid \xi, \boldsymbol{d}] \, dy_{i1}^* ... dy_{iK}^*. \quad (15)$$

Pour la famille de distributions $g_3(y_{ik}^* \mid \xi, \boldsymbol{d})$, le côté droit de (15) est souvent une expression de forme fermée.

On peut aisément calculer les valeurs de prévision bayésienne de (14) en se fondant sur un échantillon aléatoire tiré de la codistribution a postériori de ξ. Ainsi, soit $\xi^{(l)}, l = 1, ..., L,$ le produit de la simulation MCCM. On peut alors approcher la moyenne prévisionnelle a postériori (14) par

$$\hat{E}\{h(y_{i1}^*, ..., y_{iK}^*) \mid \boldsymbol{d}\} = \frac{1}{L} \sum_{l=1}^{L} E\{h(y_{i1}^*, ..., y_{iK}^*) \mid \xi^{(l)}, \boldsymbol{d}\}.$$

La variance prévisionnelle a postériori de $h(y_{i1}^*, ..., y_{iK}^*)$ étant donné $\boldsymbol{d}$ peut ainsi s'exprimer :

$$V\{h(y_{i1}^*, ..., y_{iK}^*) \mid \boldsymbol{d}\} = E[V\{h(y_{i1}^*, ..., y_{iK}^*) \mid \xi, \boldsymbol{d}\} \mid \boldsymbol{d}] + V[E\{h(y_{i1}^*, ..., y_{iK}^*) \mid \xi, \boldsymbol{d}\} \mid \boldsymbol{d}].$$

On peut aussi approcher la variance prévisionnelle par

$$\hat{V}\{h(y_{i1}^*, ..., y_{iK}^*) \mid \boldsymbol{d}\} = \frac{1}{L} \sum_{l=1}^{L} V\{h(y_{i1}^*, ..., y_{iK}^*) \mid \xi^{(l)}, \boldsymbol{d}\} + \frac{1}{L} \sum_{l=1}^{L} [E\{h(y_{i1}^*, ..., y_{iK}^*) \mid \xi^{(l)}, \boldsymbol{d}\} - \hat{E}\{h(y_{i1}^*, ..., y_{iK}^*) \mid \boldsymbol{d}\}]^2.$$

4. Observations

Dans cet article, nous avons conçu un MMLG à deux variables pour deux paramètres canoniques inconnus. Il le fallait pour obtenir des estimations lorsque la taille d'échantillon est aléatoire et qu'on a besoin d'estimations de N_{ik} en sus d'estimations de η_{ik}. Nous avons élaboré notre modèle à l'aide de deux MMLG simultanés dans une structure hiérarchisée de Bayes. Le modèle que nous proposons a pour avantage d'être applicable à une grande diversité de problèmes. L'introduction d'une taille d'échantillon aléatoire et l'estimation des tailles de population représentent des techniques utiles dans bien des applications.

Bien sûr, nous pensons qu'il y a un rapport inverse entre les taux de succès de la chasse et le nombre d'expéditions dans un comté. Pour modéliser le phénomène, nous pouvons songer à un modèle ARC double de représentation des relations entre η_{ik} et N_{ik} comme le modèle de Kim, Sun et Tsutakawa (2001). Un modèle ARC à plus de deux variables est peut-être un autre moyen à notre disposition.

Pour chacun des modèles spatiaux, nous avons posé une corrélation commune ρ_k, à l'échelle de l'État. On pourrait juger plus raisonnable d'introduire des termes supplémentaires de corrélation dans différentes régions de productivité délimitées par les services de conservation du Missouri. Ce serait là un apport intéressant mais complexe au modèle. Il faut également dire que la structure spatiale que nous avons utilisée est semblable à celle de He et Sun (2000) et d'Oleson et He (2004), études où cette modélisation spatiale a donné de bons résultats.

Il serait peut-être bon de tenir compte de la distance entre le lieu de résidence et le lieu de chasse des chasseurs pour mieux estimer la « pression de chasse ». La plupart des chasseurs ne s'éloignent pas trop de leur lieu de résidence et une telle information pourrait être intégrée au cadre hiérarchisé.

À noter que l'estimation de la récolte est supérieure au dénombrement effectué aux postes de contrôle. Une explication partielle en est que les chasseurs qui réussissent mieux ont tendance à répondre au questionnaire postal. Nous faisons une recherche pour corriger le biais de non-réponse.

Remerciements

Cette étude a été partiellement soutenue par le projet W-13-R du programme d'aide au rétablissement de la faune du gouvernement fédéral américain. Pour ses travaux, Sun a reçu la subvention SES-0351523 de la National Science Foundation et la subvention R01-CA100760 de la NIH. Les auteurs aimeraient remercier de leurs précieuses observations le Dr Larry Vangilder et M. Jeff Beringer des services de conservation de l'État du Missouri. Ils remercient enfin vivement le rédacteur en chef, un rédacteur adjoint et deux examinateurs pour leurs commentaires constructifs sur les versions a priori de la présente étude.

Bibliographie

Agresti, A. (2002). *Categorical data analysis.* New York : John Wiley & Sons, Inc.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society*, *Methodological*, Séries B, 36, 192-236.

Carlin, B.P., et Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman & Hall Ltd.

Clayton, D., et Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43, 671-681.

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

Das, K., Jiang, J. et Rao, J.N.K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, 32, 818-840.

Durbin, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. Dans *New developments in survey sampling*, (Éds., N.L. Johnson et H.J. Smith). New York : Wiley-Interscience, 629-651.

Gelfand, A.E., et Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Gelman, A., Carlin, J., Stern, H. et Rubin, D. (1995). *Bayesian Data Analysis.* Chapman & Hall, London, U.K.

Ghosh, M., Natarajan, K, Stroud, T. et Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93, 273-282.

Gilks, W.R., Richardson, S. et Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice.* Chapman & Hall.

Gilks, W.R., et Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41, 337-348.

He, C.Z., et Sun, D. (1998). Hierarchical Bayes estimation of hunting success rates. *Environmental and Ecological Statistics*, 5, 223-236.

He, C.Z., et Sun, D. (2000). Hierarchical Bayes estimation of hunting success rates with spatial correlations. *Biometrics*, 56, 360-367.

Heidelberger, P., et Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31, 1109-1144.

Jiang, J., et Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, 15, 1-96.

Jiang, J., Lahiri, P. et Wan, S.-M. (2002). A unified jackknife theory for empirical best prediction with *M*-estimation. *The Annals of Statistics*, 30, 1782-1810.

Kim, H., Sun, D. et Tsutakawa, R.K. (2001). A bivariate Bayes method for improving the estimates of mortality rates with a twofold conditional autoregressive model. *Journal of the American Statistical Association*, 96, 1506-1521.

Kimmel, R.O. (2001). Regulating spring wild turkey hunting based on population and hunting quality. Dans *Proceeding of the National Wildlike Turkey Symposium*, 8, 243-250.

Lohr, S.L. (1999). *Sampling*: *Design and analysis.* Duxbury Press.

Malec, D., Sedransk, J., Moriarity, C.L. et LeClere, F.B. (1997). Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association*, 92, 815-826.

Oleson, J.J., et He, C.Z. (2004). Space-time modeling for the Missouri Turkey Hunting Survey. *Environmental and Ecological Statistics*, 11, 85-101.

Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Rao, J.N.K. (2003). *Small Area Estimation.* Wiley-Interscience.

Rao, J.N.K. (2005). Inferential issues in small area estimation: Some new developments. *Statistics in Transition*, 7, 513-526.

Robert, C.P., et Casella, G. (1999). *Monte Carlo Statistical Methods.* Springer-Verlag.

Smith, B.J. (2005). Bayesian Output Analysis program (BOA), version 1.1.5. http://www.public-health.uiowa.edu/boa.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. et Van Der Linde, A. (2002). Bayesian measures of model complexity and fit (Pkg: 583-639). *Journal of the Royal Statistical Society*, *Methodological*, Séries B, 64, 583-616.

Sun, D., Speckman, P.L. et Tsutakawa, R.K. (2000). Random effects in generalized linear mixed models (GLMMs). Dans *Generalized Linear Models*: *A Bayesian Perspective*, (Éds., D.K. Dey, S.K. Ghosh et B.K. Mallick). New York : Marcel Dekker, 23-39.

Vangilder, L.D., Sheriff, S.L. et Olsen, G.S. (1990). Characteristics, attitudes, and preferences of Missouri's spring turkey hunters. Dans *Proceeding of the National Wildlife Turkey Symposium*, 6, 167-176.

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of the Royal Statistical Society*, *Statistical Methodology*, Séries B, 62, 159-180.

Woodard, R., He, C.Z. et Sun, D. (2003). Bayesian estimation of hunting success rate and harvest for spatially correlated post-stratified data. *Biometrical Journal*, 45, 985-1005.

You, Y., et Rao, J. (2003). Pseudo hierarchical Bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference*, 111, 197-208.

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2007
Vol. 33, N° 2, pp. 209-220
Statistique Canada, N° 12-001-X au catalogue

Estimation pour petits domaines du revenu moyen des ménages en fonction des modèles au niveau des unités pour les données d'enquêtes par panel

Enrico Fabrizi, Maria Rosaria Ferrante et Silvia Pacei [1]

Résumé

Le panel communautaire de ménages (PCM) est une enquête par panel qui porte sur un large éventail de sujets concernant les conditions socio-économiques et les conditions de vie. Plus précisément, cette enquête permet de calculer le revenu équivalent disponible des ménages, qui constitue une variable clé de l'étude de l'inégalité économique et de la pauvreté. Pour obtenir des estimations fiables de la moyenne de cette variable pour des régions données de pays, il faut avoir recours aux méthodes d'estimation pour petits domaines. Dans le présent document, nous nous attardons sur les prédicteurs linéaires empiriques du revenu équivalent moyen en fonction de l'emprunt d'information des « modèles au niveau des unités », d'une région à l'autre et d'une période à l'autre. En nous appuyant sur une étude par simulation basée sur les données du PCM, nous comparons les estimateurs suggérés avec les estimateurs transversaux, fondés sur les modèles et fondés sur le plan de sondage. Dans le cas de ces prédicteurs empiriques, nous comparons également trois différents types d'estimateurs de l'EQM. Les résultats indiquent que les estimateurs qui sont rattachés aux modèles qui tiennent compte de l'autocorrélation des unités entrainent d'importants gains d'efficience, même en l'absence de covariables dont on connaît la moyenne de population.

Mots clés : Panel communautaire de ménages; revenu équivalent moyen; modèles linéaires mixtes; prédicteur linéaire sans biais empirique; estimation de l'EQM.

1. Introduction

Depuis quelques années, le milieu universitaire s'intéresse de plus en plus à l'analyse des disparités économiques régionales, qui entravent sérieusement la croissance économique nationale, et donc, la cohésion sociale. Ce phénomène est particulièrement présent au sein de l'Union européenne, où les disparités régionales sont un des traits distinctifs de l'économie. Ce regain d'intérêt envers les économies locales a attisé la demande en matière d'information statistique régionale et stimulé les recherches sur la répartition du revenu, la pauvreté et l'exclusion sociale au niveau infranational.

Dans les années 90, Eurostat (l'organisme statistique de l'UE) a lancé le Panel communautaire de ménages (PCM), une enquête annuelle par panel menée auprès des ménages européens au moyen de méthodes normalisées dans tous les pays membres de l'UE (Betti et Verma 2002; Eurostat 2002). Le PCM a pris fin en 2001, après huit cycles. À l'heure actuelle, on est en train de le remplacer par l'Enquête européenne sur le revenu et les conditions de vie (EU-SILC), qui ressemble au PCM à bien des égards, mais dont on n'a pas encore publié les données. Le PCM portait sur un large éventail de sujets et permettait entre autres de calculer le revenu équivalent disponible des ménages, qui constitue une variable clé de l'étude de l'équité économique et de la pauvreté.

Le PCM avait pour objet de fournir des estimations fiables pour les grandes régions des pays de la NUTS1 (NUTS renvoie à la « Nomenclature des unités territoriales statistiques », qui est définie en fonction de certains principes décrits sur le site Web d'EUROSTAT au http://europa.eu.int/comm/Eurostat/ramon/Noix/home regions en.html). Malheureusement, la NUTS1 renvoie à des régions (cinq groupes de régions administratives dans le cas de l'Italie) qui sont trop vastes pour permettre la mesure efficace de la disparité du revenu des régions ou pour fournir des renseignements utiles aux fins de la gouvernance régionale. Par conséquent, pour obtenir des estimations à un niveau géographique plus précis, on a employé une méthode d'estimation pour petits domaines. Le problème réside dans le choix d'une méthode appropriée et efficace.

Dans le présent document, afin d'amalgamer les renseignements provenant des enquêtes précédentes, les variables auxiliaires connexes et les petites régions, nous envisageons plusieurs extensions possibles du modèle de régression à erreurs emboîtées au niveau des unités, qui est bien connu (voir Battese, Harter et Fuller 1988), pour estimer le revenu équivalent moyen des ménages. En nous appuyant sur les données du PCM, nous illustrons comment les modèles proposés pourraient s'avérer utiles et améliorer l'efficience des estimations pour petits domaines en misant sur la corrélation du revenu des ménages individuels au fil du temps.

1. Enrico Fabrizi, DMSIA, University of Bergamo, via dei Caniana 2, 24127, Bergamo, Italy. Courriel : enrico.fabrizi@unibg.it; M.R. Ferrante, Department of Statistics, University of Bologna, via Belle Arti 41, 40126, Bologna, Italy. Courriel : ferrante@stat.unibo.it; S. Pacei, Department of Statistics, University of Bologna, via Belle Arti 41, 40126, Bologna, Italy. Courriel : pacei@stat.unibo.it.

À la section 2, nous présentons une organisation générale de l'estimation pour petits domaines au moyen des données de l'enquête par panel et examinons brièvement les méthodes d'estimations pour petits domaines fondées sur le plan de sondage et sur le modèle. Dans cette section, nous élaborons les meilleurs prédicteurs linéaires sans biais empiriques (MPLSBE) et leurs estimateurs d'erreur quadratique moyenne (EQM) pour certains modèles transversaux et chronologiques au niveau des unités, grâce à la théorie disponible sur les MPLSBE en matière d'estimation pour petits domaines (pour plus de détails, voir Rao 2003, et Jiang et Lahiri 2006a). Soulignons que les modèles transversaux et chronologiques ont été envisagés dans la documentation sur les petits domaines, mais uniquement dans le contexte de la modélisation au niveau de la région (voir Rao et Yu 1994; Ghosh, Nangia et Kim 1996; Datta, Lahiri, Maiti et Lu 1999; Datta, Lahiri et Maiti 2002; Pfeffermann 2002, entre autres).

À la section 3, nous passons brièvement en revue le PCM et décrivons comment nous nous servons des données de cette enquête pour effectuer une étude par simulation Monte Carlo, en vue de comparer différents estimateurs pour petits domaines et leurs estimateurs d'EQM. À la section 4, nous faisons état des résultats de la simulation Monte Carlo. Nous mentionnons que la simulation avait pour objet d'évaluer les propriétés fondées sur le plan de sondage de tous les estimateurs, même s'ils sont dérivés en tant que prédicteurs fondés sur le modèle. Nous constatons que les MPLSBE donnent de très bons résultats comparativement aux estimateurs fondés sur le plan de sondage, même si notre pseudo-population comporte des signes de non-normalité. Toutefois, l'effet de la non-normalité de la pseudo-population semble avoir un effet sur à l'efficience des estimateurs de l'EQM. Dans notre simulation, les estimateurs de l'EQM par la série de Taylor (voir Prasad et Rao 1990; Datta et Lahiri 1999, entre autres) et par la méthode bootstrap paramétrique (voir Butar et Lahiri 2003) se révèlent plus vulnérables à la non-normalité que la méthode jackknife de Jiang, Lahiri et Wan (2002). Nous concluons le document avec quelques observations finales.

2. Les méthodes d'estimation sur petits domaines considérées

Pour décrire les données de l'échantillon, y_{dti} indique la valeur d'une variable de l'étude pour l'unité i appartenant à la région d pendant la période $t\,(d = 1,...,m;\ t = 1,...,T;\ i = 1,...,n_{dt})$. De plus, $\mathbf{x}_{dti}$ correspond aux valeurs du vecteur des covariables associées à chaque y_{dti} (et dont le premier élément est égal à 1), et $\mathbf{X}_s = \{\mathbf{x}'_{dti}\}$ représente la matrice $n \times p$ des valeurs des covariables pour l'échantillon au complet $(n = \sum_{d,t} n_{dt})$. Supposons que nous voulons prédire les moyennes régionales de la variable ciblée pendant la période finale $T : \bar{Y}_{dT}, (d = 1, ..., m)$. Supposons également que les vecteurs des valeurs de la population moyenne des covariables sont connus pour la période T; nous désignons ces vecteurs comme suit : $\bar{\mathbf{X}}'_{dT}$ $(d = 1, ..., m)$.

2.1 Estimateurs fondés sur le plan de sondage

Une première solution au problème de l'estimation pour petits domaines consiste à utiliser des estimateurs directs, c'est-à-dire des estimateurs utilisant seulement les valeurs y obtenues de la région (et de la période) auxquelles se rapporte le paramètre. L'estimateur direct le plus simple de la moyenne de population est la moyenne pondérée. Nous désignons cet estimateur direct comme suit : $\bar{y}_{dT,\,\mathrm{DIR}}$ $(d = 1, ..., m)$, et nous nous en servirons comme point de repère dans les sections qui suivent.

On peut généralement définir les estimateurs synthétiques comme des estimateurs sans biais pour une plus grande région comportant des erreurs-types acceptables. Ces estimateurs servent à calculer les estimations pour petits domaines, en supposant que les petits domaines ont les mêmes caractéristiques que les plus grands. En outre, lorsqu'il existe de l'information au sujet des variables auxiliaires, on peut obtenir un estimateur synthétique en particulier (l'estimateur de régression) en appliquant un modèle de régression à toutes les données de l'échantillon. Soulignons que l'estimateur synthétique se rapporte à une région par rapport aux variables auxiliaires et non par rapport à la variable à l'étude.

Par exemple, si nous considérons uniquement les observations du dernier cycle $(t = T)$, le modèle de régression simple serait formulé comme suit :

$$y_{dTi} = \mathbf{x}'_{dTi}\,\beta + e_{dTi}$$

$$E(e_{dTi}) = 0, \quad E(e^2_{dTi}) = \pi^2.$$

Pour tenir compte de la complexité du plan d'échantillonnage, on peut obtenir l'estimation pondérée par les moindres carrés $\tilde{\beta}_w$ de β, ce qui signifie que l'estimateur de régression synthétique sera calculé comme suit :

$$\bar{y}_{dT,\,\mathrm{RSYN}} = \bar{\mathbf{X}}'_{dT}\,\tilde{\beta}_w \quad d = 1, \ldots, m. \qquad (1)$$

En général, les estimateurs synthétiques ont des variances très faibles, mais ils peuvent être entachés de forts biais lorsque le modèle qui s'applique à l'échantillon en entier n'est pas bien adapté aux données régionales. Les estimateurs composites sont des moyennes pondérées des estimateurs direct et synthétique. Soit la formule suivante de l'estimateur composite :

$$\bar{y}_{dT,\,\mathrm{COMP}} = \phi_{dT}\,\bar{y}_{dT,\,\mathrm{DIR}} + (1 - \phi_{dT})\,\bar{y}_{dT,\,\mathrm{RSYN}}, \qquad (2)$$

où

$$\phi_{dT} = \frac{\mathrm{EQM}_D(\bar{y}_{dT,\,\mathrm{RSYN}})}{\mathrm{EQM}_D(\bar{y}_{dT,\,\mathrm{DIR}}) + \mathrm{EQM}_D(\bar{y}_{dT,\,\mathrm{RSYN}})}$$

et EQM_D signifie que l'erreur quadratique moyenne (EQM) est évaluée par rapport à la distribution de la randomisation. La sélection de ϕ_{dT} donne lieu aux estimateurs composites $\bar{y}_{dT,\,\mathrm{COMP}}$ qui sont à peu près optimaux à l'égard de l'EQM_D (voir Rao 2003, section 4.3). En pratique, les quantités de la formule de ϕ_{dT} sont inconnues et peuvent être estimées à partir des données. On peut obtenir des estimateurs sans biais et convergents pour l'$\mathrm{EQM}_D(\bar{y}_{dT,\,\mathrm{DIR}}) = V_D(\bar{y}_{dT,\,\mathrm{DIR}})$ en utilisant des formules standard. Un estimateur approximatif sans biais fondé sur le plan de sondage de l'$\mathrm{EQM}_D(\bar{y}_{dT,\,\mathrm{RSYN}})$ peut être obtenu au moyen des formules proposées dans le livre de Rao (2003, section 4.2.4). En particulier, nous calculons l'approximation comme suit :

$$\mathrm{eqm}_D(\bar{y}_{dT,\,\mathrm{RSYN}}) \approx (\bar{y}_{dT,\,\mathrm{RSYN}} - \bar{y}_{dT,\,\mathrm{DIR}})^2 - v_D(\bar{y}_{dT,\,\mathrm{DIR}}),$$

où eqm_D et v_D représentent les estimateurs des valeurs EQM_D et V_D correspondantes. Plus précisément, v_D est l'estimateur sans biais selon le plan de sondage ordinaire de V_D lorsque nous calculons sa moyenne par rapport à d, comme d'habitude, pour obtenir un estimateur plus stable. En fait, l'un des problèmes associés à mse_D est qu'il peut même être négatif.

De plus, afin d'améliorer la fiabilité de l'estimateur, on peut avoir recours à un estimateur direct modifié empruntant de l'information d'une région à l'autre pour l'estimation du coefficient de régression. Lorsque de l'information auxiliaire est disponible, l'estimateur de régression généralisée (GREG),

$$\bar{y}_{dT,\,\mathrm{GREG}} = \bar{\mathbf{X}}'_{dT}\tilde{\beta}_w + \frac{\sum_{j\in s_{dT}} w_j e_j}{\sum_{j\in s_{dT}} w_j}, \qquad (3)$$

corrige approximativement le biais de l'estimateur synthétique au moyen du terme $(\sum_{j\in s_{dT}} w_j)^{-1}\sum_{j\in s_{dT}} w_j e_j$, en fonction des résidus de régression e_j.

2.2 Les estimateurs fondés sur le modèle

Les estimateurs fondés sur le modèle que nous avons envisagés sont axés sur la spécification de modèles explicites pour les données de l'échantillon, qui sont simulées par un processus hypothétique de génération des données. Par conséquent, le problème associé à l'estimation de $\bar{Y}_{dT}$ se résume à un problème de prédiction. Par ailleurs, les erreurs quadratiques moyennes et les autres propriétés statistiques des estimateurs sont habituellement évaluées par rapport au processus de génération des données. Nous nous sommes concentrés sur le modèle « au niveau des unités » en fonction de modèles rattachant y_{dti} à un vecteur de covariables $\mathbf{x}_{dti}$. L'utilisation de modèles explicites comporte plusieurs avantages, dont le plus important est la possibilité de mettre à l'essai les hypothèses sous-jacentes.

