Nº 12-001-X au catalogue

Techniques d'enquête

Juin 2011

Statistique Canada
Statistics Canada
Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca. Vous pouvez également communiquer avec nous par courriel à infostats@statcan.gc.ca ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :

Service de renseignements	1-800-263-1136
Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
Télécopieur	1-877-287-4369

Appels locaux ou internationaux :

Service de renseignements	1-613-951-8116
Télécopieur	1-613-951-0581

Programme des services de dépôt

Service de renseignements	1-800-635-7943
Télécopieur	1-800-565-7757

Comment accéder à ce produit ou le commander

Le produit nº 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Ce produit 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 les moyens suivants :

- Téléphone (Canada et États-Unis) 1-800-267-6677
- Télécopieur (Canada et États-Unis) 1-877-287-4369
- Courriel infostats@statcan.gc.ca
- Poste Statistique Canada
Finances
Immeuble R.-H.-Coats, 6e étage
150, 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 que les employés observent. 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 sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Statistique Canada
Division des méthodes d'enquêtes auprès des entreprises

Techniques d'enquête

Juin 2011

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2011

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.

Juin 2011

N° 12-001-XIF au catalogue
ISSN 1712-5685

N° 12-001-XPB au catalogue
ISSN 0714-0045

Périodicité : semestriel

Ottawa

This publication is available in English upon request (catalogue no. 12-001-X)

Note de reconnaissance

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

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods. La revue est également citée par SCOPUS sur les bases de données Elsevier Bibliographic Databases.

COMITÉ DE DIRECTION

Président J. Kovar

Anciens présidents D. Royce (2006-2009)
G.J. Brackstone (1986-2005)
R. Platek (1975-1986)

Membres G. Beaudoin
S. Fortier (Gestionnaire de la production)
J. Gambino
M.A. Hidiroglou
H. Mantel

COMITÉ DE RÉDACTION

Rédacteur en chef M.A. Hidiroglou, *Statistique Canada*
Rédacteur en chef délégué H. Mantel, *Statistique Canada*

Ancien rédacteur en chef J. Kovar (2006-2009)
M.P. Singh (1975-2005)

Rédacteurs associés

J.-F. Beaumont, *Statistique Canada*
J. van den Brakel, *Statistics Netherlands*
J.M. Brick, *Westat Inc.*
P. Cantwell, *U.S. Bureau of the Census*
R. Chambers, *Centre for Statistical and Survey Methodology*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
B. Hulliger, *University of Applied Sciences Northwestern Switzerland*
D. Judkins, *Westat Inc*
D. Kasprzyk, *NORC at the University of Chicago*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistique Canada*
P. Lynn, *Univerity of Essex*
D.J. Malec, *National Center for Health Statistics*
G. Nathan, *Hebrew University*
J. Opsomer, *Colorado State University*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
J.N.K. Rao, *Carleton University*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
P. do N. Silva, *Escola Nacional de Ciências Estatísticas*
P. Smith, *Office for National Statistics*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
M. Thompson, *University of Waterloo*
V.J. Verma, *Università degli Studi di Siena*
K.M. Wolter, *Iowa State University*
C. Wu, *University of Waterloo*
W. Yung, *Statistique Canada*
A. Zaslavsky, *Harvard University*

Rédacteurs adjoints C. Bocci, P. Dick, G. Dubreuil, S. Godbout, D. Haziza, Z. Patak, S. Rubin-Bleuer et Y. You, *Statistique Canada*

POLITIQUE DE RÉDACTION

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

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préférablement en Word au rédacteur en chef, (rte@statcan.gc.ca, Statistique Canada, 150 Promenade du Pré Tunney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue ou sur le site web (www.statcan.gc.ca).

Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 $ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada : États-Unis 12 $ CA (6 $ × 2 exemplaires); autres pays, 20 $ CA (10 $ × 2 exemplaires). Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et statisticiens du Québec. Des versions électroniques sont disponibles sur le site internet de Statistique Canada : www.statcan.gc.ca.

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 37, numéro 1, juin 2011

Table des matières

Articles réguliers

J. Michael Brick, Ismael Flores Cervantes, Sunghee Lee et Greg Norman
Erreurs non dues à l'échantillonnage dans les enquêtes téléphoniques à base de sondage double 1

James O. Chipperfield, Glenys R. Bishop et Paul Campbell
Estimation du maximum de vraisemblance pour les tableaux de contingence et la régression logistique en présence de données incorrectement appariées 17

Yong You et Qian M. Zhou
Estimation sur petits domaines hiérarchique bayésienne sous un modèle spatial avec application à des données d'enquête sur la santé 31

Hukum Chandra et Ray Chambers
Estimation sur petits domaines sous linéarisation 45

Sophie Baillargeon et Louis-Paul Rivest
Élaboration de plans stratifiés en R à l'aide du programme *stratification* 59

Jae Kwang Kim et Cindy Long Yu
Estimation de la variance par répliques sous échantillonnage à deux phases 73

Stanislav Kolenikov et Gustavo Angeles
Rentabilité des enquêtes avec échantillonnage en grappes répétées 83

Paul Knottnerus
À propos de l'efficacité de l'échantillonnage à probabilité proportionnelle à la taille aléatoire 103

Communications brèves

Éric Lesage
Utilisation des équations estimantes pour réaliser un calage sur des paramètres complexes 111

Autres revues 117

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, juin 2011
Vol. 37, N° 1, pp. 1-16
Statistique Canada, N° 12-001-X au catalogue

Erreurs non dues à l'échantillonnage dans les enquêtes téléphoniques à base de sondage double

J. Michael Brick, Ismael Flores Cervantes, Sunghee Lee et Greg Norman [1]

Résumé

Les enquêtes téléphoniques à base de sondage double deviennent fréquentes aux États-Unis en raison de l'incomplétude de la liste de numéros de téléphone fixe causée par l'adoption progressive du téléphone mobile. Le présent article traite des erreurs non dues à l'échantillonnage dans les enquêtes téléphoniques à base de sondage double. Alors que la plupart des publications sur les bases de sondage doubles ne tiennent pas compte des erreurs non dues à l'échantillonnage, nous constatons que ces dernières peuvent, dans certaines conditions, causer des biais importants dans les enquêtes téléphoniques à base de sondage double. Nous examinons en particulier les biais dus à la non-réponse et à l'erreur de mesure dans ces enquêtes. En vue de réduire le biais résultant de ces erreurs, nous proposons des méthodes d'échantillonnage à base de sondage double et de pondération. Nous montrons que le facteur de composition utilisé pour combiner les estimations provenant de deux bases de sondage joue un rôle important dans la réduction du biais de non-réponse.