Les modèles linéaires mixtes sont très souvent utilisés pour estimer les moyennes ou les totaux des variables continues sur des petits domaines. Un modèle linéaire mixte général peut se décrire comme suit :

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_1 v_1 + \ldots + \mathbf{Z}_s v_s + \mathbf{e}, \qquad (4)$$

où $\mathbf{y} = \{y_{dti}\}$ est le vecteur de dimension n des observations des échantillons, β est un vecteur $p\times 1$ des effets fixes, v_j est un vecteur $q_j \times 1$ des effets aléatoires $(j = 1, \ldots, s)$, $\mathbf{e} = \{e_{dti}\}$ est un vecteur des erreurs; nous présumons que $\mathbf{X}$ est de rang $p, \mathbf{Z}_i = \{\mathbf{z}'_{dti}\}$ est une matrice $n \times q_j$ de l'incidence de l'effet aléatoire j^{e}. Nous supposons que $E(v_j) = 0, V(v_j) = \mathbf{G}_j, E(\mathbf{e}) = 0, V(\mathbf{e}) = \mathbf{R}$ (toutes les espérances sont par rapport au modèle (4)) et que $v_1, \ldots, v_s, \mathbf{e}$ sont mutuellement indépendants.

Ainsi, la matrice de variance-covariance $\mathbf{y}$ se calcule comme suit :

$$\mathbf{V} = V(\mathbf{y}) = \sum_{j=1}^{s} \mathbf{Z}_j \mathbf{G}_j \mathbf{Z}'_j + \mathbf{R} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R},$$

où $\mathbf{Z} = [\mathbf{Z}_1 \,|\, \ldots \,|\, \mathbf{Z}_s]$. On présume généralement que les matrices $\mathbf{G}$, $\mathbf{R}$ dépendent d'un vecteur $\mathbf{k}$ des composantes de variance ψ, ce qui nous permet d'écrire : $\mathbf{V}(\psi) = \mathbf{Z}\mathbf{G}(\psi)\mathbf{Z}' + \mathbf{R}(\psi)$.

Soulignons qu'au niveau des observations individuelles, le modèle (4) peut être révisé comme suit : $y_{dti} = \mathbf{x}'_{dti}\beta + \mathbf{z}'_{1dti}v_1 + \ldots + \mathbf{z}'_{sdti}v_s + e_{dti}$.

Nous tenons compte de diverses spécifications pour les modèles linéaires mixtes, qui peuvent aussi être considérés comme des cas spéciaux du modèle général (4). Par souci de simplicité, nous avons adopté la notation au niveau des unités pour décrire les modèles à l'étude. Le premier modèle :

$$MM1: \quad y_{dti} = \mathbf{x}'_{dti}\beta + v_d + \alpha_t + e_{dti}, \qquad (5)$$

peut être obtenu à partir de la formule (4) en écrivant $s = 2, q_1 = m, q_2 = T$ $\mathbf{G}_1 = \sigma_v^2\mathbf{I}_m$, $\mathbf{G}_2 = \sigma_\alpha^2\mathbf{I}_T$, $\mathbf{R} = \sigma_e^2\mathbf{I}_n$. Ce modèle tient compte des effets régionaux et temporels indépendants les uns des autres, ce qui signifie qu'on présume que les effets régionaux n'évoluent pas au fil du temps. Cette structure d'effets aléatoires correspond à l'hypothèse d'une covariance constante entre les unités d'une même région, observée à deux moments donnés.

Le deuxième modèle :

$$MM2: \quad y_{dti} = \mathbf{x}'_{dti}\beta + \delta_{dt} + e_{dti}, \qquad (6)$$

correspond au cas particulier où $s = 1, q_1 = mq$, $\mathbf{G}_1 = \sigma_\delta^2\mathbf{I}_{mq}, \mathbf{R} = \sigma_e^2\mathbf{I}_n$. Nous tenons compte des effets de l'interaction entre les régions et les périodes, c'est-à-dire que certains effets régionaux ne sont pas constants au fil du temps.

Le troisième modèle :

$$MM3: \quad y_{dti} = \mathbf{x}'_{dti}\beta + v_d + \alpha^*_t + e_{dti}, \tag{7}$$

est obtenu en supposant que $s=2, q_1=m, q_2=T$ $\mathbf{G}_1=\sigma_v^2\mathbf{I}_m$, $\mathbf{R}=\sigma_e^2\mathbf{I}_n$, tandis que l'élément générique $g_2(h,k)$ de $\mathbf{G}_2$ est $g_2(h,k)=\sigma_\alpha^2\rho_\alpha^{|h-k|}, h,k=1,...,T$. Certains effets régionaux et temporels sont indépendants les uns des autres, tout comme c'est le cas du modèle *MM*1, mais nous présumons que les effets temporels suivent un processus AR(1).

Le quatrième modèle :

$$MM4: \quad y_{dti} = \mathbf{x}'_{dti}\beta + \delta^*_{dt} + e_{dti}, \tag{8}$$

est semblable au modèle *MM*2 en ce qu'il est caractérisé par des effets régionaux variant au fil du temps, mais nous présumons en outre que ces effets suivent un processus AR(1). Donc, à condition d'ordonner les observations par région, en ce qui concerne la formule générale (4), nous obtenons $s=1, q_1=mq, \mathbf{G}_1=\mathrm{diag}(\mathbf{G}_{1d}), \mathbf{R}=\sigma_e^2\mathbf{I}_n$ où $\mathbf{G}_{1d}, \quad d=1,...,m,$ est une matrice $T\times T$ dont l'élément générique $g_{1r}(h,k)=\sigma_\delta^2\rho_\delta^{|h-k|}, h,k=1,...,T$.

La dernière spécification :

$$MM5: \quad y_{dti} = \mathbf{x}'_{dti}\beta + v_d + \alpha_t + e^*_{dti}, \tag{9}$$

peut être obtenue au moyen de (4) avec $s=2, q_1=m,$ $q_2=T, \mathbf{G}_1=\sigma_v^2\mathbf{I}_m, \mathbf{G}_2=\sigma_\alpha^2\mathbf{I}_T$. À condition d'ordonner les observations selon le ménage et la période $\mathbf{R}=\mathrm{diag}(\mathbf{R}_{di})$ où $\mathbf{R}_{di}$ est une matrice $T\times T$ dont l'élément générique est donné par $r_{di}(h,k)=\sigma_e^2\rho_e^{|h-k|}, h,k=1,...,T$. Comme dans le modèle *MM*1, il existe des effets régionaux et temporels indépendants les uns des autres, mais nous supposons que les erreurs sont autocorrélées selon un processus AR(1).

Pour évaluer les conséquences de l'utilisation des cycles précédents de l'enquête sur l'efficacité de l'estimateur, le modèle linéaire transversal mixte (*SMM*) fondé exclusivement sur les données du dernier cycle *T* a servi de point de repère :

$$SMM: \quad y_{dTi} = \mathbf{x}'_{dTi}\beta + \vartheta_d + e_{dTi} \tag{10}$$

où $\vartheta_d \overset{\text{IID}}{\sim} N(0,\sigma_\vartheta^2), \quad e_{jdT} \overset{\text{IID}}{\sim} N(0,\sigma_e^2)$.

Encore une fois, il s'agit d'un cas particulier de (4) obtenue pour $s=2, q_1=m, \mathbf{G}_1=\sigma_v^2\mathbf{I}_m$ et $\mathbf{R}=\sigma_e^2\mathbf{I}_n$. Soulignons que (10) est le modèle standard de régression à erreurs emboîtées de Battese et coll. (1988).

Nous tenons également compte des modèles linéaires de variance de l'erreur aléatoire (voir Rao 2003; section 5.5.2) obtenus en remplaçant $\mathbf{x}'_{dti}\beta$ dans les formules (5) - (10) par un point d'ordonnée à l'origine général de θ. Ces modèles seront désignés comme suit : *MM*1*, *MM*2*, *MM*3*, *MM*4*, *MM*5*, *SMM**. Toutes les hypothèses au sujet des effets aléatoires et des résidus demeurent inchangées. Ce dernier groupe de modèles nous permet d'explorer les gains d'efficience découlant de la mise à profit de la répétition de l'observation de la même unité lorsqu'il n'existe pas de covariables au niveau de population.

En ce qui concerne l'estimation pour petits domaines, l'objectif est de prédire les combinaisons scalaires linéaires des effets fixes et aléatoires du type $\eta=\mathbf{m}'\beta+\mathbf{k}'v$ où **m** et **k** renvoient respectivement aux vecteurs $p\times 1$ et $q\times 1$, et où $q=\sum_j q_j$. Nous obtenons le meilleur prédicteur linéaire sans biais (MPLSB) de η en estimant (β, v) pour réduire l'EQM du modèle dans tous les estimateurs linéaires :

$$\tilde{\eta}^{\text{MPLSB}}(\psi) = \mathbf{m}'\tilde{\beta}(\psi) + \mathbf{k}'\tilde{v}(\psi). \tag{11}$$

Lorsque les composantes de la variance de ψ sont inconnues, nous pouvons les estimer à partir des données et les substituer dans la formule (11), en vue d'obtenir les « MPLSB empiriques » (ou MPLSBE) $\tilde{\eta}^{\text{MPLSBE}}(\hat{\psi}) = \mathbf{m}'\hat{\beta}(\hat{\psi}) + \mathbf{k}'\hat{v}(\hat{\psi})$ (pour plus de détails, voir Rao 2003, chapitre 6, et Jiang et Lahiri 2006b).

En ce qui a trait à l'estimation de ψ, plusieurs méthodes ont été proposées dans la documentation, telles que le maximum de vraisemblance (MV) et le maximum de vraisemblance restreint (MVR), qui présument de la normalité des termes aléatoires, et la méthode MINQUE proposée par Rao (1972), qui est non paramétrique. Pour les besoins du présent document, nous avons opté pour la méthode MVR, ce qui signifie donc que nous avons supposé la normalité.

2.3 Les mesures de l'incertitude associée aux prédicteurs fondés sur les modèles linéaires mixtes

Dans la documentation pour petits domaines, diverses approches sont proposées pour résoudre la problématique de l'estimation de l'EQM des MPLSBE, en tenant compte de la variabilité des composantes estimées de la variance et de la covariance.

Une méthode populaire est fondée sur une approximation de l'EQM par la série de Taylor sous l'hypothèse de la normalité (Prasad et Rao 1990; Datta et Lahiri 1999). Plus récemment, grâce aux ordinateurs à haute vitesse et aux logiciels puissants, des méthodes de rééchantillonnage ont été proposées. Par exemple, Butar et Lahiri (2003) présentent une méthode « bootstrap » paramétrique qui est fondée sur l'hypothèse de normalité, mais moins lourde du point de vue de l'analyse que la méthode de la série de Taylor. Jiang et coll. (2002) décrivent une méthode jackknife générale, qui nécessite une hypothèse de répartition plus faible que la normalité (linéarité postérieure). Notre objectif est d'effectuer une comparaison

empirique du rendement de ces trois estimateurs dans un contexte où le nombre de régions est modéré et l'hypothèse de normalité pourrait ne pas s'avérer tout à fait adéquate. Voici une brève description des trois méthodes d'estimation.

Soit la formule $\mathrm{EQM}[\tilde{\eta}^{\mathrm{MPLSBE}}(\hat{\psi})]=E(\tilde{\eta}^{\mathrm{MPLSBE}}(\hat{\psi})-\eta)^2$, où l'espérance renvoie au modèle (4). Il est possible de démontrer que, dans l'optique de la normalité,

$$\begin{aligned}&\mathrm{EQM}[\tilde{\eta}^{\mathrm{MPLSBE}}(\hat{\psi})]\\&\quad = g_1(\psi)+g_2(\psi)+E(\tilde{\eta}^{\mathrm{MPLSBE}}-\tilde{\eta}^{\mathrm{MPLSB}})^2 \qquad (12)\end{aligned}$$

où $g_1(\psi)=\mathbf{k}'(\mathbf{G}-\mathbf{GZ}'\mathbf{V}^{-1}\mathbf{ZG})\mathbf{k}'$ et $g_2(\psi)=\mathbf{d}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{d}$, et $\mathbf{d}=\mathbf{m}'-\mathbf{k}'\mathbf{GZ}'\mathbf{V}^{-1}\mathbf{X}$ (voir Rao 2003, chapitre 3). Au moyen de l'approximation suivante, fondée sur un argument de la série de Taylor

$$\begin{aligned}&E(\tilde{\eta}^{\mathrm{MPLSBE}}-\tilde{\eta}^{\mathrm{MPLSBE}})^2\\&\quad\approx \mathrm{tr}[(\partial\mathbf{b}'/\partial\psi)\mathbf{V}(\partial\mathbf{b}'/\partial\psi)'\bar{\mathbf{V}}(\psi)]=g_3(\psi)\end{aligned}$$

où $\mathbf{b}'=\mathbf{k}'\mathbf{GZ}'\mathbf{V}^{-1}$, on peut faire une approximation de deuxième ordre de (12) :

$$\mathrm{EQM}[\tilde{\eta}^{\mathrm{MPLSBE}}(\hat{\psi})]\approx g_1(\psi)+g_2(\psi)+g_3(\psi). \qquad (13)$$

Soulignons qu'ici, $\approx$ signifie que les termes omis sont de l'ordre de $o(m^{-1})$. On obtient un estimateur asymptotiquement sans biais de (13), selon Prasad et Rao (1990), grâce à la formule suivante :

$$\mathrm{eqm}_{\mathrm{PR}}(\tilde{\eta}^{\mathrm{MPLSBE}})=g_1(\hat{\psi})+g_2(\hat{\psi})+2g_3(\hat{\psi}). \qquad (14)$$

Datta et Lahiri (1999) démontrent que, en supposant la normalité avec l'estimation de ψ, au moyen des méthodes MVR ou MV, $\mathrm{eqm}_{\mathrm{PR}}(\tilde{\eta}^{\mathrm{MPLSBE}})$ estime $\mathrm{EQM}[\tilde{\eta}^{\mathrm{MPLSBE}}(\hat{\psi})]$ avec un biais de l'ordre $o(m^{-1})$.

Butar et Lahiri (2003) proposent une estimation paramétrique bootstrap selon le modèle (13), en s'appuyant sur une hypothèse de normalité. Nous adaptons leur estimateur aux modèles que nous analysons, en présumant de l'utilisation du modèle bootstrap suivant :

$$\begin{aligned}&\text{i)}\quad \mathbf{y}^*\mid\mathbf{v}^*\overset{\mathrm{ind}}{\sim}N[\mathbf{X}\hat{\beta}+\mathbf{Z}\mathbf{v}^*,\mathbf{R}(\hat{\psi})]\\&\text{ii)}\quad \mathbf{v}^*\overset{\mathrm{ind}}{\sim}N[\mathbf{0},\mathbf{G}(\hat{\psi})]\end{aligned} \qquad (15)$$

où $\mathbf{v}=(v_1',\ldots v_s')'$. La méthode du bootstrap paramétrique est ensuite utilisée à deux reprises, une fois pour estimer les deux premiers termes de (13), ce qui permet de corriger le biais de $g_1(\hat{\psi})+g_2(\hat{\psi})$, et une fois pour estimer $g_3(\psi)$.

L'estimateur de (13) suivant est proposé :

$$\begin{aligned}&\mathrm{eqm}_{\mathrm{BL}}(\tilde{\eta}^{\mathrm{MPLSBE}})\\&\quad=2[g_1(\hat{\psi})+g_2(\hat{\psi})]-E_B[g_1(\hat{\psi}^*)+g_2(\hat{\psi}^*)]\\&\qquad+E_B[\tilde{\eta}(\mathbf{y},\hat{\beta}(\hat{\psi}^*),\hat{\psi}^*)-\tilde{\eta}(\mathbf{y},\hat{\beta}(\hat{\psi}),\hat{\psi})]\end{aligned} \qquad (16)$$

où $\hat{\psi}^*$ est identique à $\hat{\psi}$ sauf qu'il est calculé par rapport à $\mathbf{y}^*$ plutôt que $\mathbf{y}$, et E_B est l'espérance par rapport au modèle bootstrap (15).

L'estimateur bootstrap (16) ne nécessite pas la dérivation analytique de $g_3(\hat{\psi})$, qui peut s'avérer fort laborieuse lorsque $\mathbf{G}$ et $\mathbf{R}$ ont des structures complexes.

Jiang et coll. (2002) ont présenté un estimateur jackknife général de la variance des prédicteurs empiriques des modèles linéaires et non linéaires mixtes, avec une estimation *M*. Dans l'optique de la problématique qui nous occupe, l'estimateur qu'ils proposent peut s'énoncer comme suit :

$$\begin{aligned}\mathrm{eqm}_{\mathrm{JLW}}(\tilde{\eta}^{\mathrm{MPLSBE}})&=g_1(\hat{\psi})-\frac{m-1}{m}\sum_{j=1}^{m}[g_1(\hat{\psi}_{-j})-g_1(\hat{\psi})]\\&\quad+\frac{m-1}{m}\sum_{j=1}^{m}(\tilde{\eta}_{-j}^{\mathrm{MPLSBE}}-\tilde{\eta}^{\mathrm{MPLSBE}})^2 \qquad (17)\end{aligned}$$

où $\hat{\psi}_{-j}$ est l'estimation de ψ calculée au moyen de toutes les données sauf celles de la région j. Parallèlement, $\tilde{\eta}_{-j}^{\mathrm{MPLSBE}}=\tilde{\eta}^{\mathrm{MPLSBE}}(\mathbf{y}_{-j},\hat{\beta}(\hat{\psi}_{-j}),\hat{\psi}_{-j})$.

Il est important de souligner que, si l'on se fie aux résultats de la simulation signalés dans Jiang et coll. (2002), l'$\mathrm{eqm}_{\mathrm{JLW}}$ est considéré plus robuste que l'$\mathrm{eqm}_{\mathrm{PR}}$ en ce qui a trait aux écarts par rapport à l'hypothèse de normalité, qui risque également d'être essentielle à l'estimateur $\mathrm{eqm}_{\mathrm{BL}}$.

3. L'étude par simulation fondée sur les données du panel communautaire de ménages

La population cible du PCM est composée de tous les ménages d'un vaste sous-ensemble de pays membres de l'UE. Bien que les lignes directrices générales de l'enquête aient été instaurées par Eurostat, une certaine souplesse a été permise, de sorte que le plan d'échantillonnage comporte des différences d'un pays à l'autre. Pour ce qui est de l'Italie, l'enquête est fondée sur un plan stratifié à deux degré, selon lequel les strates ont été formées par le regroupement des UPE (municipalités) en fonction de la région géographique (NUTS2) et de la taille de la population. Pour plus de détails sur l'enquête, voir Eurostat (2002).

Le PCM traite la non-réponse des unités, l'érosion de l'échantillon et les nouvelles entrées au moyen de la pondération et de l'imputation. Étant donné que l'érosion pourrait donner lieu à des estimations biaisées du revenu si elle n'apparaît pas de façon aléatoire, l'effet de la pauvreté sur la propension au décrochage a été étudié (Rendtel, Behr et Sisto 2003; Vandecasteele et Debels 2004), et les résultats de ces études révèlent que dans certains pays, notamment en Italie, cet effet disparaît sous le contrôle des variables de pondération.

Nous avons dirigé notre attention vers les huit cycles du PCM disponibles pour l'Italie (1994-2001). Étant donné que notre objectif est de déterminer si l'utilisation de plusieurs observations successives du même ménage pourrait être bénéfique aux fins de l'estimation pour petits domaines, nous n'avons pas abordé le problème d'érosion et avons seulement tenu compte des ménages qui participent à tous les cycles de l'enquête.

La variable qui nous intéresse est le revenu disponible du ménage après impôt au moment du dernier cycle (2001). Dans les études sur la pauvreté et l'inégalité, on a souvent recours au concept de « revenu équivalent », établi en fonction d'une échelle d'équivalences pour éviter les problèmes de comparaison attribuables aux différences relatives à la composition des ménages. Nous tenons compte de l'échelle modifiée de l'OCDE, qui est largement utilisée et qui a également été adoptée par Eurostat (2002) dans ses publications sur le revenu, la pauvreté et l'exclusion sociale. Selon cette échelle, on obtient le revenu équivalent en divisant le revenu du ménage disponible par le nombre (k) « d'adultes équivalents », définis comme $k = 1 + 0,5a + 0,3c$, où a est le nombre d'adultes en plus du « chef du ménage » et c correspond au nombre d'enfants de 13 ans et moins. En général, le revenu équivalent peut être perçu comme la quantité de revenu dont devrait disposer une personne vivant seule afin d'atteindre le même niveau de bien-être économique dont elle jouit au sein de son ménage.

Parmi les nombreuses covariables du volumineux questionnaire du PCM, nous avons sélectionné seulement celles dont les moyennes régionales étaient disponibles par l'entremise des résultats du recensement italien de 2001. Par conséquent, les covariables choisies sont les suivantes : le pourcentage d'adultes; le pourcentage de personnes occupées; le pourcentage de chômeurs; le pourcentage de membres du ménage ayant un niveau de scolarité élevé/moyen/faible; la typologie des ménages (présence d'enfants, présence de personnes âgées, *etc.*); le nombre de pièces par habitant et le titre d'occupation du logement (en logement, en propriété, *etc.*).

Comme nous l'avons mentionné, le présent document a pour objectif de comparer le rendement de différents estimateurs dans l'environnement contrôlé d'un exercice de simulation. Plusieurs ouvrages de la documentation existante ont comparé les estimateurs pour petits domaines au moyen de la méthode de Monte Carlo, où les échantillons proviennent des populations synthétiques fondées soit sur les données des recensements (Falorsi, Falorsi et Russo 1994; Ghosh et coll. 1996), soit sur la répétition des enregistrements des unités d'échantillonnage (Falorsi, Falorsi et Russo 1999; Lehtonen, Särndal et Veijanen 2003; Singh, Mantel et Thomas 1994). Étant donné que le revenu du ménage n'est pas mesuré par le recensement italien (et qu'il n'est pas non plus obtenu des résultats des autres recensements effectués par les pays de l'UE), nous avons considéré les données du PCM comme la pseudo-population, avant d'établir les échantillons au moyen d'une méthode d'échantillonnage stratifié avec probabilité proportionnelle à la taille, la variable de la taille étant obtenue à partir des poids de l'enquête. Cette solution n'est peut-être pas aussi bonne que celle qui consiste à utiliser les données de la population réelle d'un recensement, mais nous espérons qu'elle est tout de même plus réaliste que l'extrapolation de valeurs démographiques du revenu des ménages à partir d'un modèle paramétrique.

Les échantillons de 1 000 unités de Monte Carlo (à peu près 15 % de la taille réelle de l'échantillon du PCM) ont été établis en fonction de la population synthétique par échantillonnage aléatoire stratifié sans remplacement, les strates étant données par les 21 régions de la NUTS2. Par conséquent, ces régions sont traitées comme des domaines planifiés (comme c'est le cas dans le PCM) dont la taille de l'échantillon des petits domaines est établie à l'avance, de manière à ce que les fractions de l'échantillonnage reflètent le suréchantillonnage des petits domaines exactement de la même façon que le plan d'échantillonnage réel du PCM. Les échantillons régionaux que nous avons obtenus comportent de 14 à 112 unités, la taille moyenne s'établissant à 48. Par conséquent, dans notre simulation $n = 1\,000$, le nombre de petits domaines correspond à celui des régions italiennes ($m = 21$) et le nombre de moments ponctuels correspond aux cycles disponibles du PCM ($T = 8$).

La répartition du revenu du ménage équivalent de notre pseudo-population (c'est-à-dire la répartition obtenue au moyen de l'estimation pondérée des données de l'échantillon du PCM) est caractérisée par une moyenne globale de 22 547 euros et un coefficient de variation de 0,59. La répartition est désaxée vers la droite (même si l'asymétrie n'est pas extrême; coefficient d'asymétrie : $\gamma_1 = \mu_3 / \sigma^3 \cong 2,5$) et d'aplatissement ($\kappa = \mu_4 / \sigma^4 \cong 14,3$). La différence entre la moyenne et la médiane représente 9 % de la moyenne. Caractéristique intéressante, on note des écarts importants d'une région administrative à l'autre (ces régions représentent les petits domaines qui nous intéressent dans le cadre de notre étude). La moyenne du revenu du ménage équivalent oscille de 16 604 à 27 011, ce qui signifie que la région la plus riche affiche un revenu équivalent moyen de 62 % supérieur à celui de la région la plus pauvre. Par ailleurs, le coefficient de variation (de 0,28 à 0,84), l'asymétrie (γ_1 de 0,1 à 4,6) et l'aplatissement (κ de -0,7 à 32,9) démontrent que la distribution de la variable qui nous intéresse fluctue largement d'une région à l'autre.

Pour justifier les spécifications sélectionnées de la portion des effets aléatoires des modèles linéaires mixtes

envisagés (voir la section 2.2), les manuels académiques (voir Verbeke et Molenberghs 2000, chapitre 9) recommandent souvent la méthode suivante, qui a été suivie dans ce cas-ci : d'abord, nous appliquons à nos données la méthode standard de régression des moindres carrés ordinaires en utilisant toutes les covariables disponibles; ensuite, nous analysons les résidus obtenus pour nous aider à repérer les effets aléatoires. Cette analyse préliminaire a été effectuée séparément sur plusieurs échantillons aléatoires de 1000 unités, établis conformément au plan de rééchantillonnage susmentionné.

Le R^2 rajusté de la régression des moindres carrés ordinaires se rapproche de 0,35 dans chacun des échantillons observés. Ce résultat plutôt faible est attribuable à la nature du phénomène à l'étude (le revenu du ménage n'est pas facile à prédire), à l'information contenue dans l'enquête, et à la contrainte de la nécessité d'inclure seulement les covariables pour lesquelles le chiffre de population peut être obtenu du recensement.

La figure 1 contient des diagrammes des quartiles des résidus par région et par cycle, créés pour un des échantillons de Monte Carlo (les constatations sont très semblables d'un échantillon à l'autre). L'analyse des diagrammes suggère une corrélation infrarégionale et à l'intérieur des cycles, et donc la nécessité de préciser les modèles, comprenant les effets des régions et des cycles. Lorsque l'on analyse les résidus, les avantages de l'inclusion des effets interactifs (c'est-à-dire les effets régionaux fluctuant au fil du temps) sont moins clairs.

Par ailleurs, les résidus affichent une certaine autocorrélation, la moyenne du coefficient d'autocorrélation calculée pour tout les historiques individuels de valeurs résiduelles se chiffrant à 0,27. Bien que ce niveau d'autocorrélation ne soit pas très élevé, par souci d'intégralité, nous avons décidé de tenir également compte des modèles comportant des erreurs autocorrélées ou des effets aléatoires. Après avoir mis à l'essai diverses structures d'autocorrélation (ARMA(p, q), modèle linéaire généralisé, *etc*.), nous avons constaté que le processus autorégressif de l'ordre 1 est le mieux adapté à nos données.

L'asymétrie apparente des résidus suggère également que l'hypothèse de normalité pour les erreurs ne s'avère pas tout à fait adéquate. Nous maintenons cette hypothèse pour tous les modèles que nous précisons, et nous nous servons des estimateurs MVR pour les composantes de variance. En fait, nous pouvons nous attendre à ce que les écarts de la normalité aient un léger effet sur les valeurs ponctuelles des prédicteurs. Les formules des MPLSB peuvent être dérivées sans qu'il n'y ait normalité; qui plus est, nous avons de très bonnes raisons de nous attendre à ce que les estimateurs MVR (et MV) de ψ donnent de bons résultats, même si la normalité ne tient pas le coup (pour plus de détails, voir Jiang 1996). Les écarts de la normalité peuvent avoir des conséquences plus graves sur l'estimation de l'EQM, et nous nous pencherons sur ce problème à la section 4.2 ci-dessous

4. Résultats

4.1 Les estimateurs ponctuels

Tous les calculs nécessaires dans l'exercice de simulation décrit à la section 3 ont été effectués au moyen de la version 9.1 de SAS pour Windows. Les estimateurs MPLSBE sont obtenus à l'aide de Proc MIXED, et la création des échantillons est fondée sur Proc SURVEYSELECT.

Étant donné que l'estimation précise des paramètres régionaux est l'objectif principal de l'estimation pour petits domaines, nous avons d'abord évalué le rendement des estimateurs décrits pour prédire les valeurs des régions individuelles. Nous avons aussi évalué l'ampleur du surrétrécissement rattaché à chaque estimateur. En fait, les estimations pour petits domaines devraient refléter (au moins de façon approximative) la variabilité globale des paramètres régionaux sous-jacents.

Mentionnons que notre essai de simulation a pour objet d'évaluer les propriétés fondées sur le plan de sondage des estimateurs, c'est-à-dire que les échantillons aléatoires sont générés à partir d'une population fixe.

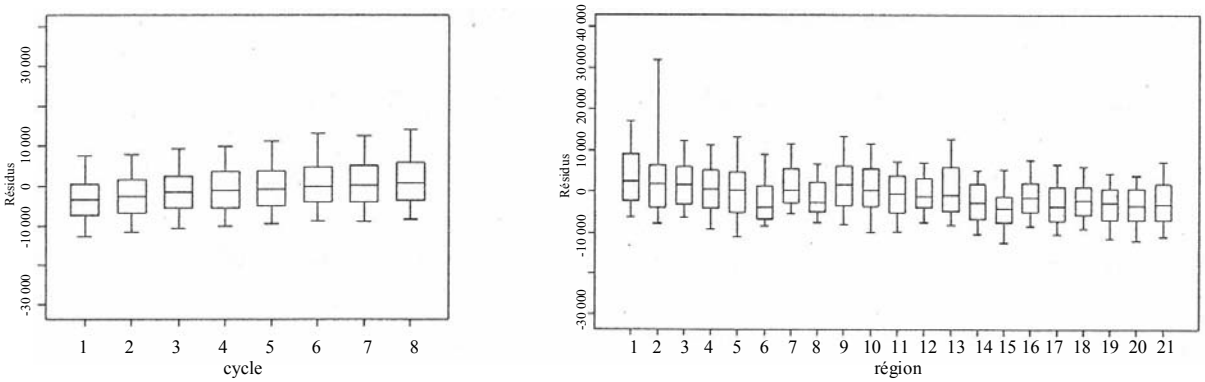


Figure 1 Diagramme des quartiles des résidus par cycle (à gauche) et par région (à droite)

Pour évaluer le rendement des estimateurs, nous avons adopté une approche souvent mentionnée dans la documentation (voir Rao 2003; section 7.2.6), au moyen de deux indicateurs : le biais relatif absolu moyen (BRAM) et la moyenne des erreurs quadratiques moyennes relatives (MEQMR) :

$$\begin{aligned} \text{BRAM} &= m^{-1}\sum_{d=1}^{m}\left|R^{-1}\sum_{r=1}^{R}\left(\frac{\tilde{y}_{dT(r)}}{\overline{Y}_{dT}}-1\right)\right| \\ \text{MEQMR} &= m^{-1}\sum_{d=1}^{m}\left\{R^{-1}\sum_{r=1}^{R}\left(\frac{\tilde{y}_{dT(r)}}{\overline{Y}_{dT}}-1\right)^2\right\} \end{aligned} \tag{18}$$

où $\tilde{y}_{dT(r)}$ représente l'estimation pour la région *d*, la période *T* et l'échantillon répété *r*, tandis que $\overline{Y}_{dT}$ est la moyenne de la population estimée. Soulignons que BRAM mesure le biais d'un estimateur, alors que MEQMR en mesure la précision. Le nombre de répétitions de *R* est établi à 500, ce qui devrait suffire pour obtenir les estimations Monte Carlo stables des valeurs prévues et des variances, une méthode souvent utilisée dans les études de simulation au sujet de l'estimation pour petits domaines (Heady, Higgins et Ralphs 2004; EURAREA Consortium 2004).

Les gains d'efficience associées à chacun des estimateurs régionaux sont évaluées au moyen du ratio de son MEQMR par rapport à celui de certains estimateurs dont nous nous servons comme points de repère. Plus précisément, tous les estimateurs sont comparés à la moyenne pondérée $\overline{y}_{dT,\,\text{DIR}}$ et nous désignons ce ratio MEFF_{Dir}. De plus, les estimateurs des MPLSBE associés aux modèles (5) - (9), qui s'appuient sur les données des cycles précédents, sont comparés aux estimateurs MPLSBE associés au modèle transversal (10), afin d'évaluer les gains d'efficience découlant de l'utilisation des cycles précédents. Dans ce cas, le ratio est désigné $\text{MEFF}_{\text{Trans}}$.

En ce qui concerne l'évaluation du degré de rétrécissement, nous avons comparé l'écart-type empirique des valeurs de la population du domaine :

$$\text{ETE} = \sqrt{m^{-1}\sum_{d=1}^{m}(\overline{Y}_{dT}-\overline{Y}_{T})^2},$$

où $\overline{Y}_{T}$ correspond aux valeurs de la moyenne de population des régions *m* pendant la période *T*, avec l'écart-type empirique des valeurs régionales estimées, qui sont obtenues par la formule qui suit dans le cadre d'une étude par simulation :

$$\text{ete} = R^{-1}\sum_{r=1}^{R}\left[\sqrt{m^{-1}\sum_{d=1}^{m}(\tilde{y}_{dT(r)}-\overline{\tilde{y}}_{T(r)})^2}\right],$$

où $\overline{\tilde{y}}_{T(r)}$ est la moyenne des valeurs estimées pour les *m* régions pendant la période *T* dans le cadre de l'exécution de la simulation *r*. La comparaison s'effectue au moyen de l'indicateur

$$\text{ETER} = \frac{\text{ete}}{\text{ETE}} - 1 \tag{19}$$

qui révèle en quoi l'écart-type empirique associé à un estimateur en particulier diffère de celui de la population.

Le tableau 1 indique les valeurs en pourcentage de BRAM, MEQMR, MEFF et ETER obtenues pour l'estimateur direct, les estimateurs fondés sur le plan de sondage selon les formules (2) et (3) et les estimateurs des MPLSBE dérivés à partir des modèles (5) - (10).

Tous les estimateurs fonctionnent considérablement mieux que $\overline{y}_{dT,\,\text{DIR}}$ en ce qui concerne MEQMR et donnent des valeurs MEFF_{Dir} de moins de 100 %. Nous constatons également que les estimateurs fondés sur le plan de sondage donnent de moins bons résultats que les estimateurs MPLSBE pour ce qui est de MEQMR, et que les gains d'efficience démontrées par MEFF_{Dir} sont particulièrement élevées dans certains cas (plus de 50 %). Ce résultat souligne l'exactitude supérieure des estimateurs fondés sur le modèle en question.

L'estimateur MPLSBE le plus fiable est celui du modèle *MM*5, où les effets régionaux et temporels sont indépendants et les résidus sont autocorrélés conformément à un processus AR(1), ce qui entraîne des gains d'efficience d'environ 60 % par rapport à l'estimateur direct. Vient ensuite l'estimateur MPLSBE associé au modèle *MM*1, qui diffère du premier estimateur en raison de l'absence de résidus autocorrélés.

Tableau 1 Indicateurs de rendement - information auxiliaire disponible

Modèle	BRAM%	MEQMR%	MEFF_{Dir}%	$\text{MEFF}_{\text{Trans}}$%	ETER%
DIR	0,0	0,787	100,0	-	15,6
COMP	2,7	0,552	70,1	-	-9,8
GREG	0,2	0,543	68,2	-	10,0
SMM	2,3	0,377	47,7	100,0	-8,7
*MM*1	3,1	0,358	45,3	95,0	2,4
*MM*2	2,4	0,427	54,1	113,4	-4,4
*MM*3	2,6	0,380	48,3	101,2	4,7
*MM*4	2,6	0,429	54,2	113,6	-8,0
*MM*5	2,9	0,318	40,4	84,7	-7,0

En ce qui concerne le biais, l'estimateur GREG donne la plus faible valeur de BRAM, comme on pouvait s'y attendre (Särndal, Swensson et Wretman 1992, chapitre 7; Veijanen, Lehtonen et Särndal 2005). Les estimateurs restants donnent une valeur de BRAM semblable. Parmi les estimateurs MPLSBE, ceux qui sont associés aux modèles *MM*1 et *MM*5 sont plus efficaces en ce qui concerne MEQMR, mais ils sont généralement plus biaisés que l'estimateur du *SMM*. Cette situation est probablement attribuable au fait que notre évaluation du rendement se limite au dernier cycle; pour ce sous-ensemble de données, nous nous attendons à ce que l'ajustement de la régression sous-jacente de *SMM*, en fonction du dernier cycle seulement, soit meilleur que l'ajustement fondé sur l'ensemble des données. Pour ce qui est des estimateurs MPLSBE, la colonne $\text{MEFF}_{\text{Trans}}$ indique comment les gains d'efficience des prédicteurs, en fonction de l'emprunt d'information au fil du temps, s'avèrent parfois positifs, parfois négatifs. Les modèles *MM*2 et *MM*4 (voir les formules (6) et (8)), où les effets de l'interaction entre la région et la période sont présents, sont apparemment inadéquats, étant donné que les prédicteurs associés aux deux modèles donnent des résultats plutôt médiocres. Le rendement du prédicteur associé à *MM*3 (voir (7)) est légèrement inférieur à celui du prédicteur associé au modèle transversal. Ce résultat plutôt surprenant est probablement attribuable au petit nombre de cycles, qui ne permet pas d'estimer efficacement le coefficient de corrélation entre les divers effets temporels consécutifs.

Comme nous l'avons déjà mentionné, l'estimateur associé au modèle *MM*5 est celui qui offre le meilleur rendement : il est considérablement plus efficace que celui de *SMM*, et affiche un ratio $\text{MEFF}_{\text{Trans}}$ d'environ 85 %, ce qui représente un gain d'efficience d'environ 15 % si l'on tient compte de plus d'un cycle. L'estimateur MPLSBE associé à *MM*1 se révèle également plus efficace que celui de *SMM*, mais dans ce cas-ci, le gain d'efficience s'établit à seulement 5 %. Ces résultats confirment le fait que les données au niveau du ménage à plusieurs moments consécutifs peuvent être utilisées, au moyen de certains types de modèles longitudinaux, pour produire des estimations plus efficientes.

Si nous passons à l'indicateur de rétrécissement mentionné à la dernière colonne du tableau, nous voyons que l'estimateur direct surestime l'écart-type de la population des moyennes régionales de 15 %. On observe le même effet, quoique quelque peu atténué, pour ce qui est de l'estimateur GREG, dont l'écart-type est gonflé de 10 %. En revanche, l'estimateur COMP a tendance à rétrécir les estimations vers le centre de la répartition, ce qui entraîne une réduction de l'écart-type des moyennes régionales d'environ 10 % par rapport à la population. Ces résultats s'alignent à ceux qu'ont obtenu d'autres auteurs qui ont comparé les mêmes types d'estimateurs (Heady et coll. 2004; Spjøtvoll et Thomsen 1987). Les résultats obtenus pour les estimateurs des MPLSBE sont plus encourageants, étant donné que la différence calculée en pourcentage est toujours inférieure à 10 % en chiffres absolus. Ainsi, à cet égard, tous les estimateurs MPLSBE semblent acceptables. De plus, nous pouvons nous attendre à ce que les estimateurs MPLSB soient sous-dispersés comparativement aux paramètres de population correspondants. En pareil cas, l'indicateur ETER prend des valeurs positives pour certains estimateurs longitudinaux MPLSBE, parce qu'il est calculé seulement en fonction du dernier cycle, tandis que les modèles longitudinaux visent à prédire $m \times T$ paramètres.

Le tableau 2 résume les résultats au sujet des estimateurs MPLSBE qui sont associés aux modèles de variance de l'erreur aléatoire, comme décrit au dernier paragraphe de la section 2. Lorsque les modèles ne comportent aucune variable auxiliaire, l'avantage de l'emprunt d'information selon la période et la région ressort indépendamment de l'avantage associé aux covariables.

Comme prévu, les gains d'efficience mesurés par MEFF_{Dir} sont inférieurs à ceux du tableau 1, bien que les réductions de MEQMR demeurent importantes. Le classement des prédicteurs associés à la spécification des divers effets aléatoires demeure le même que celui du tableau 1, le prédicteur associé au modèle $MM5^*$ s'avérant le plus efficient, comme l'indique la colonne MEQMR%. Les gains d'efficience associés à ce dernier estimateur par rapport à l'estimateur direct se chiffrent à environ 43 %.

Tableau 2 Indicateurs de rendement – aucune information auxiliaire disponible

Modèle	BRAM%	MEQMR%	MEFF_{Dir}%	$\text{MEFF}_{\text{Trans}}$%	ETER%
SMM^*	2,7	0,575	72,8	100,0	-7,6
$MM1^*$	2,9	0,556	70,3	96,6	7,5
$MM2^*$	2,8	0,639	80,8	111,0	-3,0
$MM3^*$	3,7	0,574	72,6	99,7	8,6
$MM4^*$	3,5	0,691	87,2	119,8	-6,7
$MM5^*$	3,0	0,445	56,2	77,2	-6,3

Pour ce qui est du biais, les estimateurs MPLSBE découlant des modèles sans covariables ont tendance à être plus biaisés que les modèles correspondants comportant des covariables.

L'analyse de la colonne $\text{MEFF}_{\text{Trans}}$ révèle que la réduction de MEQMR permise par certains des modèles empruntant de l'information au fil du temps est plus prononcée lorsque que des covariables sont incluses; en effet, elle atteint 22 % dans le meilleur exemple du modèle $MM5^{*}$.

Ce dernier résultat est vraiment encourageant. En fait, dans le contexte de l'estimation pour petits domaines, l'absence des totaux connus des covariables de la population peut s'avérer fort contraignante quand vient le temps d'obtenir des estimations fiables. La réduction de MEQMR observée relativement à la considération de plusieurs cycles d'une enquête par panel révèle que les estimations peuvent être améliorées grâce à l'emprunt d'information au fil du temps, lorsqu'il n'est pas possible de mettre à profit l'information auxiliaire.

Pour ce qui est des résultats du rétrécissement, on peut aussi les considérer acceptables dans ce cas-ci, et on constate un lien entre les résultats obtenus pour les estimateurs MPLSBE dérivés de modèles analogues avec ou sans covariables.

4.2 Comparaison de différents estimateurs de l'EQM associée aux estimateurs MPLSBE

À la section 2.3, nous avons examiné trois types d'estimateurs de l'EQM associés aux estimateurs MPLSBE. Dans la présente section, nous allons comparer le rendement de ces trois estimateurs au moyen d'un exercice de simulation. Comme nous nous attarderons sur une estimation de l'EQM plutôt qu'une comparaison des estimateurs MPLSBE dérivés de différents modèles, nous tiendrons seulement compte du prédicteur associé au modèle *MM5*, dont le rendement a été jugé supérieur aux autres à la section précédente.

Supposons que nous désignons le prédicteur de $\bar{Y}_{dT}$ par $\tilde{\eta}_{dT}^{\text{MPLSBE}}$ et son erreur quadratique moyenne par $\text{EQM}(\tilde{\eta}_{dT}^{\text{MPLSBE}})$. L'estimateur suivant :

$$\text{eqm}_{\text{ACT}}(\tilde{\eta}_{dT}^{\text{MPLSBE}}) = \frac{1}{R}\sum_{r=1}^{R}(\tilde{\eta}_{dT(r)}^{\text{MPLSBE}} - \tilde{\bar{\eta}}_{dT}^{\text{MPLSBE}})^2 + (\tilde{\bar{\eta}}_{dT}^{\text{MPLSBE}} - \bar{Y}_{dT})^2,$$

où $\tilde{\eta}_{dT(r)}^{\text{MPLSBE}}$ est $\tilde{\eta}_{dT}^{\text{MPLSBE}}$ calculé en fonction du r^{e} échantillon répété et $\tilde{\bar{\eta}}_{dT}^{\text{MPLSBE}} = R^{-1}\sum_{r=1}^{R}\tilde{\eta}_{dT(r)}^{\text{MPLSBE}}$, servira de point de repère pour comparer le rendement des estimateurs de l'erreur quadratique moyenne décrits à la section 2.3, parce que l'erreur quadratique moyenne réelle est inconnue.

Comme dans le cas des estimateurs ponctuels, tous les calculs sont effectués au moyen de SAS. Pour déterminer l'estimateur Prasad-Rao (14), les résultats de l'énoncé ESTIMATE de Proc MIXED sont utilisés et l'option KENWARDROGER est activée. La somme $g_1(\hat{\psi}) + g_2(\hat{\psi})$ est obtenue des résultats de Proc MIXED. L'option KENWARDROGER permet le calcul du facteur d'inflation de l'EQM décrit dans Kenward et Rogers (1986), lequel est équivalent à $2g_3(\hat{\psi})$ (voir aussi Rao 2003, section 6.2.7).