Mots clés : Biais de non-réponse ; erreur de mesure ; calage ; répartition de l'échantillon ; composite.

1. Introduction

Les enquêtes téléphoniques à base de sondage double comportant l'échantillonnage de numéros de téléphone fixe et de numéros de téléphone mobile ont pris de l'importance aux États-Unis en vue de réduire le biais de sous-dénombrement dû à l'incomplétude de la liste de numéros de téléphone fixe. Blumberg et Luke (2009) montrent que le pourcentage de ménages ne possédant pas de ligne téléphonique fixe mais dotés d'au moins un téléphone mobile a augmenté de façon spectaculaire ces dernières années, pour atteindre 20 % à la fin de 2008. D'autres pays font également part d'un accroissement considérable du pourcentage de personnes abonnées seulement à la téléphonie mobile (par exemple Kuusela, Callegaro et Vehovar 2008 ; Vicente et Reis 2009).

Le présent article s'appuie sur des données provenant de la California Health Interview Survey (CHIS) et de huit enquêtes réalisées pour le compte du Pew Research Center for the People & the Press (enquêtes Pew) en vue d'étudier les effets des erreurs non dues à l'échantillonnage dans les enquêtes téléphoniques à base de sondage double. La CHIS 2007, menée auprès des Californiens adultes, a été réalisée à la fin de 2007. Cette enquête combine un sondage par téléphone fixe classique et un échantillon de présélection de numéros de téléphone mobile, dans lequel seuls les adultes ayant indiqué que le ménage dont ils faisaient partie ne possédait pas de numéro de téléphone fixe ont été interviewés. Les enquêtes Pew sont des sondages nationaux comportant l'interview d'un adulte pour chacun des numéros de téléphone résidentiels provenant de l'échantillon de téléphones fixes et de celui de téléphones mobiles. Ces enquêtes sont décrites plus en détail plus loin. La réalisation de ces enquêtes téléphoniques à base de sondage double a mis en relief un certain nombre de questions importantes associées à l'effet des erreurs non dues à l'échantillonnage – erreurs qui n'ont pas été examinées complètement dans d'autres études.

À la section suivante, nous passons en revue le plan de sondage, la pondération et les méthodes d'estimation de la variance élaborées pour les enquêtes à base de sondage double, et décrivons CHIS 2007 et les enquêtes téléphoniques Pew à base de sondage double qui sont utilisées tout au long de l'article. À la troisième section, nous discutons de l'erreur non due à l'échantillonnage dans les enquêtes téléphoniques à base de sondage double, et des effets que ces erreurs peuvent avoir sur le biais des estimations. La non-réponse et les erreurs de mesure revêtent une importance particulière dans les enquêtes à base de sondage double. À la quatrième section, nous étudions les méthodes d'échantillonnage et d'estimation qui peuvent être utilisées pour réduire le biais dans les enquêtes téléphoniques à base de sondage double et nous donnons les conditions sous lesquelles ces approches d'échantillonnage et d'estimation peuvent être les plus utiles. Dans cette section, nous proposons trois estimateurs en vue de réduire le biais dû à la non-réponse différentielle dans le domaine de chevauchement des bases de sondage. À la dernière section, nous résumons certaines constatations concernant les enquêtes téléphoniques à base de sondage double et avançons des hypothèses quant à l'applicabilité de ces constatations à d'autres enquêtes à base de sondage double.

1. J. Michael Brick, Westat et Joint Program in Survey Methodology à l'Université du Maryland. Courriel : mikebrick@westat.com ; Ismael Flores Cervantes, Westat ; Greg Norman, Westat, 1600 Research Boulevard, Rockville, Maryland, 20850, États-Unis ; Sunghee Lee, Institute for Social Research, University of Michigan, 426 Thompson St. Ann Arbor, MI 48104, États-Unis.

2. Contexte

La plupart des publications traitant des enquêtes à base de sondage double portent sur la théorie statistique relative à l'efficacité du plan de sondage et de l'estimation. Nous résumons certains résultats importants concernant l'échantillonnage, la pondération et l'estimation de la variance, puis nous discutons de l'application de ces méthodes aux enquêtes téléphoniques à base de sondage double.

2.1 Échantillonnage

Soit les deux bases de sondage désignées par *A* et *B*. Nous supposons que les échantillons tirés de ces bases, S_A et S_B, sont indépendants. Le domaine des unités qui ne figurent que dans *A* est désigné par *a*, le domaine des unités comprises uniquement dans *B* est désigné par *b*, et l'intersection contenant les unités chevauchantes est désignée par *ab*. Dans notre application aux enquêtes téléphoniques, *A* est la base de numéros de téléphone fixe, *B* est la base des numéros de téléphone mobile, *a* est le domaine des ménages n'ayant que des numéros de téléphone fixe, *b* est le domaine des ménages n'ayant que des numéros de téléphone mobile, et *ab* est le domaine des ménages ayant les deux types de service téléphonique. De nombreuses caractéristiques importantes des enquêtes à base de sondage double dépendent du traitement réservé aux unités qui pourraient se trouver dans les deux bases de sondage (domaine *ab*).

Une approche d'enquête à base de sondage double avec présélection vise à rendre $ab = \varnothing$ en éliminant toutes les unités chevauchantes avant l'échantillonnage, après l'échantillonnage mais avant la collecte des données, durant la collecte des données, ou après la collecte des données. Lohr (2009) donne des exemples d'enquêtes à base de sondage double réalisées selon chacune de ces approches.

Brick, Edwards et Lee (2007) et Fleeman (2007) décrivent la présélection dans les enquêtes téléphoniques à base de sondage double. Bien qu'aux États-Unis, les numéros de téléphone puissent être partitionnés selon qu'il s'agit de numéros de téléphone mobile ou de téléphone fixe, cette base de sondage ne permet pas de déterminer si ces numéros correspondent à des ménages n'ayant que des téléphones fixes (*a*), de ménages n'ayant que des téléphones mobiles (*b*), ou de ménages ayant les deux types de service (*ab*). Dans les enquêtes décrites par Brick, Edwards et Lee (2007) et Fleeman (2007), les ménages tirés de la liste de numéros de téléphone mobile (*B*) ont été écartés durant la collecte des données s'ils déclaraient posséder une ligne fixe. Cette approche de sélection est celle qui a été adoptée pour la CHIS 2007.

Selon une deuxième approche, appelée enquête à base de sondage double avec chevauchement, les unités faisant partie du domaine de chevauchement peuvent être échantillonnées à partir des deux bases de sondage. Dans ce cas, des méthodes d'estimation doivent être appliquées pour éviter de produire des estimations biaisées parce que les unités chevauchantes ont de multiples chances de sélection. Steeh (2004), Brick, Brick, Dipko, Presser, Tucker et Yuan (2007), et Kennedy (2007) discutent des enquêtes téléphoniques à base de sondage double avec chevauchement. Dans ces cas, tous les répondants sont interviewés, quelle que soit la base de sondage dans laquelle ils ont été sélectionnés. L'approche avec chevauchement est utilisée dans les enquêtes Pew.

2.2 Estimation

Dans une enquête avec présélection, la production de pondérations pour l'estimation des totaux et des caractéristiques de l'ensemble de la population est simple, du moins en l'absence d'erreur non due à l'échantillonnage. Puisque $ab = \varnothing$ et que l'échantillonnage est indépendant, les unités échantillonnées dans chaque base de sondage se voient attribuer un poids égal à l'inverse de leur probabilité de sélection dans la base de sondage dont elles proviennent. Une estimation globale du total s'obtient en faisant la somme des estimations de domaine pondérées, $\hat{y}_{scr} = \hat{y}_A + \hat{y}_b$, où $\hat{y}_A = \sum_{i \in S_A} d_i y_i$ et $\hat{y}_b = \sum_{i \in S_B} d_i \delta_i(b) y_i$, où d_i est l'inverse de la probabilité de sélection et $\delta_i(b) = 1$ si i est dans le domaine *b* et 0 autrement. L'estimation de la variance est également simple, puisque les deux bases de sondage sont des strates et que l'on peut appliquer les méthodes d'estimation de la variance appropriées pour les échantillons stratifiés. Dans le cas des enquêtes téléphoniques, les unités provenant de l'échantillon de numéros de téléphone fixe sont pondérées et ajoutées aux unités pondérées provenant de l'échantillon de numéros de téléphone mobile, après avoir attribué un poids nul aux unités tirées de la base de numéros de téléphone mobile qui possèdent un téléphone fixe.

Même en l'absence d'erreurs non dues à l'échantillonnage, la présélection durant la collecte des données a des conséquences. Par exemple, les ménages tirés de *B* qui sont écartés ne sont pas admissibles à l'interview, ce qui augmente le coût de la collecte des données et la variance des totaux estimés (Kish 1965, chapitre 11). Les unités qui sont écartées devraient aussi être traitées correctement comme des unités échantillonnées dans l'estimation de la variance.

Les enquêtes avec chevauchement sont plus complexes, parce que les unités peuvent être échantillonnées dans l'une ou l'autre base de sondage. Une approche d'estimation consiste à combiner les deux estimations de domaine, $\hat{y}_a$ et $\hat{y}_b$, avec une moyenne des estimations pour la population chevauchante produites d'après les bases de sondage distinctes. Si $\hat{y}_{ab}^A$ et $\hat{y}_{ab}^B$ sont les estimations pondérées du domaine de chevauchement provenant des bases de sondage *A* et *B*, respectivement, un estimateur moyen ou composite

est donné par $\hat{y}_{ave} = \hat{y}_a + \hat{y}_b + \lambda\hat{y}_{ab}^A + (1-\lambda)\hat{y}_{ab}^B$, avec $0 \le \lambda \le 1$. À l'instar de Lohr (2009), nous donnons à ces estimateurs le nom d'estimateurs moyens. Si nous supposons que $\hat{y}_a$ et $\hat{y}_b$ sont sans biais pour le domaine *a* et pour le domaine *b*, respectivement et que $\hat{y}_{ab}^A$ et $\hat{y}_{ab}^B$ sont tous deux sans biais pour le domaine *ab*, alors $\hat{y}_{ave}$ est un estimateur sans biais du total. Les estimations des moyennes et d'autres quantités peuvent être produites en utilisant des pondérations, celles pour les unités comprises dans *ab* tirées de *A* étant multipliées par λ et celles pour les unités tirées de *B* étant multipliées par $(1-\lambda)$. Le choix du facteur de composition, λ, a été étudié par de nombreux chercheurs et des valeurs particulières en vue de réduire la variance des estimations ont été proposés par Hartley (1962, 1974) et par Fuller et Burmeister (1972). Tous les estimateurs moyens requièrent que le domaine puisse être précisé pour chacune des unités échantillonnées.

L'estimation de la variance en se servant de l'estimateur moyen est relativement simple si λ est fixe et indépendant de l'échantillon sélectionné. Dans ce cas, $V(\hat{y}_{ave}) = V(\hat{y}_a + \lambda\hat{y}_{ab}^A) + V(\hat{y}_b + (1-\lambda)\hat{y}_{ab}^B)$, et chacune de ces variances peut être calculée en utilisant des méthodes d'estimation de la variance appropriées pour les échantillons distincts. Si λ dépend de l'échantillon, comme dans le cas des estimateurs de Hartley et de Fuller et Burmeister, l'estimation de la variance est plus compliquée. Les estimateurs moyens avec un facteur λ fixe ont été utilisés dans la plupart des enquêtes téléphoniques à base de sondage double avec chevauchement. Nous discutons de cette approche plus loin pour les enquêtes Pew.