L'estimateur $\text{eqm}_{\text{BL}}(\tilde{\eta}_{dT}^{\text{MPLSBE}})$ est fondé sur le ré-échantillonnage. Donc, l'évaluation de son rendement dans le cadre d'un exercice de Monte Carlo nécessite la mise en œuvre de deux simulations emboîtées : pour chaque $r\,(r = 1,...,R)$, nous exécutons les répétitions R_{BOOT} nécessaires pour estimer les attentes à l'égard du modèle bootstrap. Pour atténuer le fardeau de calcul, nous établissons que $R_{\text{BOOT}} = 150$. Butar et Lahiri (2003) proposent une approximation analytique de l' eqm_{BL}, mais uniquement dans le cas des modèles qui ne sont pas aussi complexes que celui qui nous occupe.

Aussi bien pour $\text{eqm}_{\text{BL}}(\tilde{\eta}_{dT}^{\text{MPLSBE}})$ que pour $\text{eqm}_{\text{JLW}}(\tilde{\eta}_{dT}^{\text{MPLSBE}})$, nous avons préparé des codes SAS spéciaux utilisant les résultats de Proc MIXED.

Pour comparer les trois estimateurs de l'EQM, nous employons les mêmes mesures qui sont utilisées pour évaluer le rendement des estimateurs ponctuels, BRAM et MEQMR. Compte tenu du risque de sous-estimation des estimateurs de l'EQM, nous nous intéressons également à tout indice de biais associé aux estimateurs en question. Par conséquent, dans le cas des estimateurs de l'EQM, en plus de la moyenne des valeurs absolues des estimations obtenues pour le biais de chaque région (BRAM), nous calculons également la moyenne de ces estimations sans tenir compte de la valeur absolue (BRAM′), de manière à mieux comprendre si les estimateurs donnés ont effectivement tendance à sous-évaluer l'EQM. Donc, les mesures calculées s'énoncent comme suit :

$$\text{BRAM} = m^{-1}\sum_{d=1}^{m}\left|R^{-1}\sum_{r=1}^{R}\left(\frac{\text{eqm}_{*}(\tilde{\eta}_{dT}^{\text{MPLSBE}})}{\text{eqm}_{\text{ACT}}(\tilde{\eta}_{dT}^{\text{MPLSBE}})} - 1\right)\right|$$

$$\text{BRAM}' = m^{-1}\sum_{d=1}^{m}\left\{R^{-1}\sum_{r=1}^{R}\left(\frac{\text{eqm}_{*}(\tilde{\eta}_{dT}^{\text{MPLSBE}})}{\text{eqm}_{\text{ACT}}(\tilde{\eta}_{dT}^{\text{MPLSBE}})} - 1\right)\right\},$$

$$\text{MEQMR} = m^{-1}\sum_{d=1}^{m}\left\{R^{-1}\sum_{r=1}^{R}\left(\frac{\text{eqm}_{*}(\tilde{\eta}_{dT}^{\text{MPLSBE}})}{\text{eqm}_{\text{ACT}}(\tilde{\eta}_{dT}^{\text{MPLSBE}})} - 1\right)^2\right\}$$

où le symbole* renvoie aux procédures d'estimation envisagées, soit PR, BL et JLW. Les résultats des comparaisons en fonction des $R = 500$ itérations de MC sont indiqués dans le tableau 3.

En ce qui concerne MEQMR et BRAM, les trois estimateurs se comportent de façon semblable, et aucun d'entre eux ne se démarque nettement des deux autres. Néanmoins, la colonne BRAM′ démontre clairement que

l'eqm_{PR} et l'eqm_{BL} sous-estiment systématiquement l'EQM_{ACT}, tandis que l'eqm_{JLW} ne le fait pas. Ce phénomène est probablement attribuable à l'échec de l'hypothèse de normalité pour les termes d'erreurs. En fait, comme nous l'avions prédit à la section 3, le revenu équivalent est une variable désaxée vers la droite, ce qui semble également le cas des résidus de régression **e**. La normalité est une hypothèse essentielle à la dérivation de l'eqm_{PR} et de l'eqm_{BL}, tandis que l'on pourrait s'attendre ce que l'eqm_{JLW} soit plus robuste à cet égard. Nos conclusions sont conformes aux prédictions théoriques et aux résultats des simulations décrits dans Jiang et coll. (2002). Même si Bell (2001) a indiqué que l'eqm_{JLW} pourrait être négatif dans certains ensembles de données en raison de la correction du biais, ce scénario de s'est jamais produit dans nos simulations. Dans tous les ensembles de données répétés, le deuxième terme de (17) donne un résultat positif, en plus de contribuer considérablement à l'estimation de l'EQM dans la plupart des cas. Jiang et Lahiri (2006b) décrivent les modifications de (17) en cas de valeurs négatives.

Tableau 3 Rendement des estimateurs de l'EQM de $\tilde{\eta}_{dT}^{\text{MPLSBE}}$ au moyen du modèle *MM*5

Estimateur	BRAM	BRAM'	MEQMR
eqm_{PR}	0,378	-0,383	0,238
eqm_{BL}	0,377	-0,318	0,228
eqm_{JLW}	0,337	0,036	0,261

Bref, dans le cas du problème qui nous occupe, l'eqm_{JLW} semble la plus appropriée des trois méthodes d'estimation de $\text{EQM}(\tilde{\eta}_{dT}^{\text{MPLSBE}})$. Cette conclusion pourrait revêtir une importance particulière pour toute application de la théorie des modèles linéaires mixtes fondés sur la normalité aux ensembles de données où les hypothèses de normalité pour les termes d'erreurs ne s'avèrent pas tout à fait adéquates.

Nous avons répété l'exercice de simulation pour le modèle transversal sans covariables SMM^*, qui est souvent considéré dans les simulations visant à comparer diverses méthodes d'estimation. À cette fin, soulignons que pour ce modèle, le ratio $\hat{\sigma}_e^2 / \hat{\sigma}_v^2$ est d'environ 12, ce qui entraîne un prédicteur MPLSBE caractérisé par $\gamma_i = \hat{\sigma}_v^2 n_i (\hat{\sigma}_v^2 n_i + \hat{\sigma}_e^2)^{-1}$ de 0,54 à 0,9. Signalons en outre que certaines régions sont caractérisées par la présence de valeurs aberrantes (le coefficient d'asymétrie γ_1 varie de 0,1 à 4,6).

Selon cette hypothèse, les estimateurs de l'EQM ont un comportement fort différent de celui qui est illustré dans le cas du modèle *MM*5. Les résultats sont indiqués dans le tableau 4.

Tous les estimateurs surestiment l'EQM réelle, mais la surestimation de mse_{JLW} est la moins prononcée des trois. Grâce à une analyse détaillée des résultats relatifs aux régions individuelles, nous concluons que les valeurs de BRAM′ (qui représente la différence la plus apparente par rapport aux résultats du tableau 3) sont caractérisées par une forte surestimation de l'EQM réelle dans les régions comportant les plus faibles niveaux d'asymétrie et d'aplatissement. Pour ces régions $\hat{\sigma}_e^2$ surestime largement la variation réelle des données, ce qui entraîne la surestimation de $g_1(\sigma_v^2, \sigma_e^2)$. Cette situation est probablement attribuable au fait que l'échec de la normalité (l'excès d'aplatissement) entraîne la surestimation de σ_e^2. Ce problème ne s'est pas manifesté dans le cas du modèle *MM*5, grâce à la présence de covariables et à la modélisation AR(1) des résidus individuels.

Tableau 4 Rendement des estimateurs de l'EQM des $\tilde{\eta}_{dT}^{\text{MPLSBE}}$ en fonction du modèle SSM*

Estimateur	BRAM	BRAM'	MEQMR
mse_{PR}	0,449	0,262	0,503
mse_{BL}	0,376	0,213	0,376
mse_{JLW}	0,354	0,149	0,335

5. Remarques finales et autres points

Les résultats obtenus révèlent que, dans l'ensemble, les estimateurs MPLSBE dérivés des spécifications du modèle linéaire mixte au niveau des unités qui empruntent de l'information d'une période à l'autre et d'une région à l'autre donnent lieu à des gains d'efficience importants comparativement à l'estimateur direct et aux autres estimateurs populaires fondés sur le plan de sondage, tels que l'estimateur composite optimal et l'estimateur GREG. En outre, l'erreur quadratique moyenne de certains estimateurs longitudinaux MPLSBE en question est, en moyenne sur l'ensemble des régions, beaucoup plus faible que celle des estimateurs analogues transversaux MPLSBE. Parmi les spécifications des modèles utilisés pour dériver les estimateurs MPLSBE, ceux dont les effets temporels et régionaux sont indépendants, qu'ils tiennent compte ou non de l'autocorrélation des résidus, semblent être les plus efficaces (gains d'efficience d'environ 55-60 % comparativement à l'estimateur direct). Ces résultats se maintiennent également lorsque les covariables sont supprimées; en fait, on peut obtenir des estimations pour petits domaines fiables même en l'absence de covariables, à condition que des observations répétées de la même unité à divers moments soient disponibles. De toute manière, l'effet de rétrécissement associé aux estimateurs MPLSBE semble modéré, ce qui atténue la nécessité de l'estimation d'ensemble ou multiple (Rao 2003, chapitre 9). En ce qui concerne l'estimation de l'EQM pour les estimateurs régionaux en question, nous avons constaté que l'estimateur jackknife donne en moyenne les meilleurs résultats sur l'ensemble des régions, et qu'il est donc plus résistant à tout écart par

rapport aux hypothèses généralisées des modèles linéaires mixtes. Cette conclusion peut revêtir une importance particulière pour toute application de la théorie des modèles linéaires mixtes fondés sur la normalité aux ensembles de données où les hypothèses de normalité ne s'avèrent pas tout à fait adéquates, comme c'est le cas des données sur le revenu.

Remerciements

La recherche a été financée en partie par Miur-PRIN 2003 *Statistical analysis of changes of the Italian productive sectors and their territorial structure*, sous la coordination du professeur C. Filippucci. Les travaux d'Enrico Fabrizi ont été subventionnés en partie par l'université de Bergamo (subventions nº 60FABR06 et 60BIFF04).

Nous remercions ISTAT d'avoir eu la gentillesse de nous fournir les données dont nous nous sommes servis pour réaliser nos travaux.

Bibliographie

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Bell, W. (2001). Discussion with "Jackknife in the Fay-Herriott model with an example". *Proceeding of the Seminar on Funding Opportunity in Survey Research*, 98-104.

Betti, G., et Verma, V. (2002). Non-monetary or lifestyle deprivation, in EUROSTAT (2002). *Income, Poverty Risk and Social Exclusion in the European Union*, Second European Social Report, 87-106.

Butar, F., et Lahiri, P. (2003). On measures of uncertainty of empirical bayes small-area estimators. *Journal of Statistical Planning and Inference*, 112, 63-76

Datta, G.S., et Lahiri, P. (1999). A unified measure of uncertainty of estimated best linear predictors in small area estimation problems. *Statistica Sinica*, 10, 613-627.

Datta, G.S., Lahiri, P. et Maiti, T. (2002). Empirical bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference*, 102, 83-97.

Datta, G.S., Lahiri, P., Maiti, T. et Lu, K.L. (1999). Hierarchical bayes estimation of unemployment rates for the States of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.

EURAREA CONSORTIUM (2004). EURAREA. Enhancing Small Area Estimation Techniques to meet European Needs, Project Reference Volume, téléchargeable au http://www.statistics.gov.uk/eurarea/download.asp.

EUROSTAT (2002). European social statistics - Income, poverty and social exclusion. 2ième rapport.

Falorsi, P.D., Falorsi, S. et Russo, A. (1994). Comparaison empirique de méthods d'estimation pour petites régions pour l'enquête sur la population active italienne. *Techniques d'enquête*, 20, 179-184.

Falorsi, P.D., Falorsi, S. et Russo, A. (1999). Small area estimation at provincial level in the Italian Labour Force Survey. *Journal of the Italian Statistical Society*, 1, 93-109.

Ghosh, M., Nangia, N. et Kim, D. (1996). Estimation of median income of four-person families: A bayesian time series approach. *Journal of the American Statistical Association*, 91, 1423-1431.

Heady, P., Higgins, N. et Ralphs, M. (2004). Evidence-Based Guidance on the Applicability of Small Area Estimation Techniques. Article présenté à l' European Conference on Quality and Methodology in Official Statistics, Mainz, Allemangne, 24-26 mai.

Jiang, J. (1996). REML estimation: Asymptotic behavior and related topics. *The Annals of Statistics*, 24, 255-286.

Jiang, J., Lahiri, P. et Wan, S.M. (2002). A unified jackknife theory for empirical best prediction with *M*-estimation. *The Annals of Statistics*, 30, 1782-1810.

Jiang, J., et Lahiri, P. (2006a). Estimation of finite population domain means - A model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.

Jiang, J., et Lahiri, P. (2006b). Mixed model prediction and small area estimation. Discussion sollicitée par l'éditeur, *Test*, 15, 1, 1-96.

Kenward, M.G., et Roger, J.H. (1986). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53, 983-997

Lehtonen R., Särndal C.-E. et Veijanen A. (2003) L'effet du choix d'un modèle dans l'estimation par domaine, dont les petits domaines. *Techniques d'enquête*, 29, 1, 37-49.

Pfeffermann, D. (2002). Small area estimation - New developments and directions. *Revue internationale de Statistique*, 70, 125-143.

Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of mean squared errors of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Rao, C.R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, 67, 112-115.

Rao, J.N.K. (2003). *Small Area Estimation*, New York: John Wiley & Sons, Inc.

Rao, J.N.K., et You, M. (1994). Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics*, 22, 511-528.

Rendtel, U., Behr, A. et Sisto, J. (2003). Attrition effect in the European Community household panel, CHINTEX PROJECT, European Commission.

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

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'enquête*, 20, 35-46.

Spjøtvoll, E., et Thomsen, I. (1987). Application of some empirical bayes methods to small area statistics. *Bulletin of the International Statistical Institute*, 4, 435-450.

Vandecasteele, L., et Debels, A (2004). Modelling attrition in the European Community Household Panel: The effectiveness of weighting, 2ième Conférence Internationale des utilisateurs de l'ECHP, EPUNet 2004, Berlin, juin 24-26.

Veijanen, A., Lehtonen, R. et Särndal, C.-E. (2005). The Effect of Model Quality on Model-Assisted and Model-Dependent Estimators of Totals and Class Frequency for Domains, Article présenté à la Conférence de SAE2005, Challenges in Statistics Production for Domains and Small Areas, 28-31 août 2005, Jyväskylä, Finland.

Verbeke, G., et Molenberg, H. (2000). *Linear Mixed Models for Longitudinal Data*, New York: Springer-Verlag.

Techniques d'enquête, décembre 2007
Vol. 33, N° 2, pp. 221-232
Statistique Canada, N° 12-001-X au catalogue

Estimation de la couverture du recensement de la population de l'an 2000 en Suisse : méthodes et résultats

Anne Renaud [1]

Résumé

Les défauts de couverture sont estimés et analysés pour le recensement de la population de l'an 2000 en Suisse. La composante de sous-couverture est estimée sur la base d'un échantillon indépendant du recensement et d'un appariement avec le recensement. La composante de sur-couverture est estimée sur la base d'un échantillon tiré dans la liste du recensement et d'un appariement avec le reste du recensement. Les composantes de sur- et sous-couverture sont ensuite combinées pour obtenir une estimation de la couverture nette résultante. Cette estimation est basée sur un modèle de capture-recapture, nommé système dual, combiné avec un modèle synthétique. Les estimateurs sont calculés pour la population entière et différents sous-groupes, avec une variance estimée par un jackknife stratifié. Les analyses de couverture sont complétées par une étude des appariements entre l'échantillon indépendant et le recensement afin de déterminer les erreurs de mesure et de localisation potentielles dans les données du recensement.

Mots clés : Recensement; erreurs de couverture; système dual; plan d'échantillonnage à plusieurs degrés; erreurs de mesure.

1. Introduction

Dans tout recensement, certaines personnes ne sont pas recensées alors qu'elles devraient l'être et d'autres sont comptées deux fois ou n'auraient pas dû être recensées. Il y a donc de la sous-couverture et de la sur-couverture, dont le bilan est très souvent une sous-couverture nette. La sous-couverture nette est par exemple estimée à 1,6 % aux États-Unis en 1990 (Hogan 1993), 2,2 % au Royaume-Uni en 1991 (Brown, Diamond, Chambers, Buckner et Teague 1999) et 3 % au Canada en 2001 (Statistique Canada 2004). Aux États-Unis en 2000, la couverture nette correspond par contre à une sur-couverture de 0,5 % (Hogan 2003). Les défauts de couverture peuvent fortement varier entre sous-groupes de la population. Aux États-Unis en 2000, on note par exemple que les Noirs avaient une sous-couverture nette de 1,8 % alors que les Blancs avaient une sur-couverture de 1,1 %. De plus, les valeurs varient souvent entre classes d'âge et régions par exemple. Ces défauts de couverture, et les autres erreurs telles les erreurs de mesure, conduisent à une image biaisée de la population. Elles sont donc étudiées afin d'avoir une information sur la qualité des données disponibles et trouver des pistes pour améliorer les relevés censitaires auprès de la population.

Le recensement de la population de l'an 2000 en Suisse donne une image de la population au 5 décembre 2000. Pour la première fois, on estime les défauts de couverture d'un recensement suisse. La sous-couverture, la sur-couverture et la couverture nette résultante du recensement 2000 sont toutes trois analysées. La sous-couverture est estimée sur la base d'un échantillon de personnes S_P, indépendant du recensement, sur lequel on organise une enquête de couverture quelques mois après le recensement (relevé entre avril et mai 2001). Les données de cette enquête sont alors appariées avec les données du recensement afin de déterminer si la personne de S_P a été recensée ou pas. La sur-couverture est estimée sur la base d'un échantillon de personnes S_E tiré parmi les enregistrements du recensement. Une recherche de doubles et d'autres enregistrements erronés permet alors de déterminer si l'enregistrement correspond à une vraie personne à recenser ou pas. La couverture nette est estimée sur la base d'un modèle de capture-recapture nommé système dual (Wolter 1986, Fienberg 1992). L'estimateur dual est appliqué dans des cellules homogènes et les résultats recombinés en suivant un modèle synthétique pour obtenir des résultats pour différents domaines de la population (Hogan 2003). Le but du projet n'est pas d'ajuster les chiffres du recensement mais d'obtenir des informations sur la qualité du recensement de l'an 2000 et les potentiels d'améliorations pour les recensements suivants.

Cet article présente les différentes étapes des estimations et les résultats. Les sections 2 et 3 décrivent les jeux de données et les estimateurs de couverture. La section 4 expose les détails de la construction des différents statuts utilisés dans les estimateurs. La section 5 décrit l'approche utilisée pour comparer les valeurs relevées dans le recensement et dans l'enquête pour les personnes appariées de S_P. Les sections 6 et 7 présentent les résultats numériques et la conclusion.

1. Anne Renaud, Service de méthodes statistiques, Office fédéral de la statistique, Espace de l'Europe 10, CH-2010 Neuchâtel, Suisse. Courriel : Anne.Renaud@bfs.admin.ch.

2. Les trois jeux de données

2.1 Recensement

Le recensement de l'an 2000 a été réalisé sous l'égide de l'Office fédéral de la statistique, avec la date de référence du 5 décembre 2000. Des informations ont été relevées pour 7,3 millions d'habitants, 3,1 millions de ménages, 3,8 millions de logements et 1,5 millions de bâtiments. Les différents niveaux ont ensuite été liés par des identificateurs communs lors du traitement des données.

Le relevé des personnes et ménages était sous la responsabilité des 2 896 communes politiques suisses. Ces dernières avaient le choix entre différentes formes de relevé :

- CLASSIC : agents recenseurs;
- SEMI-CLASSIC : préimpression de questionnaires sur la base du registre communal des habitants, envoi par la poste et collecte par des agents recenseurs;
- TRANSIT : préimpression de questionnaires, envoi et retour par la poste;
- FUTURE : idem à TRANSIT avec liens entre les ménages et les logements fournis par la commune;
- TICINO : similaire à TRANSIT mais restreint au canton du Tessin.

La majorité des communes SEMI-CLASSIC, TRANSIT, FUTURE et TICINO offraient également la possibilité de remplir les questionnaires sur Internet. Les 2 208 communes SEMI-CLASSIC, TRANSIT, FUTURE et TICINO qui ont utilisé la préimpression des questionnaires sur la base des registres communaux des habitants contiennent près de 96 % de la population. Les travaux d'envois et de contrôle des retours de la plupart de ces communes étaient organisés dans un centre national.

Le jeu de données des personnes comporte 7 452 075 entrées. Il a la particularité de comporter deux enregistrements pour la même personne si cette dernière a deux domiciles (2,3 % de la population, par exemple étudiant avec un domicile chez ses parents et un domicile proche de son école). Dans le cas de deux domiciles, l'un est codé comme *domicile économique* et l'autre comme *domicile civil*. Le domicile économique est l'endroit où la personne passe le plus de temps par semaine et le domicile civil est l'endroit où la personne a ses papiers officiels (acte de naissance pour les Suisses, permis de séjour pour les étrangers). Dans le cas d'un unique domicile, ce dernier est le domicile économique et le domicile civil tout à la fois. La *population résidante* en Suisse de 7 280 010 personnes est définie par l'ensemble des enregistrements au domicile économique.

Les ménages sont classifiés en *ménages privés*, *collectifs* et *administratifs*. Les ménages privés sont par exemple des familles, des couples ou encore des personnes vivant seules. Les ménages collectifs sont par exemple des groupes de locataires d'une résidence pour personnes âgées ou d'un internat, ou les détenus d'une prison. Les ménages administratifs rassemblent les personnes sans domicile fixe, les gens en voyage et les personnes - par bâtiment ou commune - n'ayant pas pu être assignées à des ménages privés ou collectifs (2,4 % de la population résidante).

Les données du recensement ne comportent aucune imputation au niveau des enregistrements car les communes ont envoyé les informations de base pour les non-répondants (unit nonresponse). Des valeurs sont cependant imputées dans les cas de données manquantes ou d'incohérence dans les questionnaires (item nonresponse).

La *population d'intérêt* pour les estimations de couverture est la population résidante (domicile économique) dans des ménages privés et administratifs. Les ménages collectifs, qui représentent 2,3 % de la population résidante recensée, sont exclus des estimations.

2.2 Échantillon S_P, enquête de couverture et appariement (sous-couverture)

L'objectif de taille de l'échantillon S_P est fixé à environ 50 000 personnes. Faute de bases existantes en Suisse, cette valeur a été déterminée de manière approximative sur la base des expériences faites à l'étranger. Les résultats australiens de 1996 ont notamment été utilisés car le plan d'échantillonnage de leur enquête de couverture était similaire à celui prévu en Suisse en 2000 (ABS 1997).