D'autres méthodes d'estimation qui ont été prises en considération pour une enquête avec chevauchement comprennent l'estimateur pour base de sondage simple (Bankier 1986 ; Kalton et Anderson 1986 ; Skinner 1991) et l'estimateur du pseudo-maximum de vraisemblance (Skinner et Rao 1996 ; Lohr et Rao 2000 ; Lohr et Rao 2006). Lohr (2009) passe ces estimateurs en revue. Presque toutes les enquêtes téléphoniques avec chevauchement que nous avons vues utilisent certaines versions de l'estimateur moyen, qui est le point de concentration de la présente étude.

2.3 Applications aux enquêtes téléphoniques

Les données provenant de la CHIS 2007 sont utilisées pour illustrer les problèmes que pose une enquête téléphonique à base de sondage double s'appuyant sur une approche de présélection. La CHIS 2007 est une enquête téléphonique ayant pour cible la population californienne qui a été réalisée par le UCLA Center for Health Policy Research en collaboration avec le California Department of Public Health, le California Department of Health Care Services et le Public Health Institute. La collecte des données de la CHIS 2007 a été effectuée par Westat de la fin de 2007 au début de 2008.

Dans l'échantillon de numéros de téléphone fixe de la CHIS 2007, on a échantillonné et interviewé un adulte par ménage. Dans l'échantillon de numéros de téléphone mobile, les personnes vivant dans un ménage possédant un numéro de téléphone fixe ont été écartées ; un adulte par ménage a été échantillonné et interviewé dans l'échantillon de numéros de téléphone mobile s'il vivait dans un ménage classé comme possédant seulement un téléphone mobile. Tous les ménages répondants, y compris ceux éliminés de la base de numéros de téléphone mobile, ont répondu à des questions sur la situation concernant les types de téléphone et sur leur usage. Près de 49 000 adultes provenant de l'échantillon de numéros de téléphone fixe ont été interviewés, ainsi que 825 adultes possédant uniquement un téléphone mobile. Pour l'échantillon de numéros de téléphone fixe, le taux de réponse était de 35,5 % à l'interview menée auprès d'une personne fournissant des renseignements sur le ménage et de 59,4 % à l'interview auprès de l'adulte échantillonné. Pour la base de numéros de téléphone mobile, les taux de réponse correspondants étaient de 22,1 % et de 52,0 %. Puisque la CHIS 2007 a été réalisée en adoptant une approche de présélection, le taux de réponse publié pour l'interview des personnes fournissant des renseignements sur le ménage ne possédant qu'un téléphone mobile est de 30,5 %. Le document California Health Interview Survey (2009) contient une discussion détaillée du plan de l'étude, y compris les différences entre le taux de réponse global pour tous les ménages possédant un téléphone mobile et le taux pour les ménages possédant exclusivement un téléphone mobile.

Dans la CHIS 2007, les estimations pour l'échantillon de numéros de téléphone mobile sont calées sur la population californienne adulte possédant seulement un téléphone mobile à l'étape de la présélection (avant la correction de la pondération pour tenir compte de la non-réponse de l'adulte échantillonné). L'obtention de totaux de contrôle fiables pour le calage au niveau de l'État pose certaines difficultés dont il sera question plus loin. Les échantillons tirés des deux bases de sondage sont indépendants et traités en tant que tels jusqu'à la dernière étape, où ils sont combinés et calés sur des totaux indépendants pour l'ensemble de la population californienne adulte. Cette dernière étape de calage n'inclut pas le type de téléphone comme domaine.

Dans le cas des enquêtes téléphoniques à base de sondage double avec chevauchement, nous utilisons des données agrégées provenant de huit enquêtes réalisées par le Pew Research Center for the People & the Press de la fin de 2008 au début de 2009. (Les données provenant des enquêtes Pew ont été fournies par Scott Keeter, du Pew Research Center for the People & the Press). Toutes ces

enquêtes sont réalisées auprès de l'ensemble de la population adulte des États-Unis. Elles comprennent l'interview d'un adulte dans chaque ménage échantillonné dans l'une ou l'autre base de sondage en utilisant des questionnaires presque identiques. Pour l'ensemble de huit enquêtes, près de 11 300 interviews par téléphone fixe et 3 800 interviews par téléphone mobile ont été effectuées. Les taux de réponse aux diverses enquêtes sont fort semblables pour les échantillons de numéros de téléphone fixe et de téléphone mobile, avec un écart médian d'un point de pourcentage entre les échantillons provenant des deux bases de sondage. Sur l'ensemble de huit enquêtes et des deux bases de sondage, les taux de réponse varient entre 17 % et 24 %.

Dans les enquêtes Pew, comme dans la plupart des enquêtes téléphoniques à base de sondage double avec chevauchement, une version calée de l'estimateur moyen est employée pour produire les estimations. Pour la plupart des enquêtes, le calage est effectué à la fois sur les dénombrements de domaine pour le type de téléphone (nombres respectifs d'adultes vivant dans un ménage ne possédant qu'un téléphone mobile, dans un ménage ne possédant qu'un téléphone fixe et dans un ménage possédant à la fois un téléphone fixe et un téléphone mobile) ainsi que sur des variables démographiques. Pour les enquêtes Pew, le calage est également effectué sur des totaux de variables démographiques, dont l'âge, le niveau de scolarité, la race/l'ethnicité, la région et la densité de population pour les ménages comptant des adultes de 18 ans et plus. En outre, un calage est effectué sur les totaux selon le type de téléphone et, dans le domaine de chevauchement, selon l'usage relatif des téléphones fixe et mobile.

3. Erreurs non dues à l'échantillonnage

La théorie des enquêtes à base de sondage double a été élaborée sous des conditions idéales de réponse complète et d'absence d'autres erreurs non dues à l'échantillonnage. Les erreurs non dues à l'échantillonnage ont une incidence sur le biais et sur la précision des estimations de toute enquête, mais, pour trois raisons, leurs effets dans les enquêtes à base de sondage double peuvent différer qualitativement de ceux dans les enquêtes à base de sondage simple. Premièrement, dans les enquêtes à base de sondage double, l'erreur non due à l'échantillonnage rend souvent difficile la détermination de la probabilité de sélection de l'unité échantillonnée. Cela se produit quand l'appartenance au domaine est confirmée durant la collecte de données et que la non-réponse et les erreurs de mesure font qu'il est difficile de déterminer si l'unité échantillonnée se trouve dans le domaine de chevauchement. Deuxièmement, dans les enquêtes à base de sondage double, l'erreur non due à l'échantillonnage peut être reliée directement, parfois de manière causale, à la base de sondage, surtout si les méthodes de collecte de données diffèrent selon la base de sondage. Troisièmement, l'échantillonnage à partir de plus d'une base de sondage accroît la complexité et crée un plus grand nombre de situations dans lesquelles les erreurs non dues à l'échantillonnage pourraient avoir des effets différentiels.

3.1 Effets de la non-réponse

Brick, Dipko, Presser, Tucker et Yuan (2006) montrent que la surreprésentation du nombre d'adultes appartenant à un ménage muni uniquement d'un téléphone mobile qui se produit dans presque tous les échantillons d'enquête téléphonique à base de sondage double peut être due à l'erreur de non-réponse. Selon ces auteurs, la surreprésentation pourrait être le résultat d'une différence d'accessibilité, en ce sens que les adultes qui utilisent rarement leur téléphone mobile sont moins susceptibles de répondre à un appel sur celui-ci que ceux qui l'utilisent régulièrement. Ils n'ont pas constaté le même genre de différence de taux de réponse liée à l'usage dans l'échantillon de ménages dotés d'un téléphone fixe. Kennedy (2007) explore plus en détail ce type de biais de non-réponse en examinant les effets sur des estimations particulières.

Afin d'évaluer la différence de représentation, nous comparons les répartitions des échantillons de la CHIS 2007 et des enquêtes Pew selon la base de sondage et l'usage du téléphone à des estimations provenant de la National Health Interview Survey (NHIS). La NHIS est une enquête par interview sur place parrainée par le National Center for Health Statistics dont la collecte des données est effectuée par le U.S. Bureau of Census (les données de la NHIS ont été fournies par S. Blumberg et J. Luke sous forme de totalisations spéciales). Il s'agit de la seule enquête du gouvernement fédéral qui fournit des estimations sur la situation concernant le type de téléphone et l'usage de celui-ci (Blumberg et Luke 2009). Nous définissons l'usage pour les utilisateurs doubles (les membres des ménages dotés des deux types de service téléphonique) comme étant principalement mobile ou principalement fixe, la catégorie principalement mobile comprenant les personnes qui vivent dans les ménages qui reçoivent tous ou presque tous leurs appels sur leur téléphone mobile, et la catégorie principalement fixe comprenant les utilisateurs doubles dans les ménages qui ne reçoivent pas tous ou presque tous leurs appels sur leur téléphone mobile.

Pour que les chiffres soient plus comparables à ceux de la CHIS, dans le tableau 1, nous limitons les estimations de la NHIS à celles obtenues pour la région de l'Ouest seulement (les estimations de la NHIS pour la Californie ne sont pas disponibles). La Californie comprend 52 % des adultes vivant dans l'Ouest. Les chiffres de la NHIS sont des chiffres de population pour les six premiers mois de 2008,

c'est-à-dire une période correspondant approximativement à la période de collecte des données de la CHIS. Les chiffres de la CHIS correspondent aux proportions non pondérées dans l'échantillon (les proportions pondérées sont presque identiques). Bien qu'une approche de présélection ait été utilisée pour la CHIS, l'information sur l'usage du téléphone a été recueillie pour chaque ménage répondant compris dans l'échantillon de téléphones mobiles. Le tableau montre que, comparativement aux estimations de la NHIS, la répartition dans la base de numéros de téléphone mobile surreprésente le pourcentage d'adultes faisant partie de ménages dotés uniquement d'un téléphone mobile et sous-représente ceux faisant partie de ménages utilisant principalement le téléphone fixe. Les répondants de l'échantillon de numéros de téléphone fixe surreprésentent les utilisateurs d'un téléphone fixe uniquement et sous-représentent les utilisateurs doubles se servant principalement d'un téléphone mobile. Les différences pour la base de numéros de téléphone fixe sont plus importantes que celles observées dans une enquête menée en 2004 dont les résultats sont présentés dans Brick et coll. (2006).

Le tableau 2 donne le même type de comparaison des estimations nationales d'après la NHIS pour la deuxième moitié de 2008 aux résultats non pondérés des enquêtes Pew agrégées (toutes les enquêtes ont été réalisées auprès d'un échantillon à probabilités d'inclusion égales). Comme dans le cas de la CHIS, la répartition dans la base de numéros de téléphone mobile utilisée pour les enquêtes Pew surreprésente le pourcentage dans le groupe possédant un téléphone mobile seulement et sous-représente celui dans le groupe d'utilisateurs doubles se servant principalement d'un téléphone fixe, mais les différences sont moins marquées que dans le cas de la CHIS. La répartition d'après l'échantillon de numéros de téléphone fixe des enquêtes Pew correspond de près à celle observée pour la NHIS, avec une légère sous-représentation du groupe d'utilisateurs double se servant principalement d'un téléphone mobile.