L'échantillon S_P, indépendant du recensement, est construit en deux parties : le canton du Tessin (TICINO) et le reste de la Suisse (NORD). Les deux parties utilisent un tirage à plusieurs degrés. Le premier degré consiste en la sélection de 303 unités primaires, communes politiques pour TICINO et numéros postaux pour NORD, selon un plan stratifié avec un tirage proportionnel au nombre de bâtiments. Le deuxième degré consiste en un tirage aléatoire simple d'un nombre fixe de 60 bâtiments par unité primaire. Dans le plan NORD, ces bâtiments sont répartis dans un maximum de 3 tournées de distribution du courrier grâce à un degré d'échantillonnage intermédiaire. L'échantillonnage est ainsi construit de manière à regrouper le travail sur le terrain tout en limitant la variabilité des poids. Pour des questions pratiques et de ressources, les numéros postaux comportant une grande proportion de bâtiments sans adresses postales complètes ou codés comme inhabités sont sélectionnés avec une probabilité plus faible que les autres numéros postaux. Il s'agit principalement de numéros postaux dans des régions rurales ou encore de zones industrielles, peu enclins à des défauts de couverture importants. Des listes exhaustives de ménages sont établies sur le terrain avec l'aide des employés postaux dans

l'échantillon d'environ 16 000 bâtiments avant de passer à un sous-échantillonnage de bâtiments permettant d'atteindre un total de environ 27 000 ménages. Pour plus d'information sur l'échantillonnage et la procédure d'enquête, voir Renaud (2001) et, plus détaillé, Renaud et Eichenberger (2002).

L'enquête de couverture consiste à contacter les 27 000 ménages; par téléphone si un numéro est trouvé et en face-à-face si cela n'est pas le cas. On relève les variables permettant un appariement avec le recensement et la définition de sous-groupes intéressants pour l'étude de la couverture (variables sociodémographiques, adresses). Le relevé porte sur tous les membres de tous les ménages des bâtiments sélectionnés.

L'échantillon final S_P contient $n_P = 49\,883$ personnes dans la population d'intérêt (domicile économique et ménage privé). 88 % des ménages ont été atteints par téléphone et 12 % en face-à-face. La pondération dépend de l'échantillonnage et d'un ajustement pour la non-réponse. L'ajustement pour la non-réponse repose sur un modèle d'homogénéité dans des cellules construites sur la base des strates d'échantillonnage et de la connaissance ou non de l'existence d'un numéro de téléphone (interviews prévu par téléphone ou en face-à-face). Il intègre également une estimation de la proportion de vrais ménages parmi les ménages à contacter. Une partie non négligeable des ménages à contacter étaient en effet formée de logements de vacances, de commerces ou encore d'entreprises. Aucun calage n'est appliqué car les données auxiliaires disponibles ne sont pas indépendantes du recensement. Il n'y a pas de non-réponse partielle. Les détails de la pondération sont documentés dans Renaud et Potterat (2004).

Sur la base des questions posées durant l'enquête et des divers contrôles de plausibilité, on fait l'hypothèse que les données de S_P sont correctes et utilisables pour l'appariement avec le recensement. Les critères de qualité utilisés sont les suivants :

- *complétude* : l'enregistrement permet d'identifier la personne;
- *pertinence* : la personne devait bien être recensée;
- *unicité* : la personne est listée une seule fois;
- *appartenance à la population d'intérêt*; la personne est listée au domicile économique et dans un ménage privé;
- *localisation* : la personne est listée à la bonne adresse au jour du recensement.

L'appariement entre l'échantillon S_P et le recensement permet de déterminer le *statut d'appariement* P_j de chaque élément j de S_P. Le statut P_j vaut 1 si l'élément est apparié dans le recensement (personne recensée) et 0 si cela n'est pas le cas (personne non recensée). Dans notre cas, les données relevées durant l'enquête de couverture, les données finales du recensement et les images des questionnaires du recensement sont utilisées pour l'appariement automatique, l'appariement manuel et les contrôles. Aucune interview complémentaire n'a lieu en plus de l'enquête de couverture. Les personnes ayant déménagé entre le jour du recensement et le jour de l'enquête sont échantillonnées à leur adresse au jour de l'enquête puis recherchées en priorité à l'adresse qu'elles ont indiqué pour le jour du recensement. Aucun cas n'est indéterminé à la fin du processus.

2.3 Échantillon S_E et recherche des enregistrements erronés (sur-couverture)

L'objectif de taille de l'échantillon S_E est fixé à environ 55 000 personnes. Cette valeur, un peu supérieure à celle de S_P, n'influe que peu sur le traitement des données car il n'y a pas de travail sur le terrain ni d'interview complémentaire au recensement.

L'échantillon S_E est sélectionné dans les données du recensement selon un tirage à deux degrés. Seuls les éléments faisant partie de la population d'intérêt sont éligibles (enregistrements au domicile économique sans les membres des ménages collectifs). Les unités primaires de S_E sont identiques aux unités primaires de S_P (numéros postaux et communes). La liste des numéros postaux du plan NORD utilisée pour S_P ne correspond cependant pas exactement à la liste des numéros postaux présents dans les données du recensement. Les enregistrements du recensement se trouvant dans des numéros postaux inexistants dans la liste utilisée pour S_P sont donc redistribués dans des numéros existants en tenant compte de la localisation géographique (assignation de numéros postaux fictifs pour l'échantillonnage). Au deuxième degré, on tire des enregistrements de la population d'intérêt selon un plan aléatoire simple, sans degrés intermédiaires. L'allocation est choisie de façon à obtenir des poids constants dans les strates d'échantillonnage des unités primaires. Au final, l'échantillon comporte $n_E = 55\,375$ enregistrements (Renaud 2003).

On fait l'hypothèse que les enregistrements de S_E permettent d'identifier les personnes (complétude) car il y a peu d'imputation dans les données du recensement et que la plupart des questionnaires étaient préimprimés à partir des registres des habitants. La pertinence et l'unicité sont déterminées lors d'un appariement entre S_E et le reste du recensement selon un procédé similaire à l'appariement entre S_P et le recensement. Dans notre cas, il s'agit d'une recherche de doublets ou de triplets des éléments de S_E, complétée par une analyse de cas suspects dans S_E. Un élément j est considéré comme pertinent s'il n'est pas considéré comme erroné dans l'analyse des suspects (par exemple note sur le questionnaire indiquant que la personne

est partie à l'étranger). Un élément j est considéré unique si aucun double/triple n'est détecté dans le recensement. Aucune interview complémentaire n'a lieu auprès du S_E. Il n'y a donc pas d'information complémentaire au recensement sur les personnes de S_E (localisation réelle ? type effectif de domicile et de ménage ?). La recherche des doublets/triplets et des cas suspects aboutit à un *statut d'énumération* E_i pour chaque élément j de S_E. Le statut E_i vaut 1 si l'élément devait bien être énuméré dans le recensement (valeur par défaut) et 0 s'il ne devait pas l'être. En pratique, il peut prendre des valeurs entre 0 et 1 si le cas n'est pas déterminé de manière précise. Ainsi, les doublets et triplets reçoivent respectivement les valeurs 1/2 et 1/3 s'il n'y a pas d'informations permettant de déterminer l'enregistrement correct parmi les enregistrements détectés. Ces cas, rares, correspondent à des personnes ayant remplis plusieurs questionnaires au recensement sans qu'un lien n'ait été fait entre eux durant le traitement des données.

3. Estimateurs de la couverture

3.1 Sous-couverture et sur-couverture

Le *taux de sous-couverture* est estimé par $\hat{R}_{\text{sous}} = 1 - \hat{R}_m$, où $\hat{R}_m$ est l'estimation du *taux d'appariements corrects* basée sur l'échantillon S_P. De façon similaire, on définit le *taux de sur-couverture* $\hat{R}_{\text{sur}} = 1 - \hat{R}_c$, avec $\hat{R}_c$ l'estimation du *taux d'enregistrements corrects* basée sur l'échantillon S_E. Les taux d'appariements corrects et d'enregistrements corrects sont estimés par les moyennes pondérées des statuts d'appariement P_j et d'énumération E_j :

$$\hat{R}_m = \frac{\sum_{j \in S_P} w_{P,j} P_j}{\sum_{j \in S_P} w_{P,j}} \quad \text{et} \quad \hat{R}_c = \frac{\sum_{j \in S_E} w_{E,j} E_j}{\sum_{j \in S_E} w_{E,j}}, \quad (1)$$

avec $w_{P,j}$ le poids de l'élément j de l'échantillon S_P et $w_{E,j}$ le poids de l'élément j de l'échantillon S_E. Nous notons que le dénominateur de $\hat{R}_c$ est la somme des poids $w_{E,j}$ de S_E et non pas le nombre C d'enregistrements connus dans le recensement afin d'avoir un estimateur potentiellement moins biaisé.

L'estimation des taux de sous-couverture et de sur-couverture dans un domaine d est donnée par $\hat{R}_{\text{sous},d} = 1 - \hat{R}_{m,d}$ et $\hat{R}_{\text{sur},d} = 1 - \hat{R}_{c,d}$, avec

$$\hat{R}_{m,d} = \frac{\sum_{j \in S_P} w_{P,j} P_j I_{jd}}{\sum_{j \in S_P} w_{P,j} I_{jd}} \text{ et } \hat{R}_{c,d} = \frac{\sum_{j \in S_E} w_{E,j} E_j J_{jd}}{\sum_{j \in S_E} w_{E,j} J_{jd}}. \quad (2)$$

Les identificateurs I_{jd} et J_{jd} prennent la valeur 1 si l'élément j, respectivement de S_P et S_E, se trouve dans le domaine d et la valeur 0 sinon.

3.2 Couverture nette

Le *taux de sous-couverture nette* est estimé par $\hat{R}_{\text{sousnet}} = 1 - \hat{R}_{\text{net}}$ avec $\hat{R}_{\text{net}} = C / \hat{N}$ l'estimation du *taux de couverture nette*, C le nombre recensé dans la population d'intérêt et $\hat{N}$ l'estimation du vrai total dans la population d'intérêt. Si $\hat{R}_{\text{sousnet}}$ est négatif, nous sommes en présence d'une sur-couverture nette.

L'estimation du vrai total $\hat{N}$ est basée sur le modèle dual (Wolter 1986). Ce modèle repose sur un principe de capture (recensement) et recapture (enquête de couverture). Il est appliqué dans des cellules d'estimation $k = 1, ..., K$ afin de satisfaire au mieux les hypothèses du modèle; voir discussion plus bas. Ainsi, l'estimation du vrai total $\hat{N}$ est composée de la somme des vrais totaux estimés $\hat{N}_k$ dans des cellules d'estimations disjointes recouvrant l'ensemble de la population d'intérêt $k = 1, ..., K$:

$$\hat{N} = \sum_{k=1}^{K} \hat{N}_k. \quad (3)$$

Les totaux estimés $\hat{N}_k$ ont la forme donnée par le modèle dual :

$$\hat{N}_k = [N_{1+,k}] \left[\frac{N_{+1,k}}{N_{11,k}} \right], \quad (4)$$

avec $N_{1+,k}$ le total des enregistrements correctement comptés dans la cellule k durant la capture (recensement), $N_{+1,k}$ le total dans k durant la recapture (estimé sur la base de l'échantillon S_P) et $N_{11,k}$ le nombre d'enregistrements communs aux deux listes (estimé sur la base des appariements entre S_P et le recensement).

Les différents termes de l'équation (4) sont estimés sur la base des estimations de sous-couverture et de sur-couverture. Il s'agit d'une extension du modèle de Wolter (1986) proche de celle utilisée notamment par Hogan (2003). Ainsi, le total des enregistrements correctement comptés dans le recensement $N_{1+,k}$ est estimé par le produit du total recensé C_k par le taux d'enregistrements corrects $\hat{R}_{c,k}$ afin de tenir compte de la sur-couverture. De plus, le rapport entre le total dans la recapture $N_{+1,k}$ et le nombre d'enregistrements communs aux deux listes $N_{11,k}$ est estimé par l'inverse du taux d'appariement $\hat{R}_{m,k}$ entre l'enquête de couverture et le recensement afin de tenir compte de la sous-couverture. On obtient :

$$\hat{N}_k = [C_k \hat{R}_{c,k}][\hat{R}_{m,k}^{-1}] = C_k [\hat{R}_{c,k} \hat{R}_{m,k}^{-1}] = C_k \hat{F}_k, \quad (5)$$

avec $\hat{F}_k = \hat{R}_{c,k} \hat{R}_{m,k}^{-1}$ le *facteur de correction de la couverture* dans la cellule k. Le facteur $\hat{F}_k$ combine les effets de sur-couverture et de sous-couverture de la cellule k estimés sur la base des échantillons S_P et S_E. On note qu'une sous-couverture dans un domaine peut être compensée par une sur-couverture dans le même domaine.

Ainsi une sous-couverture nette nulle dans un domaine ne signifie pas qu'aucun défaut de couverture existe dans ce domaine.

Les estimations proposées reposent sur les hypothèses du modèle dual, le choix des cellules d'estimations, et le choix des statuts définissant les estimateurs $\hat{R}_{c,k}$ et $\hat{R}_{m,k}$. Le modèle dual est intéressant car il prend en compte le fait que certaines personnes ne sont atteintes ni par le recensement (capture) ni par l'enquête de couverture (recapture). Cependant, une série de contraintes doivent être respectées afin d'éviter des biais d'estimation. L'enquête de couverture et le recensement doivent être totalement indépendants. L'appariement doit être de très bonne qualité. Le modèle doit être appliqué dans des cellules avec des personnes ayant la même probabilité d'être énumérées dans le recensement, respectivement dans l'enquête; voir la section 3.3. Finalement, la population ne doit pas trop changer entre le jour du recensement et celui de l'enquête. De leur côté, les estimateurs $\hat{R}_{c,k}$ et $\hat{R}_{m,k}$ sont basés sur la qualité de l'appariement et de la recherche des enregistrements erronés. De plus, il s'agit de faire en sorte que la définition d'un appariement correct dans S_P et celle d'un enregistrement correct dans S_E soient identiques, *i.e.*, équilibre entre sur- et sous-couverture (« balancing »); voir la section 4. Tous ces éléments sont pris en compte dans la mesure du possible dans les présentes estimations.

L'estimation de la sous-couverture nette dans un domaine d a la forme $\hat{R}_{\text{sousnet},d} = 1 - \hat{R}_{\text{net},d} = 1 - C_d / \hat{N}_d$ avec C_d le nombre recensé dans le domaine et $\hat{N}_d$ l'estimation du vrai total. L'estimation du vrai total $\hat{N}_d$ est basée sur un modèle nommé *synthétique* qui suppose que le facteur de correction est fixe dans chaque cellule $k = 1, \ldots, K$:

$$\hat{N}_d = \sum_{k=1}^{K} \hat{N}_{k,d} = \sum_{k=1}^{K} C_{k,d}\, \hat{F}_k. \qquad (6)$$

$C_{k,d}$ est le nombre recensé dans la population d'intérêt dans l'intersection entre la cellule k et le domaine d et $\hat{F}_k$ est le facteur de correction de la couverture dans la cellule k. L'hypothèse du modèle synthétique est respectée si le comportement de tout sous-ensemble dans la cellule est identique à celui de la cellule entière. Cette homogénéité doit être contrôlée au mieux par le choix des cellules. On reprend ici les cellules homogènes définies pour le modèle dual.

3.3 Cellules d'estimation

Les cellules d'estimation $k = 1, \ldots, K$ sont construites de façon à grouper les éléments ayant des probabilités d'énumération homogènes dans le recensement, respectivement dans l'enquête, (hypothèse duale) et des taux de couverture nette homogènes (hypothèse synthétique). On désire un minimum de 100 personnes par cellule dans S_E et S_P afin de contrôler la variance et limiter le biais d'estimation. Les variables définissant les cellules sont sélectionnées à l'aide d'un modèle de régression logistique et d'une méthode de discrimination appliqués sur les données de S_P (variable binaire : P_j). Les trois variables les plus influentes sont croisées : nationalité en 2 catégories, état civil en 2 catégories et tailles de la commune en 3 catégories. Les autres variables sont ensuite intégrées successivement en faisant des regroupements lorsque les tailles de cellules sont trop petites (langue officielle de la commune en 2 catégories, classe d'âge en 7 catégories et sexe en 2 catégories). Au final, nous obtenons 121 cellules d'estimations; voir Renaud (2004) pour plus de détails.

3.4 Variance des estimateurs de couverture

La variance des estimateurs est estimée par un jackknife stratifié appliqué sur les unités primaires - identiques - de S_P et S_E. On note que la variance de la sous-couverture estimée $\hat{R}_{\text{sous}} = 1 - \hat{R}_m$ est égale à la variance du taux d'appariement estimé $\hat{R}_m$. De même la variance de la sur-couverture $\hat{R}_{\text{sur}} = 1 - \hat{R}_c$ est égale à celle du taux d'enregistrement correct $\hat{R}_c$ et la variance de la sous-couverture nette $\hat{R}_{\text{sousnet}} = 1 - \hat{R}_{\text{net}}$ est égale à celle du taux de couverture nette $\hat{R}_{\text{net}}$.

Soit θ le paramètre d'intérêt prenant la forme d'une moyenne pondérée de statuts dans le cas de la sous-couverture et de la sur-couverture, et la forme d'une fonction linéaire de quotients entre deux moyennes pondérées dans le cas de la sous-couverture nette. Son estimateur est $\hat{\theta}$.

Soient $h = 1, \ldots, H$ la strate utilisée au premier degré de l'échantillonnage, $i = 1, \ldots, m_h$ le numéro de l'unité primaire dans la strate h (numéro postal pour NORD ou commune pour TICINO), et $j = 1, \ldots, n_{hi}$ le numéro de la personne dans l'unité primaire i de h. Pour les besoins du jackknife, les échantillons S_P et S_E sont partitionnés, dans chaque strate h, en m_h sous-ensembles correspondant aux personnes dans les unités primaires $\alpha = 1, \ldots, m_h$.

Soit $\hat{\theta}_{(h\alpha)}$ l'estimateur ayant la même forme que $\hat{\theta}$ mais calculé sur l'échantillon auquel on a retiré l'unité primaire α de la strate h. On note que les estimateurs $\hat{R}_{m(h\alpha),k}$ et $\hat{R}_{c(h\alpha),k}$, $k = 1, \ldots, K$ sont combinés pour former $\hat{R}_{\text{net}(h\alpha)}$:

$$\hat{R}_{\text{net}(h\alpha)} = C\left[\sum_{k=1}^{K} C_k \frac{\hat{R}_{c(h\alpha),k}}{\hat{R}_{m(h\alpha),k}}\right]^{-1}. \qquad (7)$$

Les poids corrigés w'_{hij} utilisés pour le calcul des $\hat{R}_{m(h\alpha)}$ et $\hat{R}_{c(h\alpha)}$ ont la forme :

$$w'_{hij} = \begin{cases} 0 & \text{si} \quad i = \alpha \\ w_{hij}\dfrac{m_h}{m_h - 1} & \text{si} \quad \alpha \in h \text{ et } i \neq \alpha \\ w_{hij} & \text{si} \quad \alpha \notin h. \end{cases} \qquad (8)$$

Cette forme de correction est préférée au quotient entre la somme des poids des éléments dans la strate et la somme des poids sans l'unité primaire α car elle permet de prendre en compte la variabilité due au nombre inconnu d'éléments dans la strate.

L'estimateur du jackknife devient :

$$\hat{\theta}_{JK} = \frac{\sum_h \sum_{\alpha=1}^{m_h} \hat{\theta}_{h\alpha}}{\sum_h m_h}, \qquad (9)$$

avec les pseudo valeurs $\hat{\theta}_{h\alpha} = m_h \hat{\theta} - (m_h - 1)\hat{\theta}_{(h\alpha)}$. L'estimateur de sa variance peut prendre différentes formes; voir par exemple Shao and Tu (1995). Nous appliquons la forme suivante :

$$v(\hat{\theta}_{JK}) = \sum_h \frac{m_h - 1}{m_h} \sum_{\alpha=1}^{m_h} (\hat{\theta}_{(h\alpha)} - \hat{\theta}_{(h.)})^2, \qquad (10)$$

avec $\hat{\theta}_{(h.)} = \sum_{\alpha=1}^{m_h} \hat{\theta}_{(h\alpha)} / m_h$. Finalement, on utilise $v(\hat{\theta}_{JK})$ comme estimateur de la variance de $\hat{\theta}$. Les estimations dans des sous-groupes utilisent la même forme de l'estimateur avec l'intégration d'un indicateur de domaine dans la construction des $\hat{\theta}_{(h\alpha)}$. Aucune correction pour la population finie n'est appliquée dans les estimations. De plus on ne tient pas compte d'autres variabilités telle que celle induite par le modèle de pondération pour la non-réponse dans S_P.

Des problèmes, tels le manque de stabilité de l'estimation dans les strates avec peu d'unités primaires, sont apparus durant l'application. Les essais de partage de certaines unités primaires et une comparaison avec la linéarization de Taylor ou un jackknife simple permettent cependant de penser que les estimateurs de variance par le jackknife stratifié présentés dans ce document sont plutôt conservateurs.

4. Choix des statuts d'appariement et d'énumeration corrects

Un élément clé des estimations de couverture est la définition de *statut d'appariement correct* pour les éléments de S_P et du *statut d'énumération correct* pour les éléments de S_E. Ces statuts corrects sont définis à partir des statuts de bases P_j et E_j déterminés durant les appariements.

Un appariement avec un élément du recensement faisant partie d'un ménage collectif est-t-il accepté comme appariement correct pour un élément de S_P ou s'agit-il d'une sous-couverture de la population d'intérêt ? Un double hors de la population d'intérêt pour un élément de S_E est-il vraiment considéré comme un double, et donc une sur-couverture, ou devrait-il être exclu? Une définition claire s'impose. De plus, les statuts utilisés dans les estimations de la sous-couverture nette doivent être choisis de manière à satisfaire l'équilibre entre la sur- et la sous-couverture; voir la notion d'équilibrage ou de « balancing » par exemple dans Hogan (2003). Un appariement $(P_j = 1)$ avec un élément hors population d'intérêt peut par exemple être refusé comme appariement correct (statut d'appariement correct = 0, pas de sous-couverture) uniquement si la recherche des enregistrements corrects détecterait cet élément également comme incorrect car hors population (statut d'énumération correct = 0, pas de sur-couverture).