Tableau 1
Répartition en pourcentage des adultes selon l'usage du téléphone d'après la CHIS de 2007 et la NHIS

Usage du téléphone	NHIS Ouest – Adultes dans les ménages à téléphone fixe	CHIS 2007 – Base de numéros de téléphone fixe	NHIS Ouest – Adultes dans les ménages à téléphone mobile	CHIS 2007 – Base de numéros de téléphone mobile
Téléphone fixe seulement	23,5 % (1,5 %)	34,2 % (0,2 %)	–	–
Double – principalement fixe	56,6 % (1,7 %)	53,2 % (0,2 %)	60,9 % (1,7 %)	18,5 % (0,7 %)
Double – principalement mobile	19,9 % (1,4 %)	12,7 % (0,2 %)	21,4 % (1,4 %)	31,2 % (0,9 %)
Téléphone mobile seulement	–	–	17,7 % (1.3 %)	50,3 % (0,9 %)
Total	100,0 %	100,0 %	100,0 %	100,0 %

Notes NHIS-Ouest désigne la National Health Interview Survey, région de l'Ouest, six premiers mois de 2008, avec les pourcentages de tous les ménages possédant le type de service particulier (merci à S. Blumberg et J. Luke pour cette totalisation spéciale). CHIS 2007 représente la California Health Interview Survey, dont les données ont été recueillies en 2007 et au début de 2008, avec les pourcentages non pondérés d'après les bases de numéros de téléphone fixe et de numéros de téléphone mobile. Dans l'échantillon de téléphones mobiles, l'information sur l'usage a été obtenue durant l'interview de présélection. Les erreurs-types approximatives figurent entre parenthèses.

Tableau 2
Répartition en pourcentage des adultes selon l'usage du téléphone d'après les enquêtes Pew et la NHIS

Usage du téléphone	NHIS – Adultes dans les ménages à téléphone fixe	Enquêtes Pew – Base de numéros de téléphone fixe	NHIS – Adultes dans les ménages à téléphone mobile	Enquêtes Pew – Base de numéros de téléphone mobile
Téléphone fixe seulement	19,4 % (0,7 %)	23,0 % (0,4 %)	–	–
Double – principalement fixe	58,8 % (0,8 %)	62,7 % (0,5 %)	58,8 % (0,8 %)	42,3 % (0,8 %)
Double – principalement mobile	19,3 % (0,7 %)	14,4 % (0,3 %)	18,5 % (0,7 %)	24,0 % (0,7 %)
Téléphone mobile seulement	–	–	22,7 % (0,7 %)	33,7 % (0,8 %)
Total	100,0 %	100,0 %	100,0 %	100,0 %

Notes NHIS désigne la National Health Interview Survey, six derniers mois de 2008, avec les pourcentages de tous les ménages possédant le type de service particulier. Les enquêtes Pew regroupent huit enquêtes réalisées pour le Pew Research Center for the People & the Press d'octobre 2008 à mars 2009, avec les pourcentages non pondérés provenant des bases de numéros de téléphone fixe et de numéros de téléphone mobile. (Merci à S. Keeter d'avoir fourni ces données). Les erreurs-types approximatives figurent entre parenthèses.

Les répartitions des réponses selon la base de sondage et selon l'usage du téléphone que l'on observe pour ces enquêtes corroborent, dans les deux cas, la conjecture concernant l'accessibilité de Brick et coll. (2006). Cette conjecture implique un classement des répondants selon le degré d'accessibilité et la probabilité de répondre – dans la base de numéros de téléphone mobile, le classement allant des personnes les plus susceptibles de répondre à celles les moins susceptibles de le faire est le suivant : téléphone mobile seulement, principalement le téléphone mobile et principalement le téléphone fixe. Le problème particulier que pose l'utilisation de deux bases de sondage est que le classement dans la base des numéros de téléphone fixe est différent (téléphone fixe seulement, principalement téléphone fixe et principalement téléphone mobile) et que les unités comprises dans le domaine de chevauchement des deux bases de sondage pourraient avoir des taux de réponse et des biais très différents.

Pour examiner le biais de non-réponse dans le cas d'une enquête à base de sondage double avec chevauchement, supposons que l'on poststratifie les échantillons de numéros de téléphone fixe et de numéros de téléphone mobile selon les totaux de domaine pour la situation concernant le type de téléphone avant de produire une estimation globale moyenne. L'estimateur poststratifié est donné par

$$\hat{y}_{ps} = \frac{N_a}{\hat{N}_a}\hat{y}_a + \frac{N_b}{\hat{N}_b}\hat{y}_b + \lambda g^A \hat{y}_{ab}^A + (1-\lambda) g^B \hat{y}_{ab}^B, \quad (1)$$

où le facteur de poststratification est $N_a / \hat{N}_a$, pour l'échantillon de numéros de téléphone fixe seulement, $N_b / \hat{N}_b$, pour l'échantillon de numéros de téléphone mobile seulement, et les facteurs de poststratification propres à la base de sondage pour le domaine de chevauchement sont $g^A = N_{ab} / \hat{N}_{ab}^A$ et $g^B = N_{ab} / \hat{N}_{ab}^B$ pour les échantillons de numéros de téléphone fixe et de téléphone mobile, respectivement. Les estimateurs d'Horvitz-Thompson (HT) du nombre d'unités sont $\hat{N}_a$ pour le domaine des téléphones fixes seulement, $\hat{N}_b$ pour le domaine des téléphones mobiles seulement, et $\hat{N}_{ab}^A$ et $\hat{N}_{ab}^B$ pour le domaine de chevauchement des deux échantillons. Puisque nous nous concentrons sur le chevauchement, nous écrivons

$$\hat{y}_{ps,ab} = \lambda g^A \hat{y}_{ab}^A + (1-\lambda) g^B \hat{y}_{ab}^B. \quad (2)$$

Cet estimateur poststratifié diffère de l'approche proposée par Lohr et Rao (2000), qui calculent la moyenne puis font la poststratification, au lieu de faire la poststratification puis de calculer la moyenne. Les deux approches sont cohérentes et approximativement sans biais en l'absence d'erreurs non dues à l'échantillonnage.