Les critères de définition des statuts corrects sont construits à partir des informations disponibles pour les éléments de S_P et S_E. Pour ce qui concerne S_P, nous partons de l'hypothèse que les enregistrements du recensement qui ont été appariés à des éléments de S_P permettent d'identifier les personnes (complétude) et que ces personnes devaient bien être recensées (pertinence). On estime aussi qu'ils sont uniques car l'unicité, bien que pas contrôlée pour les appariements, est réalisée dans la grande majorité des cas contrôlés dans S_E. Les critères d'appartenance à la population et de localisation sont contrôlés par comparaison avec les informations relevées dans l'enquête de couverture; considérées comme référence. Aucun relevé complémentaire n'était organisé pour régler les cas peu clairs. Pour ce qui concerne S_E, nous avons à disposition le critère de complétude considéré comme respecté dans les données du recensement et les résultats concernant l'unicité et la pertinence obtenus dans l'appariement avec le reste du recensement. Pour les doubles et triples, on définit $E_j = 1/d'$, avec d' = nombre de doubles/triples dans la population d'intérêt selon le recensement. Les critères d'appartenance à la population d'intérêt et de localisation des éléments de S_E. ne peuvent être contrôlés car nous n'avons pas de données de référence complémentaires au recensement.

Pour les estimations de sous-couverture nette, il est important de satisfaire la contrainte d'équilibrage. Les critères utilisés dans la définition des statuts corrects sont donc la complétude, la pertinence et l'unicité. Les critères de l'appartenance à la population d'intérêt et de la localisation ne peuvent pas être considérés car ils ne sont pas utilisables dans la définition du statut d'énumération correct. Les critères de complétude, pertinence et unicité sont déjà intégrés dans la construction des statuts de base P_j et E_j. On fait donc les estimations avec les statuts de base P_j et E_j.

Pour les estimations n'utilisant pas le système dual et le besoin d'équilibrage, il est possible d'utiliser d'autres critères pour définir les statuts corrects. D'autres types de statuts d'appariements corrects sont notamment utilisés dans l'analyse des erreurs potentielles de mesure de la section 5 et les analyses plus détaillées des appariements et des énumérations présentées dans Renaud (2004).

5. Comparaison des appariements

5.1 Erreurs potentielles de mesure

Les erreurs de mesures ou erreurs de classification sont liées aux erreurs de couverture. En effet, une personne classifiée dans le domaine d selon le recensement (par exemple personne entre 10 et 19 ans) alors qu'elle est en réalité hors du domaine (par exemple personne de 60 ans) aboutirait à une sur-couverture dans le domaine d et une sous-couverture hors de ce domaine. Cette erreur de classification (« misclassification ») ne provoque pas d'erreur de couverture au niveau global mais une erreur au niveau de sous-groupes de la population.

Les raisons des différences entre les valeurs relevés dans deux enquêtes telles que le recensement et l'enquête de couverture peuvent être très diverses et complexes à dissocier. Il faut en effet compter sur des erreurs d'appariement, des différences résultant des méthodes de relevé (questionnaire papier ou téléphone/face-à-face) et du traitement des données, ou encore sur des différences réelles dues au décalage temporel entre les relevés (décembre 2000 ou avril-mai 2001). De plus, il est difficile de déterminer la réponse correcte si on a deux valeurs différentes : le recensement ? l'enquête ? une autre valeur pas relevée ?

Les erreurs de mesures potentielles des données du recensement sont analysées sur la base de l'ensemble des appariements entre l'échantillon indépendant S_P et le recensement. On choisit de déterminer quelles sont les variables qui montrent respectivement peu ou beaucoup de problèmes potentiels de classification, sans faire un jugement sur la qualité de l'un ou l'autre des relevés. Ces informations sont notamment utiles pour évaluer le choix des cellules d'estimations pour le système dual et choisir les sous-groupes pour lesquels les estimations des défauts de couverture sont les plus fondées.

On définit, pour la variable catégorielle X, le taux d'appariement dans le bon domaine $\hat{R}_X$ comme suit :

$$\hat{R}_X = \frac{\sum_{j \in S_P \text{ match}} w_{P,j} P_{X,j}}{\sum_{j \in S_P \text{ match}} w_{P,j}}, \qquad (11)$$

avec $w_{P,j}$ le poids de l'élément j de l'échantillon S_P apparié (S_P match) et le *statut de classification* $P_{X,j}$ qui vaut 1 si l'élément j se trouve dans la même classe dans le recensement et dans l'enquête, et 0 sinon. La valeur de $\hat{R}_X$ est estimée avec l'ensemble des éléments appariés et avec le sous-groupe des éléments sans imputation dans le recensement.

On définit également une mesure d'asymétrie $\varphi_X(d,\ d')$ pour les classes d et d' de la variable X :

$$\varphi_X(d,\ d') = \frac{\sum_{j \in S_P \text{ match}} w_{P,j} I_j(d,\ d')}{\sum_{j \in S_P \text{ match}} w_{P,j} I_j(d',\ d)}, \qquad (12)$$

avec $I_j(d,\ d') = 1$ si l'élément j se trouve dans le domaine d selon l'enquête et dans le domaine d' selon le recensement, et 0 sinon. Le facteur $\varphi_X(d,\ d')$ vaut 1 s'il y a équilibre dans les erreurs de classification, c'est-à-dire si le nombre d'éléments dans d selon l'enquête et dans d' selon le recensement est égal au nombre dans d' selon l'enquête et dans d selon le recensement. Plus il s'éloigne de 1, plus l'équilibre est rompu.

5.2 Erreurs potentielles de localisation

Les comparaisons entre le recensement et l'enquête permettent également d'étudier la localisation géographique des personnes. Dans les données du recensement, nous avons une unique adresse si la personne a un unique domicile et deux adresses - principale et secondaire - si la personne a deux domiciles. Dans les données de l'enquête, nous avons une ou deux adresses au jour du recensement, une ou deux adresses au jour de l'enquête et l'information sur un éventuel déménagement entre les deux dates. Si une personne a un unique domicile et n'a pas déménagé, son adresse principale au jour du recensement et son adresse principale au jour de l'enquête sont identiques. Elle n'a pas d'adresses secondaires.

Différentes mesures de distance sont envisagées pour déterminer les erreurs potentielles de localisation dans le recensement. Pour des raisons pratiques, notamment de données à disposition, nous définissons des aires géographiques autour de l'adresse principale de la personne telle que relevée durant l'enquête pour le jour du recensement (*adresse de référence*). Les aires sont des ensembles de communes politiques. Elles sont définies à partir des numéros postaux relevés durant l'enquête. L'*aire de base* de la personne est définie par l'ensemble des communes qui possèdent des bâtiments dans le numéro postal de son adresse de référence. La définition de cette aire utilise les données du registre suisse des bâtiments car ce dernier possède une information sur l'adresse postale et la commune des bâtiments. L'*aire étendue* comporte les communes de l'aire de base et l'ensemble des communes qui leurs sont adjacentes; voir Renaud (2004) pour des exemples.

De façon similaire aux erreurs de classification, les erreurs de localisation ne provoquent pas d'erreurs de couverture au niveau global mais des erreurs au niveau de sous-groupes tels des régions ou des types de communes. Différents taux peuvent être définis. On retiendra le taux de localisation de base et le taux de localisation étendue; tous deux pondérés par w_{Pj} le poids de l'élément j de l'échantillon S_P apparié. Le statut de localisation prend la valeur 1 si l'élément est trouvé dans l'aire de base ou respectivement étendue, et 0 sinon. On étudiera en particulier la localisation des personnes qui ont déménagé, afin

de détecter d'éventuels problèmes de décalage temporel entre le jour du recensement et le jour du relevé effectif des données du recensement.

6. Résultats

6.1 Estimations des défauts de couverture

Le taux global de sous-couverture nette est estimé à 1,41 % avec un écart-type de 0,12 %. Le taux de sur-couverture est de 0,35 % (écart-type = 0,03 %) et le taux de sous-couverture est de 1,64 % (écart-type = 0,11 %). Ces résultats sont dans l'ordre de grandeur de ceux d'autres pays, bien que plutôt dans les valeurs les plus basses; voir Tableau 1.

La sur-couverture est peu importante dans la grande majorité des domaines étudiés. Le plus haut taux est observé pour les personnes entre 20 et 31 ans (0,93 % avec un écart-type de 0,09 %); voir Tableau 2. La sous-couverture est par contre élevée dans plusieurs domaines. On note par exemple un taux de 8,03 % (écart-type = 0,85 %) pour les étrangers avec des permis d'établissement temporaires (« autres permis ») et un taux de 3,50 % (écart-type = 0,50 %) pour les 20-31 ans. On note aussi un taux de sous-couverture de 2,4 % dans la région italophone du pays (langue de la commune : italien, région NUTS : Ticino, et relevé : TICINO). Les résultats sont cependant liés à une relativement grande variabilité (écart-type de env. 0,5 %) car les échantillons S_P et S_E ne comportent respectivement que 1 500 et 1 700 personnes dans cette région.

Tableau 1
Comparaison internationale des résultats globaux. Taux estimés de sur-couverture $\hat{R}_{sur}$, sous-couverture $\hat{R}_{sous}$ et sous-couverture nette $\hat{R}_{sousnet}$, avec écarts-type estimés correspondants. Références : Statistique Canada (1999, 2004), Hogan (1993, 2003), McLennan (1997) et Trewin (2003)

		Sur-couverture [%]	Sous-couverture [%]	Sous-couverture nette [%]
Suisse	2000	0,3 (0,03)	1,64 (0,11)	1,41 (0,12)
Canada	1996	0,74 (0,04)	3,18 (0,09)	2,45 (0,10)
	2001	0,96 (0,05)	3,95 (0,13)	2,99 (0,14)
États-Unis	1990	3,1	4,7	1,6 (0,10)
	2000	-	-	-0,5 (0,21)
Australie	1996	0,2	1,8	1,6 (0,10)
	2001	0,9	2,7	1,8 (0,10)

La sous-couverture nette est positive dans tous les domaines étudiés. Il n'y a donc pas de sur-couverture nette. Les plus grandes valeurs sont observées pour les étrangers avec permis permanent ou temporaire (2,89 % et 3,48 %, écarts-type = 0,32 % et 0,39 %) ainsi que chez les 20-31 ans (2,84 %, écart-type = 0,36 %). Aucune différence significative n'est observée entre hommes et femmes, entre langues et entre régions NUTS. La faible taille de l'échantillon avec la variante de relevé TICINO ne permet pas de différencier cette méthode des autres utilisées dans le pays. Des différences sont par contre significatives entre états civils, ainsi qu'entre les types et tailles de communes.

On note que le taux de sous-couverture nette est supérieur au taux de sous-couverture dans le cas des permis d'établissement permanent. Cet effet, irréaliste, est dû au choix des cellules d'estimations et au lissage qui s'en suit. La construction des cellules a en effet nécessité un regroupement des étrangers avec permis permanents et temporaires en une seule catégorie lors des agrégats permettant d'atteindre la taille minimale de 100 personnes par cellule. Par ce regroupement nous considérons les étrangers comme un groupe homogène alors qu'il ne l'est pas. Ceci montre les limites de la méthode et la difficulté de satisfaire aux hypothèses des modèles utilisés lors de l'application. Dans le cas des étrangers, on note cependant que les intervalles de confiance des taux de sous-couverture nette et de sous-couverture se recoupent. Les conséquences des faiblesses de l'application sont donc restreintes.

Notons encore que les résultats sont présentés dans des domaines définis par des variables pour lesquelles on a observé de faibles erreurs potentielles de mesure. Des résultats pour des groupes tels que définis par les caractéristiques de ménage ou de vie active ne seraient en effet que peu fiables; voir section 6.2.

La précision des résultats obtenus est en général meilleure que l'objectif fixé au début du projet. Ce dernier était en effet d'avoir un écart-type de 0,3 % pour des sous-groupes de 10 000 individus dans S_P. Dans le cas par exemple des classes d'âge 32-44 et 45-59 qui comportent entre 10 000 et 12 000 personnes les écart-types sont 0,19 et 0,14 %.

6.2 Erreurs potentielles de mesure et de classification

Parmi les 49 107 éléments appariés entre l'enquête de couverture et le recensement, 96 % ne présentent aucune différence dans le sexe, les 7 classes d'âge, les 3 classes d'état civil et les 3 classes de permis d'établissement (suisse, permanent, temporaire). Le taux d'appariement dans le bon domaine $\hat{R}_X$ vaut 99,3 % pour le sexe (avec et sans imputations), 98,3 % pour l'état civil (98,4 % parmi les valeurs non imputées) et 98,7 % pour le permis d'établissement (98,8 % parmi les valeurs non imputées). Le taux $\hat{R}_X$ vaut 99,5 % pour les classes d'âge (avec ou sans imputations). Il faut cependant noter que la date de naissance était, avec le nom et le prénom, une des principales variables dans l'appariement. Les différences d'âge sont donc possibles uniquement lors d'un appariement non automatique (assisté par ordinateur ou manuel). Trois variables montrent un taux d'appariement dans le bon domaine nettement plus

faible que celui observé pour le sexe, l'âge, le permis et l'état civil. Il s'agit de la variable sur la vie active (actif, sans emploi, non actif), la position dans le ménage (seul/seule, époux/épouse, union libre, une personne avec enfant(s), autre chef de ménage, apparenté au chef de ménage, autres; résultats restreints aux ménages privés) et la taille du ménage de la personne (selon domicile économique et dans les ménages privés). Le taux $\hat{R}_X$ vaut 90,4 % pour la vie active (91,1 % parmi les valeurs non imputées), 91,4 % pour la position dans le ménage (94,9 % parmi les valeurs non imputées) et 88,3 % pour la taille du ménage.

Tableau 2 Nombre recensé *C* et taux estimés de sur-couverture $\hat{R}_{sur}$, sous-couverture $\hat{R}_{sous}$ et sous-couverture nette $\hat{R}_{sousnet}$ pour différents domaines [%], avec les écart-types estimés (*E-T*) correspondants

Variable	Catégories	*C*	$\hat{R}_{sur}$	E-T	$\hat{R}_{sous}$	E-T	$\hat{R}_{sousnet}$	E-T
global		7 121 626	0,35	0,03	1,64	0,11	1,41	0,12
sexe	homme	3 497 940	0,37	0,04	1,74	0,13	1,46	0,13
	femme	3 623 686	0,33	0,03	1,55	0,10	1,37	0,13
classe âge	≤ 9	810 373	0,26	0,05	1,46	0,21	1,34	0,26
	10-19	833 185	0,27	0,05	1,30	0,19	1,04	0,22
	20-31	1 115 804	0,93	0,09	3,50	0,34	2,84	0,36
	32-44	1 544 721	0,33	0,05	1,65	0,16	1,43	0,19
	45-59	1 431 771	0,22	0,04	1,18	0,14	1,04	0,14
	60-79	1 146 709	0,10	0,03	0,91	0,13	0,82	0,12
	≥ 80	239 063	0,11	0,06	1,20	0,31	1,03	0,27
permis d'établissement	Suisse	5 674 266	0,33	0,03	1,28	0,09	0,98	0,10
	étranger permanent	1 020 242	0,33	0,06	1,85	0,29	2,89	0,32
	étranger temporaire	427 118	0,56	0,11	8,03	0,85	3,48	0,39
état civil	célibataire	2 975 643	0,50	0,05	2,07	0,18	1,72	0,19
	marié/e	3 377 223	0,23	0,04	1,27	0,11	1,25	0,12
	veuf/veuve	369 339	0,25	0,08	1,23	0,26	0,79	0,13
	divorcé/e	399 421	0,24	0,08	1,95	0,35	1,02	0,10
langue commune	allemand + romanche	5 128 353	0,33	0,04	1,50	0,11	1,28	0,12
	français	1 680 062	0,35	0,06	1,89	0,25	1,79	0,27
	italien	313 211	0,53	0,12	2,35	0,49	1,56	0,19
région NUTS	région lémanique	1 296 464	0,37	0,07	2,19	0,38	1,84	0,28
	espace Mittelland	1 640 489	0,35	0,09	1,39	0,15	1,25	0,10
	Nordwestschweiz	976 699	0,18	0,04	1,50	0,27	1,32	0,12
	Zurich	1 221 014	0,31	0,05	1,58	0,19	1,46	0,13
	Ostschweiz	1 020 897	0,40	0,07	1,29	0,23	1,24	0,12
	Zentralschweiz	665 904	0,36	0,06	1,57	0,25	1,19	0,12
	Ticino	300 159	0,54	0,12	2,38	0,52	1,57	0,19
taille commune	petite	1 372 958	0,34	0,05	1,50	0,15	1,12	0,14
	moyenne	2 398 256	0,41	0,07	1,32	0,16	1,07	0,19
	grande	3 350 412	0,31	0,03	2,01	0,19	1,77	0,19
type	ville	2 078 780	0,35	0,04	1,96	0,17	1,82	0,20
	agglomération	3 145 541	0,36	0,06	1,49	0,19	1,34	0,12
	rural	1 897 305	0,32	0,04	1,56	0,17	1,07	0,12
relevé	CLASSIC	265 607	0,39	0,05	1,91	0,28	1,07	0,12
	SEMI-CLASSIC	174 501	0,37	0,08	1,07	0,24	1,16	0,13
	TRANSIT + FUTURE	6 381 359	0,33	0,03	1,62	0,11	1,42	0,12
	TICINO	300 159	0,54	0,12	2,38	0,52	1,57	0,19

La mesure de l'asymétrie $\varphi_X(d, d')$ prend la valeur 1,33 pour le sexe (d = homme et d' = femme). Il y a plus de personnes codées hommes selon l'enquête qui sont codées femmes dans le recensement que de femmes selon l'enquête qui sont des hommes selon le recensement. La proportion d'hommes est légèrement supérieure dans l'enquête. Ces résultats doivent cependant être relativisés car ils sont basés sur très peu de cas; voir le Tableau 3. Un test de McNemar est juste significatif au seuil de 5 % sans tenir compte du plan, mais il ne le reste pas si le plan est pris en compte. On observe des asymétries par contre très nettes dans l'état civil. Il y a moins de célibataires dans l'enquête qui sont mariés dans le recensement que l'inverse (facteur 0,33 pour d = célibataire et d' = marié). De même, il y a moins de mariés dans l'enquête qui sont veuf/veuve ou divorcé/e dans le recensement que l'inverse (facteur 0,42 pour d = marié et d' = autres). L'asymétrie est également visible sur la variable du permis d'établissement. La tendance est d'avoir plus de Suisses dans l'enquête qui sont notés étrangers dans le recensement que l'inverse; de même plus de permis permanents dans l'enquête et permis temporaires dans le recensement que l'inverse (facteurs 5,22 pour d = suisse et d' = étranger avec permis permanent et 3,83 pour d = étranger avec permis permanent et d' = étranger avec permis temporaire). Les facteurs calculés se basent sur peu de cas. Ils donnent cependant un aperçu des différences potentielles entre un relevé par le biais du questionnaire du recensement et une enquête réalisée principalement par téléphone. La variable de la vie active comporte plus de cas divergents; voir le Tableau 4. Nous avons donc par exemple moins de personnes occupées dans l'enquête et non actives dans le recensement que l'inverse (facteur de 0,46 pour d = actif et d' = non actif). Il y a également moins de personnes sans emploi dans l'enquête et non actives dans le recensement que l'inverse (facteur 0,26 pour d = sans emploi et d' = non actif). La variable de la position dans le ménage montre également des asymétries, mais ces dernières sont basées sur peu d'éléments car la dispersion des éléments dans les cases (d, d') est importante. Les variables du recensement au niveau ménage (position dans le ménage et taille de ménage) sont influencées par le processus complexe de la formation des ménages. Elles sont moins fiables que celles qui concernent les personnes. Les valeurs au niveau ménage sont plus fiables dans l'enquête.

Tableau 3 Comparaison des valeurs relevées dans l'enquête et le recensement pour la variable sexe

sexe			Enquête		
			homme	femme	total
Recensement	non apparié	total	393	383	776
	apparié	total	24 171	24 936	49 107
	apparié	homme	23 967	166	24 133
		femme	204	24 770	24 974
	apparié	homme	6	0	6
	(variable imputée)	femme	0	13	13
	total		24 564	25 319	49 883

Tableau 4 Comparaison des valeurs relevées dans l'enquête et le recensement pour la variable vie active

vie active			Enquête				
			occupé	sans emploi	non actif	≤ 15 ans	total
Recensement	non apparié	total	424	23	217	112	776
	apparié	total	25 163	498	14 501	8 945	49 107
	apparié	occupé	23 953	188	2 007	13	26 161
		sans emploi	300	221	323	1	845
		non actif	901	89	12 143	18	13 151
		≤ 15 ans	9	0	28	8 913	8 950
	apparié	occupé	564	22	312	6	904
	(variable imputée)	sans emploi	14	8	26	1	49
		non actif	92	15	881	5	993
		≤ 15 ans	0	0	0	0	0
	total		25 587	521	14 718	9 057	49 883

6.3 Erreurs potentielles de localisation et décalage temporel

Parmi les 49 107 éléments appariés entre l'enquête de couverture et le recensement, 97,7 % sont trouvés dans l'aire de base autour de l'adresse de référence relevée durant l'enquête. Cette valeur vaut 98,1 % pour les personnes n'ayant indiqué aucun déménagement entre le jour du recensement et le jour de l'enquête. Elle vaut 83,9 % pour ceux qui ont indiqué un déménagement (1 512 personnes); voir nombres absolus dans le Tableau 5.

Il est intéressant de noter que 9,4 % des personnes du NORD n'ayant pas déménagé sont retrouvées proche de leur adresse de référence mais pas exactement dans le même bâtiment. Ces problèmes de localisation fine ont une influence négligeable sur les données du recensement. Ils montrent cependant la difficulté à identifier les bâtiments échantillonnés lors de la construction des listes des ménages sur le terrain durant l'enquête, tout comme la difficulté de l'assignation des personnes aux bâtiments durant le traitement des données du recensement. Un relevé complémentaire serait cependant nécessaire pour évaluer l'effet respectif des deux difficultés.

La localisation des personnes qui ont déménagé indique que $151 = 145 + 6$ personnes sont trouvées autour de leur adresse au jour de l'enquête et non pas autour de leur adresse au jour du recensement (9 %, pondéré). De plus, un ensemble de 688 personnes du NORD, parmi les 922 se trouvant dans les deux aires de base, sont en fait trouvées dans le bâtiment au jour de l'enquête. Un soin spécial ayant été mis durant l'enquête de couverture sur les questions concernant les adresses au jour du recensement et au jour de l'enquête, on considère ici que les adresses des personnes ayant déménagé sont de meilleure qualité dans les données de l'enquête que dans celles du recensement. Sur cette base, on déduit qu'au moins $151 + 688 = 839$ personnes ayant déménagé sur 1 512 sont énumérés dans le recensement à une adresse qu'ils n'avaient pas au jour officiel du relevé mais quelques temps après ce jour. Le décalage exact n'est pas connu car la date du déménagement n'a pas été relevée dans l'enquête.