Si nous tenons compte, dans le domaine de chevauchement, de différences de taux de réponse selon l'usage du téléphone telles que celles observées dans les enquêtes téléphoniques à base de sondage double, l'estimateur (2) est biaisé. Soit W la proportion du domaine de chevauchement correspondant à des unités utilisant le téléphone fixe principalement, et soit $\bar{Y}_{ml}$ et $\bar{Y}_{mc}$ les moyennes de population d'une caractéristique pour les utilisateurs doubles se servant principalement du téléphone fixe (désignés par *ml* pour l'anglais *mainly land*) et ceux se servant principalement du téléphone mobile (désignés par *mc* pour l'anglais *mainly cell*), respectivement. Le biais de $\hat{y}_{ps,ab}$ est

$$b(\hat{y}_{ps,ab}) \doteq WN_{ab}(\bar{Y}_{ml} - \bar{Y}_{mc}) \\ (\lambda r_{l1} r_l^{-1} + (1-\lambda) r_{c1} r_c^{-1} - 1), \quad (3)$$

où r_l est le taux de réponse des utilisateurs doubles pour l'échantillon de numéros de téléphone fixe, r_{l1} est le taux de réponse des utilisateurs doubles se servant principalement du téléphone fixe dans l'échantillon de numéros de téléphone fixe, r_c est le taux de réponse des utilisateurs doubles dans l'échantillon de numéros de téléphone mobile, et r_{c1} est le taux de réponse des utilisateurs doubles se servant principalement du téléphone fixe dans l'échantillon de numéros de téléphone mobile.

Pour dériver (3), nous commençons par définir les estimateurs de domaine comprenant les utilisateurs doubles se servant principalement du téléphone fixe et se servant principalement du téléphone mobile provenant de l'échantillon de numéros de téléphone fixe comme étant $\hat{y}_{ab}^A(ml) = \hat{N}_{ml}^A \bar{y}_{ab}^A(ml)$ et $\hat{y}_{ab}^A(mc) = \hat{N}_{mc}^A \bar{y}_{ab}^A(mc)$, ceux et provenant de l'échantillon de numéros de téléphone mobile comme étant $\hat{y}_{ab}^B(ml) = \hat{N}_{ml}^B \bar{y}_{ab}^B(ml)$ et $\hat{y}_{ab}^B(mc) = \hat{N}_{mc}^B \bar{y}_{ab}^B(mc)$. Maintenant, supposons que a) $E\bar{y}_{ab}^A(ml) = E\bar{y}_{ab}^B(ml) = \bar{Y}_{ml}$ et $E\bar{y}_{ab}^A(mc) = E\bar{y}_{ab}^B(mc) = \bar{Y}_{mc}$; b) les covariances sont telles que $\mathrm{cov}(\hat{N}_{ml}^A / \hat{N}_{ab}^A, \bar{y}_{ab}^A(ml)) = 0$; et c) les totaux de domaine prévus sont des expressions simples telles que $E\hat{N}_{ml}^A = r_{l1} N_{ml}, E\hat{N}_{ab}^A = r_l N_{ab}$, *etc.* Puisque $E(N_{ab} / \hat{N}_{ab}^A)$ $\hat{y}_{ab}^A = N_{ab} E\{(\hat{N}_{ml}^A \bar{y}_{ab}^A(ml) + \hat{N}_{mc}^A \bar{y}_{ab}^A(mc)) / \hat{N}_{ab}^A\}$, nous pouvons écrire $E(N_{ab} / \hat{N}_{ab}^A) \hat{y}_{ab}^A \doteq r_{l1} r_l^{-1} N_{ml} \bar{Y}_{ml} + r_{l2} r_l^{-1} N_{mc}$ $\bar{Y}_{mc} = N_{ab}(r_{l1} r_l^{-1} W(\bar{Y}_{ml} - \bar{Y}_{mc}) + \bar{Y}_{mc})$. Une expression correspondante peut être écrite pour $Eg^B \hat{y}_{ab}^B$. La combinaison des deux donne (3).

Ces expressions supposent que $E\bar{y}_{ab}^B(ml) = \bar{Y}_{ml}$ et $E\bar{y}_{ab}^B(mc) = \bar{Y}_{mc}$. Une autre approche qui ne nécessite pas cette hypothèse consiste à poser qu'il existe une association entre la propension à répondre et l'usage du téléphone. Dans ce cas, le biais serait une fonction de la propension à répondre des unités provenant de chaque base de sondage. Nous n'examinons pas l'approche de la propension à répondre ici.

L'expression (3) montre que, quand $0 < W < 1$, le biais de $\hat{y}_{ps,ab}$ est nul si a) $\bar{Y}_{ml} = \bar{Y}_{mc}$; ou b) $\lambda r_{l1} r_l^{-1} + (1-\lambda)$

$r_{c1} r_c^{-1} = 1$. La condition (a) est essentiellement la condition bien connue issue de la méthode à base unique. La condition (b) diffère des expressions pour une base de sondage unique parce que le biais dépend à la fois des taux de réponse relatifs et du facteur de composition, λ. Une exception a lieu quand $r_{l1} r_l^{-1} = r_{c1} r_c^{-1}$, ou de manière équivalente $r_{l1} r_{l2}^{-1} = r_{c1} r_{c2}^{-1}$, où r_{l2} est le taux de réponse des unités utilisant principalement le téléphone mobile dans l'échantillon de numéros de téléphone fixe et r_{c2} est le taux de réponse des unités utilisant principalement le téléphone mobile dans l'échantillon de numéros de téléphone mobile. Sous cette forme, l'expression est comparable à celle du biais pour une base de sondage unique qui montre que le biais est nul quand les taux de réponse sont constants.

De façon plus générale, la valeur de λ influe non seulement sur la variance de l'estimation, mais aussi sur son biais. Celui-ci peut être éliminé en choisissant

$$\lambda_0 = \frac{r_l (r_c - r_{c1})}{r_c r_{l1} - r_l r_{c1}}. \qquad (4)$$

Puisque la proportion du total de population couverte par la base de numéros de téléphone fixe est approximativement égale à la proportion couverte par la base de numéros de téléphone mobile, la valeur de $\lambda = 0{,}50$ a été utilisée dans la plupart des applications sans tenir compte de l'effet sur le biais.