7. Conclusion

Les défauts de couverture globaux du recensement de la population de l'an 2000 en Suisse sont dans l'ordre de grandeur des recensements dans d'autres pays. Des spécificités apparaissent cependant au niveau des sous-groupes (par exemple régions). Parmi les trois composantes, celle de la sous-couverture est fort intéressante car elle détecte non seulement des groupes de personnes plus ou moins bien recensés mais permet également d'analyser les erreurs de localisation et de mesures. Les estimations de sur-couverture sont de leur côté limitées par le manque d'informations complémentaires au recensement pour S_E. Elles pourraient être améliorées dans le futur par la collecte d'informations complémentaire sur les caractéristiques au jour du recensement lors d'une enquête auprès des personnes de cet échantillon (par exemple localisation et type de ménage). Les estimations de la sous-couverture nette sont basées sur plusieurs hypothèses. Les résultats dans de grands domaines semblent fiables mais certains risques, liés notamment au choix des cellules d'estimations, existent lorsque les domaines sont plus petits. Pour de futures estimations, on propose d'évaluer l'approche modèle tel qu'appliqué au Royaume-Uni au lieu des cellules d'estimations utilisées traditionnellement aux États-Unis.

Tableau 5
Comparaison de la localisation des personnes appariées. Les aires sont définies pour l'adresse au jour du recensement (selon information relevée dans l'enquête) et pour l'adresse au jour de l'enquête (également selon information relevée dans l'enquête). Présence dans l'aire de base, l'aire étendue (hors aire de base) ou hors de l'aire étendue pour les personnes n'ayant pas déménagé (fixes) et les personnes ayant déménagé (déménagements) entre le recensement et l'enquête

		Jour enquête				
		fixes	déménagements			
		base	base	étendue	hors étend.	total
Jour	base	46 689	922	69	277	1 268
recensement	étendue	258	42	4	3	49
	hors étendue	648	145	6	28	179
	manquant	0	15	1	0	16
	total	47 595	1 124	80	308	1 512

Un élément important à revoir pour de futures estimations est le choix de la population d'intérêt. Le choix de se limiter aux personnes dans des ménages privés et au domicile économique a provoqué quelques difficultés dans les estimations car une délimitation précise de cette population était difficile. Une future estimation pourrait exclure les ménages collectifs pour éviter les problèmes pratiques de relevé mais conserver tous les types de domiciles. L'ensemble des enregistrements au domicile économique serait alors traité comme un domaine.

L'estimation des défauts de couverture d'un recensement est un projet ambitieux qui a montré son intérêt. Les résultats donnent des informations sur la qualité des données du recensement 2000 et les différents problèmes de couverture. Les prochains recensements seront essentiellement basés sur des registres. Les estimations de couverture se baseront sur l'expérience acquise lors des estimations de 2000 avec de probables adaptations pour tenir compte du nouveau système de relevé.

Remerciements

Je tiens à remercier Philippe Eichenberger du Service de méthodes statistiques de l'Office fédéral de la statistique pour les discussions fructueuses durant tout le projet. Un grand merci également à Dr. Rajendra Singh et ses collègues du Decennial Statistical Studies Division du U.S. Census Bureau pour leur assistance lors du développement des méthodes et des estimations. Merci également à toutes les personnes du recensement qui ont effectués des travaux et fourni des informations pour le bon déroulement du projet, et à Paul-André Salamin du Service de méthodes statistiques pour la relecture attentive du papier.

Bibliographie

ABS (1997). The 1996 census of population and housing. Rapport annuel 1996-97, Australian Bureau of Statistics.

Brown, J.J., Diamond, I.D., Chambers, R.L., Buckner, L.J. et Teague, A.D. (1999). A methodological strategy for a one-number census in the UK, *Journal of the Royal Statistical Society*, Série A, 162(2), 247-267.

Fienberg, S.E. (1992). Bibliographie sur la modélisation à l'aide de la saisie-resaisie avec application au redressement des chiffres du recensement pour éliminer le sous-dénombrement. *Techniques d'enquête*, 18, 1, 157-169.

Hogan, H. (1993). The post enumeration survey: Operation and results. *Journal of the American Statistical Association*, 88(423), 1047-1060.

Hogan, H. (2003). L'évaluation de l'exactitude et de la couverture : théorie et conception. *Techniques d'enquête*, 29, 2, 145-156.

McLennan, W. (1997). Census of Population and Housing, Data Quality - Undercount. Autralia 1996. Article d'information, 2940.0. Australian Bureau of Statistics.

Renaud, A. (2001). Methodology of the Swiss Census 2000 Coverage Survey. *Proceedings of the Survey Research Methods Section* [CD-ROM], American Statistical Association.

Renaud, A. (2003). Estimation de la couverture du recensement de la population de l'an 2000. Échantillon pour l'estimation de la surcouverture (E-sample). Rapport de méthodes, 338-0019, Office fédéral de la statistique.

Renaud, A. (2004). Coverage estimation for the Swiss population census 2000. Estimation methodology and results. Rapport de méthodes, 338-0027, Office fédéral de la statistique.

Renaud, A., et Eichenberger, P. (2002). Estimation de la couverture du recensement de la population de l'an 2000. Procédure d'enquête et plan d'échantillonnage de l'enquête de couverture. Rapport de méthodes, 338-0009, Office fédéral de la statistique.

Renaud, A., et Potterat, J. (2004). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sous-couverture (P-sample) et qualité du cadre de sondage des bâtiments. Rapport de méthodes, 338-0023, Office fédéral de la statistique.

Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics.

Statistique Canada (1999). Couverture. Rapport technique du recensement de 1996, 92-370-XIF, Statistique Canada.

Statistique Canada (2004). Couverture. Rapport technique du recensement de 2001, 92-394-XIF, Statistique Canada.

Trewin, D. (2003). Census of Population and Housing, Data Quality - Undercount. Autralia 2001. Article d'information, 2940.0. Australian Bureau of Statistics.

Wolter, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81(394), 338-346.

Techniques d'enquête, décembre 2007
Vol. 33, N° 2, pp. 233-238
Statistique Canada, N° 12-001-X au catalogue

Utilisation d'un échantillon de convenance électronique comme complément à un échantillon probabiliste

Marc N. Elliott et Amelia Haviland [1]

Résumé

Dans le présent document, nous décrivons une méthodologie utilisée pour combiner un échantillon de convenance avec un échantillon probabiliste afin de produire un estimateur ayant une erreur quadratique moyenne (EQM) plus faible que les estimateurs fondés uniquement sur un échantillon probabiliste. Nous examinons ensuite les propriétés de l'estimateur composite obtenu, qui est en fait une combinaison linéaire des estimateurs de l'échantillon de convenance et de l'échantillon probabiliste, les poids étant fonction du biais. Nous discutons des propriétés de l'estimateur dans le contexte de l'échantillonnage de convenance électronique. Notre analyse démontre que le recours à un échantillon de convenance pour suppléer un échantillon probabiliste en vue d'améliorer l'EQM de l'estimation pourrait s'avérer utile seulement dans des circonstances restreintes. Premièrement, le biais résiduel de l'estimateur fondé sur l'échantillon de convenance doit être très faible, représentant tout au plus 0,1 de l'écart-type de la population obtenue. En cas de résultat dichotomique, cela signifie un biais ne dépassant pas cinq points de pourcentage à 50 % de prévalence, et trois points de pourcentage à 10 % de prévalence. Deuxièmement, l'échantillon probabiliste devrait contenir au moins 1 000 à 10 000 observations pour donner lieu à une estimation adéquate du biais de l'estimateur de l'échantillon de convenance. Troisièmement, il doit être rentable et faisable de recueillir au moins des milliers (et probablement des dizaines de milliers) d'observations à partir de l'échantillon électronique de convenance. Les conclusions au sujet de l'utilité limitée des échantillons de convenance lorsque le biais de l'estimateur comporte un écart-type de plus de 0,1 s'appliquent également à l'utilisation directe des estimateurs en fonction de cet échantillon.

Mots clés : Biais; estimateur composite; calage.

1. Introduction

Les enquêtes en ligne (électroniques) se font de plus en plus nombreuses et diversifiées (Couper 2000). Par exemple, les échantillons probabilistes électroniques s'appuient sur une base de sondage traditionnelle et offrent la déclaration par Internet parmi d'autres modes de déclaration ou comme unique méthode. Les échantillons probabilistes électroniques peuvent avoir des taux de réponse élevés et produire des estimateurs comportant un biais de non-réponse minime (Kypri, Stephenson et Langley 2004). En revanche, les échantillons de convenance électroniques sont fondés sur les requêtes « internes » de pages Web lancées par n'importe quel internaute qui découvre le site et décide de participer (parfois en conséquence de publicité ciblant une population qui n'est pas spécifiable), ou encore, ils dépendent de la bonne volonté de panels recrutés qui ne sont pas nécessairement représentatifs de la population visée.

Le principal attrait des échantillons de convenance électroniques réside dans les coûts marginaux par cas potentiellement très faibles. Les visites d'un site Web ne nécessitent pas de ressources humaines (contrairement aux appels téléphoniques) ou matérielles (comme dans le cas du publipostage) coûteuses, sans compter qu'elles favorisent la rapidité de la collecte de données et la réduction des coûts marginaux par cas du traitement des données. Même si l'on tient compte de quelques frais fixes, les coûts totaux par cas devraient être très faibles, surtout dans le cas des grandes enquêtes. Par ailleurs, l'inconvénient de ces échantillons ressort clairement : un biais de sélection potentiellement prononcé et non mesuré.

À notre connaissance, la plupart de la documentation sur les échantillons de convenance électroniques soit soutient que les échantillons probabilistes sont sans importance en général, soit tente de délimiter les circonstances dans lesquelles les échantillons de convenance pourraient être utiles, soit écarte complètement l'utilisation des échantillons de convenance. Nous explorons une tout autre avenue en envisageant la possibilité d'intégrer les échantillons de convenance électroniques dans le contexte de l'échantillonnage probabiliste.

Dans le présent document, nous décrivons une méthodologie qui consiste à combiner un échantillon de convenance avec un échantillon probabiliste, en vue de produire un estimateur comportant une erreur quadratique moyenne (EQM) plus faible que celle des estimateurs fondés uniquement sur l'échantillon probabiliste. Ensuite, nous examinons les propriétés de l'estimateur composite obtenu, une combinaison linéaire des échantillons de convenance et probabiliste, les poids étant déterminés par le biais. Cette démarche nous amène à formuler des recommandations au sujet de l'utilité d'ajouter des échantillons de convenance électroniques aux échantillons probabilistes. Compte tenu des coûts différentiels très

1. Marc Elliott, Ph.D. et Amelia Haviland, Ph.D., Rand Corporation, 1776 Main Street, Santa Monica, CA 90407, États-Unis.

faibles des échantillons de convenance électroniques, nous nous efforçons de repérer les situations où l'augmentation de la taille effective de l'échantillon (TEE) attribuable à l'inclusion de l'échantillon de convenance pourrait être suffisante pour justifier une approche à deux modes. Nous démontrons que les circonstances dans lesquelles un échantillon de convenance électronique supplémentaire pourrait améliorer l'EQM de façon significative sont très limitées. Bien que nous nous attardions sur les échantillons de convenance électroniques, la discussion qui suit s'applique également à d'autres méthodes de collecte de données peu coûteuses à faible couverture de population.

2. Problématique

2.1 Conditions initiales

Dans le cas de l'échantillon combiné probabiliste/de convenance, nous proposons de mener simultanément la même enquête auprès d'un échantillon probabiliste traditionnel (avec ou sans mode de déclaration électronique) et d'un échantillon de convenance électronique. Nous imaginons une enquête polyvalente menant à divers résultats. Dans le présent document, nous nous attarderons sur l'estimation des moyennes, mais il se pourrait que les travaux à venir appliquent ces résultats à d'autres paramètres, tels que les paramètres de régression. Au départ, nous examinerons les cas où le biais des estimations de l'échantillon de convenance est connu, mais nous évaluerons par la suite dans quelle mesure l'échantillon probabiliste permet de mesurer le biais inconnu dans chaque estimation des paramètres de l'échantillon de convenance.

Lorsque le biais est connu, on peut combiner les échantillons de convenance et probabiliste de manière à réduire l'EQM. Si les estimations de l'échantillon de convenance sont fortement biaisées, l'échantillon de convenance ne servira pas à grand-chose. Cette possibilité nécessite que l'échantillon probabiliste soit assez grand pour être considéré par lui-même. Par conséquent, une approche possible serait de réserver une petite partie du budget de l'échantillon probabiliste pour créer un gros échantillon de convenance supplémentaire.

Prenons par exemple le cas d'une enquête qui s'intéresse principalement aux estimations de l'ensemble de la population, mais pour laquelle il serait également souhaitable d'obtenir des estimations de sous-populations s'il était rentable d'obtenir un échantillon d'une taille suffisante pour assurer la précision des résultats. Supposons en outre qu'on pourrait faire 4 000 observations de probabilité et 10 000 observations de convenance pour le coût d'un échantillon probabiliste de 5 000. Pour un résultat donné, si le biais est prononcé, les erreurs-types augmentent modérément en raison d'une faible réduction proportionnelle de la taille de l'échantillon; si le biais est généralement petit et propre à chaque sous-population, il pourrait y avoir des retombées positives sur la précision, ce qui pourrait donner lieu à des analyses de sous-populations d'une précision acceptable.

2.2 Réduction du biais initial

Nous ferons la preuve que le biais des estimateurs de l'échantillon de convenance doit être très faible pour que l'échantillon soit utile, ce qui indique qu'il pourrait être préférable de s'attarder sur les estimations de paramètres qui sont habituellement moins biaisés que les estimations des proportions ou des moyennes de la population globale non ajustées, telles que les coefficients de régression (Kish 1985).

En outre, on pourrait réduire le biais grâce au calage sur marges de l'échantillon de convenance en fonction des valeurs connues de la population (Kalton et Kasprzyk 1986), ou en appliquant les poids des scores de propension qui modélisent la sélection des membres entre les deux échantillons aux observations de l'échantillon de convenance (Rosenbaum 2002; Rosenbaum et Rubin 1983). On peut inclure une petite série d'éléments pour permettre l'utilisation d'une de ces approches. Parmi ces éléments figurent aussi bien ceux qui prédisent les différences entre les répondants qui fournissent leurs données par voie électronique et ceux qui optent pour d'autres modes de déclaration, que les éléments adaptés au contenu de l'enquête en particulier. L'effet du plan de sondage des poids des variables obtenus réduiront la TEE pour les estimateurs de l'échantillon de convenance, mais les faibles coûts de ces observations rentabilisent la compensation des effets modérés du plan de sondage. Nous pouvons ensuite estimer le biais résiduel d'un paramètre donné, comme la différence entre l'estimation de l'échantillon probabiliste et l'estimation pondérée de l'échantillon de convenance.

3. Facteurs d'efficacité

3.1 Combinaisons linéaires des estimateurs biaisés et sans biais d'une moyenne de population

Un cas spécial d'un estimateur donné dans un résultat par Rao (2003, pages 57-58) constitue l'estimateur le plus efficace des combinaisons linéaires d'échantillons de convenance et probabiliste (pondérés). Les propriétés de cet estimateur mènent à des recommandations générales au sujet des conditions de la taille de l'échantillon probabiliste, de la taille de l'échantillon de convenance et du biais de l'estimateur de l'échantillon de convenance qui permettent à l'échantillon de convenance d'améliorer significativement la TEE de l'échantillon probabiliste.

Nous commençons en nous posant la question suivante : Quel est l'estimateur de ce type le plus efficace lorsqu'on connaît l'ampleur du biais? Nous envisagerons ensuite la possibilité d'assouplir la supposition relative à la connaissance de l'ampleur du biais.

Supposons que n_1 et n_2 correspondent respectivement à la taille effective de l'échantillon probabiliste et à celle de l'échantillon de convenance, après division de la taille nominale des échantillons par les effets du plan de sondage associés au plan d'échantillonnage et au redressement de la non-réponse. Il faut tenir compte du score de propension ou de tout autre pondération dans le cas de l'échantillon de convenance. La première population a une moyenne de μ, et une variance de σ_1^2; la deuxième a une moyenne de $\mu + \varepsilon$ et une variance de σ_2^2, où ε représente le biais connu qui reste après la pondération, et μ correspond au paramètre inconnu qui nous intéresse. Les moyennes des échantillons correspondants ont une espérance μ et $\mu + \varepsilon$ et une variance σ_i^2 / n_i où $i = 1, 2$ en supposant un modèle d'échantillonnage fondé sur une population infinie. Nous présumons que ces deux estimateurs ne sont pas corrélés, étant donné qu'ils proviennent d'échantillons indépendants.

D'après Rao (2003, pages 57-58), l'estimateur composite le plus efficace de μ prend la forme suivante :

$$\hat{\mu} = \frac{\bar{x}_2(\sigma_1^2 / n_1) + \bar{x}_1(\varepsilon^2 + \sigma_2^2 / n_2)}{\varepsilon^2 + \sigma_1^2 / n_1 + \sigma_2^2 / n_2},$$

avec le biais résiduel

$$\varepsilon_c = \varepsilon\left(\frac{\sigma_1^2 / n_1}{\varepsilon^2 + \sigma_1^2 / n_1 + \sigma_2^2 / n_2}\right)$$

et

$$\mathrm{EQM}_c = \frac{(\sigma_1^2 / n_1)(\varepsilon^2 + \sigma_2^2 / n_2)}{\varepsilon^2 + \sigma_1^2 / n_1 + \sigma_2^2 / n_2}.$$

Comme on peut le voir, l'estimateur composite est une combinaison convexe des moyennes de l'échantillon de convenance et de l'échantillon probabiliste. L'influence du premier est déterminée par le ratio de l'EQM (la variance) de la moyenne de l'échantillon probabiliste par rapport à la somme de ce terme et de l'EQM de la moyenne de l'échantillon de convenance. De même, le biais résiduel correspond au biais initial multiplié par ce même ratio, tandis que EQM_c est le produit des deux EQM divisé par leur somme. Soulignons que le biais se rapproche de zéro lorsque $\varepsilon \to 0$ (aucun biais de sélection dans l'estimation de l'échantillon de convenance) et lorsque $\varepsilon \to \infty$ (aucun poids n'a été attribué à l'échantillon de convenance).

3.2 La quantification des contributions de l'échantillon de convenance

Nous pouvons maintenant évaluer les contributions de l'échantillon de convenance en fonction du biais résiduel connu de ses estimateurs connexes. Pour ce faire, nous définirons plusieurs quantités.

Supposons que

$$\mathrm{TEE}_1 = \frac{\sigma_1^2 / n_1}{\mathrm{EQM}_c} n_1 = \left(\frac{\varepsilon^2 + \sigma_1^2 / n_1 + \sigma_2^2 / n_2}{\varepsilon^2 + \sigma_2^2 / n_2}\right) n_1$$

représente la taille effective de l'échantillon nécessaire pour obtenir une moyenne d'échantillon sans biais ayant la même EQM que l'estimateur composite. Pour simplifier cette expression, définissons le biais normalisé résiduel, $E = \varepsilon / \sigma$, et examinons le cas dans lequel les observations des populations de convenance et probabiliste ont une variance égale ($\sigma_1 = \sigma_2 = \sigma$). En pareil cas, la *croissance* de TEE_1 attribuable à l'échantillon de convenance, soit la différence entre TEE_1 avec et sans l'échantillon de convenance, est obtenue par l'équation suivante :

$$\frac{1}{1/n_2 + E^2} = n_2\left(\frac{1}{1 + n_2 E^2}\right).$$

3.3 Contribution maximale de l'échantillon de convenance

À mesure que $n_2 \to \infty$, la croissance de TEE_1 se rapproche de $1/E^2$. Cette limite, qui représente l'inverse du biais normalisé moyen, correspond à la contribution progressive maximale possible de l'échantillon de convenance à TEE_1 (CPMEC). Si la CPMEC est faible, alors un échantillon de convenance, peu importe sa taille, ne peut pas améliorer significativement l'EQM. Si la CPMEC est significative, nous devons déterminer la taille de l'échantillon de convenance nécessaire pour obtenir une grande partie de la CPMEC.

Pour développer une intuition de l'ampleur de E (biais normalisé), nous examinons le cas important d'un résultat dichotomique, pour lequel $E = \varepsilon / \sqrt{P(1-P)}$ où P correspond à la probabilité du résultat dans la population. Le tableau 1 traduit le biais d'un résultat dichotomique, des points de pourcentage au biais normalisé, puis à la CPMEC correspondante pour $P = 0{,}1$ et $P = 0{,}5$.

Lorsque la proportion se rapproche de 50 %, un biais de 2,5 points de pourcentage limite à 400 la croissance potentielle de TEE_1. La croissance minimale de TEE_1 pour compenser les frais fixes associés à l'établissement du mode de déclaration électronique varie d'un utilisateur à l'autre, mais nous croyons qu'une croissance inférieure à 100 serait rarement rentable. Le tableau 1 porte à croire que les échantillons de convenance pour lesquels des biais normalisés des estimateurs limités à l'échantillon de

convenance affichent généralement des valeurs 0,1 fois supérieures à celles des écarts-types seront rarement rentables. Pour une variable dichotomique où P se situe entre 0,1 et 0,5, cela représente un biais de 3 à 5 points de pourcentage.

Tableau 1 Contributions maximales des échantillons de convenance à l'estimation d'une proportion selon le biais en points de pourcentage

E (biais normalisé[#])	Prévalence globale du résultat 10 %	50 %	CPMEC[&]
	Biais en points de pourcentage		
0,01	0,3 %	0,5 %	10 000
0,02	0,6 %	1,0 %	2 500
0,05	1,5 %	2,5 %	400
0,10	3,0 %	5,0 %	100
0,20	6,0 %	10,0 %	25

[#] Des estimateurs des moyennes fondés uniquement sur l'échantillon de convenance

[&] La TEE amplifiée avec un échantillon de convenance infiniment grand par rapport à la non-utilisation d'un échantillon de convenance.

Peut-on obtenir facilement des biais de cette ampleur avec les estimations rajustées des échantillons de convenance? Plusieurs études ont comparé les échantillons de convenance électronique pondérés selon la propension aux enquêtes par composition aléatoire. L'une de ces études (Taylor 2000) préconisait l'utilisation autonome de ces échantillons de convenance en dépit des différences pouvant aller jusqu'à cinq points de pourcentage dans plusieurs estimations des résultats dichotomiques à l'égard des attitudes politiques, en supposant un biais normalisé de 0,05 à 0,1 si l'on se sert de la composition aléatoire comme norme d'excellence. Une autre étude (Schonlau, Zapert, Simon, Sanstad, Marcus, Adams, Spranca, Kan, Turner et Berry 2003) n'examine pas l'ampleur des différences, mais révèle que 29 des 37 éléments au sujet des préoccupations de santé comportent des différences statistiquement significatives à $p<0{,}01$. Compte tenu de la taille des échantillons (et en écartant avec optimisme tout effet de plan découlant de la pondération), on peut démontrer qu'un tel seuil de signification laisse supposer que les estimations en points de pourcentage du biais normalisé dépassent 0,05 pour les estimateurs de 78 % des éléments. Le principal résultat obtenu dans la comparaison slovène d'un échantillon probabiliste téléphonique et d'un échantillon de convenance électronique (Vehovar, Manfreda et Batagelj 1999) serait estimé avec un biais normalisé de plus de 0,10 à partir de l'échantillon de convenance, même après les corrections importantes en fonction de la pondération. Il convient de souligner que le mode de déclaration pourrait également influer sur les réponses déclarées par voie électronique comparativement au mode téléphonique chez des répondants randomisés en fonction du mode de déclaration (Fricker, Galesic, Tourangeau et Yan 2005). Ainsi, les différences entre l'échantillon de convenance électronique et l'échantillon probabiliste non électronique ne sont pas entièrement attribuables à la sélection.

3.4 Contribution réelle de l'échantillon de convenance

Bien que la croissance maximale possible (CPMEC) soit de $1/E^2$, la croissance réelle de TEE_1 peut s'exprimer comme suit : $(k/k+1)\,\mathrm{CPMEC}$ où $k=n_2E^2$. L'insuffisance de la croissance réelle de TEE_1 en raison de la CPMEC peut s'exprimer comme suit : $\mathrm{CPMEC}\text{-}\mathrm{TEE}_1 = 1/[(E^2)(1+n_2E^2)]$. Ainsi, les rendements de TEE_1 diminuent à mesure qu'augmente la taille de l'échantillon de convenance, la diminution se faisant plus rapide lorsque le biais est plus prononcé, étant donné que ce dernier finit par dominer toute autre réduction de la variance. La moitié de la CPMEC relevée est obtenue lorsque la TEE de l'échantillon de convenance est égale à la CPMEC. Par exemple, si le biais vaut 0,01 fois les écarts-types et qu'un échantillon de convenance a une TEE de 10 000 unités, alors la CPMEC est de 10 000, mais la contribution progressive réelle à TEE_1 équivaut à 5 000. Cette réalité suggère que les échantillons de convenance dont la TEE est de deux à 20 fois supérieure à la CPMEC seront suffisants dans la plupart des cas, ce qui représente de 67 % à 95 % de l'augmentation potentielle de la TEE. Ces connaissances heuristiques suggèrent à leur tour la collecte de 200 à 4 000 cas où E est relativement élevé ($E=0{,}05$ à 0,10), et de 5 000 à 200 000 cas où E est relativement faible ($E=0{,}01$ à 0,02). Le tableau 2 donne des exemples de la TEE_1 obtenue au moyen de diverses combinaisons de tailles d'échantillons et de biais.

Tableau 2 Exemples de TEE_1 selon diverses tailles d'échantillons et niveaux de biais normalisé

n_1 (taille de l'éch. probab.)	n_2 (taille de l'éch. de convenance)	E (biais normalisé [#])	TEE_1 pour l'estimation composite	TEE_1 / n_1[&]
1 000	1 000	0,01	1 909	1,909
1 000	1 000	0,10	1 091	1,091
1 000	10 000	0,01	6 000	6,000
1 000	10 000	0,10	1 099	1,099
1 000	100 000	0,01	10 091	10,091
1 000	100 000	0,10	1 100	1,100
10 000	1 000	0,01	10 909	1,091
10 000	1 000	0,10	10 091	1,009
10 000	10 000	0,01	15 000	1,500
10 000	10 000	0,10	10 099	1,010
10 000	100 000	0,01	19 091	1,909
10 000	100 000	0,10	10 100	1,010

Plusieurs estimateurs de moyennes fondés uniquement sur l'échantillon de convenance

[#] Des estimateurs de moyennes fondés uniquement sur l'échantillon de convenance

[&] TEE par rapport à la non-utilisation de l'échantillon de convenance.

3.5 La précision de l'estimation du biais

Dans le présent document, nous avons présumé que le biais des estimateurs de l'échantillon de convenance était connu; en pratique, le biais devra être estimé au moyen des données des deux échantillons. Nous examinons ensuite dans quelle mesure la taille de l'échantillon probabiliste limite également l'utilité de l'échantillon de convenance en raison de la nécessité d'estimer avec précision le biais résiduel.

Nous pouvons estimer ε comme la différence entre la moyenne de l'échantillon probabiliste et la moyenne pondérée de l'échantillon de convenance. L'erreur type de l'estimation du biais est obtenue par $\sigma_{\hat{\varepsilon}} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. Si $\sigma_1 = \sigma_2 = \sigma$, l'erreur type réelle de l'estimation du biais normalisé (E) est obtenue par l'équation suivante : $\sigma_{\hat{E}} = \sqrt{1/n_1 + 1/n_2}$. Peu importe la taille de l'échantillon de convenance, ce terme ne peut jamais être inférieur à l'inverse de la racine carrée de la taille de l'échantillon probabiliste.

Il a été démontré que l'erreur relative de l'EQM pour un estimateur composite est relativement insensible aux petites erreurs des estimations du biais (Schaible 1978), ce qui est encourageant pour les biais bien estimés. Malheureusement, à moins que la TEE des échantillons probabiliste et de convenance soit élevée, l'erreur type de l'estimation de E est inutilement élevée par rapport aux valeurs de E qui rendent utile le supplément de convenance ($E < 0{,}10$). Par exemple, supposons qu'un échantillon probabiliste dont la TEE est de 1 000 unités et qu'un échantillon de convenance dont la TEE est de 5 000 aient estimé un biais normalisé de 0,02 points de pourcentage. Si l'estimation ponctuelle était correcte, l'échantillon de convenance augmenterait TEE_1 de 1 667 unités. Cependant, cette estimation pourrait également comporter un biais réel valant 0,088 fois les écarts-types (95 % de limite de confiance supérieure), ce qui signifierait que l'augmentation est inférieure à 130.

Si nous présumons que la taille de l'échantillon de convenance sera toujours équivalente au moins au double de la taille de l'échantillon probabiliste, ces résultats signifient que pour que cette technique puisse être mise en pratique, l'échantillon probabiliste doit comporter au moins 1000 à 10 000 unités, de manière à lever l'incertitude de l'ampleur du biais des estimateurs de l'échantillon de convenance (erreurs-types de E dans la fourchette de 0,01 à 0,04).

4. Discussion

Nous décrivons un estimateur composite, qui est une combinaison linéaire de l'estimation de la moyenne d'un échantillon probabiliste sans biais et de l'estimation de la moyenne d'un échantillon de convenance électronique biaisé (pondération fondée sur les scores de propension). Nous nous servons de l'EQM de cet estimateur composite pour caractériser en fonction de la TEE les contributions de l'échantillon de convenance à un estimateur uniquement fondé sur l'échantillon probabiliste. Nous calculons ensuite la contribution maximale de l'échantillon de convenance, le rôle de la taille de l'échantillon de convenance pour approcher cette limite, et les rôles des deux tailles d'échantillons pour estimer le biais de manière suffisamment précise.

Certains spécialistes présument parfois que les petits échantillons probabilistes sont suffisants pour estimer le biais des estimations des échantillons de convenance correspondants. Les résultats que nous avons obtenus suggèrent tout autre chose. Nous démontrons que le biais normalisé des estimateurs de l'échantillon de convenance électronique, après les rajustements initiaux pour réduire le biais, doit être très petit (0,1 fois les écarts-types tout au plus, et probablement de moins de 0,05) pour que l'EQM de l'estimation globale soit significativement plus faible qu'elle ne l'aurait été sans l'échantillon de convenance. Nous démontrons en outre que des échantillons de convenance comportant des milliers ou des dizaines de milliers d'unités sont également nécessaires par souci d'efficacité. Enfin, nous démontrons la nécessité d'un gros échantillon probabiliste (de 1 000 à 10 000 unités) pour obtenir des estimations raisonnablement précises du biais résiduel des estimateurs de l'échantillon de convenance initialement corrigés en fonction du biais. Comme le biais des estimations dans une application d'enquête polyvalente a tendance à varier selon les résultats, il faut faire preuve de prudence avant de prendre la décision globale de remplacer un grand nombre d'enquêtes peu coûteuses par un plus petit nombre d'enquêtes traditionnelles.

Ce sont les grandes enquêtes qui offrent probablement la possibilité de réaliser les plus grandes économies, tout simplement en raison de leur taille. Par ailleurs, les améliorations proportionnelles de précision surviendront probablement dans les échantillons de taille intermédiaire. Les améliorations pourraient également s'avérer considérables dans les grands échantillons dont les principales inférences sont des sous-groupes plus petits. Par exemple, une enquête nationale menée auprès de 100 000 personnes pourraient faire une inférence à 200 sous-régions géographiques, comportant chacune un échantillon de 500 unités. Si l'on ajoutait à cet échantillon national un très gros échantillon de convenance électronique, estimait le biais à l'échelle nationale et admettait l'invariabilité régionale du biais, on pourrait réduire considérablement l'EQM des estimations infrarégionales en utilisant un tel estimateur composite.

Mentionnons maintenant une dernière mise en garde. Les conclusions au sujet de l'utilité limitée des échantillons de convenance lorsque le biais des estimateurs est supérieur à 0,1 fois les écarts-types ne se limitent pas aux tentatives visant à utiliser un estimateur composite. La même approche peut être appliquée pour démontrer que, dans le cas d'un estimateur fondé uniquement sur un échantillon de convenance de n'importe quelle taille comportant un biais normalisé de 0,2 (par exemple, 10 points de pourcentage pour une variable dichotomique où $P = 0,5$), l'EQM sera supérieure ou égale à celle d'une estimation d'un échantillon probabiliste de 25 unités.

Remerciements

Les auteurs tiennent à remercier John Adams, Ph.D., des commentaires qu'il nous a fournis au sujet du document manuscrit, Matthias Schonlau, Ph.D., de ses commentaires au sujet d'une version précédente du document manuscrit, et Colleen Carey, B.A., Kate Sommers-Dawes, B.A., et Michelle Platt, B.A. de leur collaboration à la rédaction du document.

Marc Elliott bénéficie entre autres de l'aide financière des Centers for Disease Control and Prevention (CDC U48/DP000056); les opinions exprimées dans le présent article ne reflètent pas nécessairement celles des Centers for Disease Control and Prevention.

Bibliographie

Couper, M.P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64, 464-494.

Fricker, S., Galesic, M., Tourangeau R. et Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, 69, 370-392.

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

Kish, L. (1985). *Survey Sampling*. New York : John Wiley & Sons, Inc.

Kypri, K., Stephenson, S. et Langley, J. (2004). Assessment of nonresponse bias in an internet survey of alcohol use. *Alcoholism*: *Clinical and Experimental Research*, 28, 630-634.

Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.

Rosenbaum, P.R. (2002). *Observational Studies*, 2ième édition, New York : Springer-Verlag.

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

Schaible, W.A. (1978). Choosing weights for composite estimators for small area statistics. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 741-746.

Schonlau, M., Zapert, K., Simon, L.P., Sanstad, K., Marcus, S., Adams, J., Spranca, M., Kan, H., Turner, R. et Berry, S. (2003). A comparison between responses from propensity-weighted web survey and an identical RDD survey. *Science Computer Review*, 21, 1-11.

Taylor, H. (2000). Does internet research 'work'? Comparing on-line survey results with telephone surveys. *International Journal of Market Research*, 42, 41-63.

Vehovar, V., Manfreda, K.L. et Batagelj, Z. (1999). Web surveys: Can weighting solve the problem? *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 962-967.

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 2007.

R.C. Bailey, *George Mason University*
M. Bankier, *Statistique Canada*
J.-F. Beaumont, *Statistique Canada*
Y. Bélanger, *Statistique Canada*
W. Bell, *U.S. Bureau of the Census*
Y. Berger, *University of Southampton*
C. Boudreau, *Medical College of Wisconsin*
M. Brick, *Westat*, *Inc.*
J. Chen, *University of Waterloo*, *Canada*
R. Clark, *University of Wollongong*, *Australie*
E. Dagum, *University of Bologna*
A. Dessertaine, *EDF*, *R&D-OSIRIS-CLAMART*
F. Dupont, *INSEE*
M. Elliott, *University of Michigan*
J. Eltinge, *U.S. Bureau of Labor Statistics*
L.R. Ernst, *Bureau of Labour Statistics*, *É.-U.*
L. Fattorini, *Università di Siena*
R. Fisher, *U.S. Department of Treasury*
W. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
D. Garriguet, *Statistique Canada*
C. Girard, *Statistique Canada*
E. Gois, *Instituto Nacional de Estatistica*, *Portugal*
B. Graubard, *National Cancer Institute*
R. Griffin, *US Bureau of the Census*
D. Haziza, *Statistique Canada*
S. Heeringa, *ISR*, *Michigan*
M. Hidiroglou, *Statistique Canada*
J. Jiang, *University of California at Davis*
W. Kalsbeek, *U. of North Carolina*
J.-K. Kim, *Department of Applied Statistics*, *Seoul*, *Corée*
P. Knottnerus, *Statistics Netherlands*
E. Korn, *National Institute of Health*
P. Kott, *National Agricultural Statistics Service*
D. Ladiray, *INSEE*
P. Lahiri, *JPSM*, *University of Maryland*
N. Laniel, *Statistique Canada*
M. Latouche, *Statistique Canada*
P. Lavallée, *Statistique Canada*
S. Lee, *University of California at Los Angeles*
S. Lele, *University of Alberta*
C. Léon, *Statistique Canada*
P.S. Lévy, *RTI International*
S. Lohr, *Arizona State University*
W.W. Lu, *Acadia University*
L. Mach, *Statistique Canada*
T. Maiti, *Iowa State University*
D. Malec, *United States Bureau of the Census*
H. Mantel, *Statistique Canada*
A. Matei, *Université de Neuchâtel*, *Suisse*
S. Merad, *Office for National Statistics*, *R.-U.*
J. Montequila, *Westat*, *Inc.*
R. Munnich, *University of Tubingen*
P.L. Nascimento Silva, *University of Southampton*
G. Nathan, *Hebrew University of Jerusalem*
J. Opsomer, *Iowa State University*
S.M. Paddock, *Rand corporation*, *É.-U.*
S. Pramanik, *JPSM*, *University of Maryland*
N. Prasad, *University of Alberta*
L. Qualité, *Université de Neuchâtel*, *Suisse*
T.J. Rao, *ISI*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
S. Rubin-Bleuer, *Statistique Canada*
M. del Mar Rueda, *University of Granada*
P. Saavedra, *ORC Macro*, *É.-U.*
F. Scheuren, *NORC*
E. Schindler, *US Bureau of the Census*
K. Sinha, *University of Illinois at Chicago*, *É.-U.*
C.J. Skinner, *University of Southampton*
P.J. Smith, *Center for Disease Control and Prevention*
M. Thompson, *University of Waterloo*
Y. Tillé, *Université de Neuchâtel*
F. Verret, *Statistique Canada*
S.G. Walker, *University of Kent*, *R.-U.*
W. Winkler, *US Bureau of the Census*
K. Wolter, *NORC*
C. Wu, *University of Waterloo*
W. Yung, *Statistique Canada*

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2007: Cécile Bourque, Louise Demers, Anne-Marie Fleury, Roberto Guido, Liliane Lanoie, Micheal Pelchat et Isabelle Poliquin (Division de la diffusion), Sheri Buck (Division du développement de systèmes), François Beaudin (Division des langues officielles et traduction) et Sophie Chartier et Gayle Keely (Division des méthodes d'enquêtes auprès des ménages). Finalement nous désirons exprimer notre reconnaissance à Christine Cousineau, Céline Ethier et Denis Lemire de la Division des méthodes d'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 23, No. 1, 2007

Challenges to the Confidentiality of U.S. Federal Statistics, 1910-1965
Margo Anderson and William Seltzer 1

Efficient Stratification Based on Nonparametric Regression Methods
Enrico Fabrizi and Carlo Trivisano 35

A Selection Strategy for Weighting Variables Under a Not-Missing-at-Random Assumption
Barry Schouten 51

Imputing for Late Reporting in the U.S. Current Employment Statistics Survey
Kennon R. Copeland and Richard Valliant 69

Incentives in Random Digit Dial Telephone Surveys: A Replication and Extension
Richard Curtin, Eleanor Singer and Stanley Presser 91

Methods for Achieving Equivalence of Samples in Cross-National Surveys: The European Social Survey Experience
Peter Lynn, Sabine Häder, Siegfried Gabler and Seppo Laaksonen 107

Book and Software Reviews 125

Contents
Volume 23, No. 2, 2007

Estimation of Nonresponse Bias in the European Social Survey: Using Information from Reluctant Respondents
Jaak Billiet, Michel Phillippens, Rory Fitzgerald and Ineke Stoop 135

Measuring Disability in Surveys: Consistency over Time and Across Respondents
Sunghee Lee, Nancy A. Mathiowetz and Roger Tourangeau 163

Quantifying Stability and Change in Ethnic Group
Ludi Simpson and Bola Akinwale 185

Seasonal Adjustment of Weekly Time Series with Application to Unemployment Insurance Claims and Steel Production
William P. Cleveland and Stuart Scott 209

Finite Population Small Area Interval Estimation
Li-Chun Zhang 223

Predicting Natural Gas Production in Texas: An Application of Nonparametric Reporting Lag Distribution Estimation
Crystal D. Linkletter and Randy R. Sitter 239

Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update
Laura Zayatz 253

Book and Software Reviews 267

Contents
Volume 23, No. 3, 2007

The Morris Hansen Lecture 2006 Statistical Perspectives on Spatial Social Science
Michael F. Goodchild 269

Using Geospatial Information Resources in Sample Surveys
Sarah M. Nusser 285

Discussion
Linda Williams Pickle 291

Optimizing the Use of Microdata: An Overview of the Issues
Julia Lane 299

Benchmarking the Effect of Cell Adjustment on Tabular Outputs: The Shortcomings of Current Approaches
Paul Williamson 319

Summary of Accuracy and Coverage Evaluation for the U.S. Census 2000
Mary H. Mulry 345

Resampling Variance Estimation in Surveys with Missing Data
A.C. Davison and S. Sardy 371

Nonresponse Among Ethnic Minorities: A Multivariate Analysis
Remco Feskens, Joop Hox, Gerty Lensvelt-Mulders and Hans Schmeets 387

Procedures for Updating Classification Systems: A Study of Biotechnology and the Standard Occupational Classification System
Neil Malhotra and Jon A. Krosnick 409

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 35, No. 1, March/mars 2007

Paul GUSTAFSON
Editorial 1

Report from the previous Editor 5

Acknowledgement of referees' services 7

Renato ASSUNCÇÃO, Andréa TAVARES, Thais CORREA & Martin KULLDORFF
Space-time cluster identification in point processes 9

Elvan CEYHAN, Carey E. PRIEBE & David J. MARCHETTE
A new family of random graphs for testing spatial segregation 27

Aurélie LABBE & Mary E. THOMPSON
Multiple testing using the posterior probabilities of directional alternatives, with application to genomic studies 51

H.a.o.l.a.n. LU, James S. HODGES & Bradley P. CARLIN
Measuring the complexity of generalized linear hierarchical models 69

Aurore DELAIGLE
Nonparametric density estimation from data with a mixture of Berkson and classical errors 89

Ozlem ILK & Michael J. DANIELS
Marginalized transition random effect models for multivariate longitudinal binary data 105

Richard A. LOCKHART, John J. SPINELLI & Michael A. STEPHENS
Cramér-von Mises statistics for discrete distributions with unknown parameters 125

P.W. FONG, W.K. LI, C.W. YAU & C.S. WONG
On a mixture vector autoregressive model 135

Jung Wook PARK, Marc G. GENTON & Sujit K. GHOSH
Censored time series analysis with autoregressive moving average models 151

Abdessamad SAIDI
Consistent testing for non-correlation of two cointegrated ARMA time series 169

Forthcoming papers/Articles à paraître 189

Complete online access to *The Canadian Journal of Statistics* 191

Volume 35, No. 2, June/juin 2007

Pierre DUCHESNE
On consistent testing for serial correlation in seasonal time series models 193

Jaakko NEVALAINEN, Denis LAROCQUE & Hannu OJA
On the multivariate spatial median for clustered data 215

Liqun WANG
A unified approach to estimation of nonlinear mixed effects and Berkson measurement error models 233

Juan Carlos PARDO-FERNÁNDEZ, Ingrid VAN KEILEGOM & Wenceslao GONZÁLEZ-MANTEIGA
Goodness-of-fit tests for parametric models in censored regression 249

Song Xi CHEN & Tzee-Ming HUANG
Nonparametric estimation of copula functions for dependence modelling 265

Victor DE OLIVEIRA
Objective Bayesian analysis of spatial data with measurement error 283

Gonzalo GARCÍA-DONATO & Dongchu SUN
Objective priors for hypothesis testing in one-way random effects models 303

Hui LI, Guosheng YIN & Yong ZHOU
Local likelihood with time-varying additive hazards model 321

Forthcoming papers/Articles à paraître 338

Complete online access to *The Canadian Journal of Statistics* 339

Volume 35, No. 3, September/septembre 2007

David F. ANDREWS
Robust likelihood inference for public policy 341

Zeny Z. FENG, Jiahua CHEN, Mary E. THOMPSON
Asymptotic properties of likelihood ratio test statistics in affected-sib-pair analysis 351

Hwashin H. SHIN, Glen TAKAHARA & Duncan J. MURDOCH
Optimal designs for calibration of orientations 365

Jiancheng JIANG, Haibo ZHOU, Xuejun JIANG & Jianan PENG
Generalized likelihood ratio tests for the structure of semiparametric additive models 381

Runze LI & Lei NIE
A new estimation procedure for a partially nonlinear model via a mixed-effects approach 399

Axel MUNK, Matthias MIELKE, Guido SKIPKA & Gudrun FREITAG
Testing noninferiority in three-armed clinical trials based on likelihood ratio statistics 413

Inkyung JUNG & Martin KULLDORFF
Theoretical properties of tests for spatial clustering of count data 433

QiQi LU & Robert B. LUND
Simple linear regression with multiple level shifts 447

Forthcoming papers/Articles à paraître 459

Volume 36 (2008): Subscription rates/Frais d'abonnement 460