Nous pouvons maintenant appliquer ces expressions pour évaluer le biais de l'estimateur sous enquête téléphonique à base de sondage double dans le cas de la CHIS, en supposant que ce biais est dû seulement à la différence de non-réponse dans le domaine de chevauchement. Si nous utilisons les données du tableau 1, $W = 0{,}74$ pour la région Ouest selon la NHIS. Nous approximons $r_{l1} r_l^{-1}$ par le facteur de poststratification relatif qui est le ratio du pourcentage d'unités de l'échantillon de numéros de téléphone fixe de la CHIS classées comme utilisant principalement le téléphone fixe au pourcentage d'adultes participant à la NHIS dans les ménages ayant un téléphone fixe qui utilisent principalement le téléphone fixe ; nous calculons de la même façon $r_{c1} r_c^{-1}$ pour les quantités ayant trait au téléphone mobile. Les quantités estimées pour la CHIS 2007 sont présentées au tableau 3, $r_{l1} r_l^{-1} \doteq 1{,}09$ pour l'échantillon de numéros de téléphone fixe et $r_{c1} r_c^{-1} \doteq 0{,}50$ pour l'échantillon de numéros de téléphone mobile. À titre d'exemple, si nous supposons que $\bar{Y}_{ml} = 0{,}3$ et $\bar{Y}_{mc} = 0{,}5$, le biais du pourcentage estimé basé sur (3) est de l'ordre de 3 points de pourcentage (un biais relatif d'environ 9 %) si $\lambda = 0{,}5$. En utilisant (4), le biais est nul quand $\lambda \doteq 0{,}84$; le biais devient négatif pour les valeurs plus grandes de λ.

Les mêmes calculs peuvent être effectués en utilisant les données provenant des enquêtes Pew, et les estimations sont également présentées au tableau 3. Les paramètres diffèrent considérablement de ceux calculés pour la CHIS. Puisque les études Pew sont nationales, l'estimation de la NHIS est $W = 0{,}81$. Les ratios des chiffres Pew à celui de la NHIS ont également une plus faible variabilité que ceux calculés pour la CHIS, avec $r_{l1} r_l^{-1} \doteq 1{,}07$ et $r_{c1} r_c^{-1} \doteq 0{,}84$. Par conséquent, le biais est de l'ordre de 1 point de pourcentage seulement quand $\lambda = 0{,}5$. Le biais est nul quand $\lambda \doteq 0{,}7$.

Tableau 3
Facteurs de poststratification relatifs dans le domaine de chevauchement pour la CHIS 2007 et les enquêtes Pew

Facteurs de poststratification relatifs*	CHIS 2007	Enquêtes Pew
$r_{l1} r_l^{-1} \doteq g^A / g^A{}_{ml}$	1,09	1,07
$r_{l2} r_l^{-1} \doteq g^A / g^A{}_{mc}$	0,50	0,84
$r_{c1} r_c^{-1} \doteq g^B / g^B{}_{ml}$	0,74	0,78
$r_{c2} r_c^{-1} \doteq g^B / g^B{}_{mc}$	2,42	1,51

* Facteur d'ajustement par poststratification pour le domaine d'usage du téléphone dans le domaine de chevauchement divisé par le facteur de poststratification du domaine de chevauchement.

Pour évaluer les biais complètement, les estimations de $\bar{Y}_{ml} - \bar{Y}_{mc}$ sont nécessaires pour des caractéristiques étudiées au moyen d'une enquête téléphonique à base de sondage double plutôt qu'en formulant des hypothèses arbitraires comme dans l'exemple qui précède. Blumberg et Luke (2009) donnent des estimations qui donnent à penser que ces différences peuvent être aussi importantes que les celles entre la population d'utilisateurs de téléphone mobile seulement et d'utilisateurs de téléphone fixe qui ont été décrites en détail ailleurs. Cependant, les estimations de la NHIS sont calculées au moyen de données venant d'une enquête avec interview sur place et non d'une enquête téléphonique à base de sondage double.

Keeter, Dimock et Christian (2008) donnent des estimations des caractéristiques des utilisateurs des deux types de téléphone selon la base de sondage, mais pas de façon suffisamment détaillée pour pouvoir calculer les biais. Les données de Keeter indiquent que les estimations pour les utilisateurs doubles sélectionnés dans la base de numéros de téléphone mobile pourraient être plus proches des estimations de la NHIS pour le domaine de chevauchement que les estimations pour les utilisateurs doubles tirés de la base de numéros de téléphone fixe. Cependant, puisque les taux de réponse dans le domaine de chevauchement sont plus variables pour les unités provenant de la base de numéros de téléphone mobile que pour celles provenant de la base de téléphone fixe, un plan de sondage avec présélection visant à réduire le biais devrait exclure les utilisateurs doubles provenant de la base de numéros de téléphone mobile plutôt que de la base de numéros de téléphone fixe quand la

variabilité des taux de réponse est plus grande dans la base de numéros de téléphone mobile.

En raison du biais éventuel dans le plan de sondage avec chevauchement, Brick et coll. (2006) proposent d'utiliser un plan de sondage avec présélection conçu pour exclure les adultes appartenant à des ménages utilisant les deux types de téléphone s'ils ont été échantillonnés dans la base de numéros de téléphone mobile. Dans un plan avec présélection, un biais persiste à cause de la différence de non-réponse chez les utilisateurs des deux types de téléphone selon l'usage du téléphone dans l'échantillon de numéros de téléphone fixe. Si nous substituons $\lambda = 1$ dans (2) et (3), le biais de $\hat{y}_{scr,ab} = g^A \hat{y}_{ab}^A$ est donné par

$$b(\hat{y}_{scr,ab}) = WN_{ab}(\overline{Y}_{ml} - \overline{Y}_{mc})(r_{l1}\, r_l^{-1} - 1). \qquad (5)$$

Le biais pour ce plan de sondage et cet estimateur équivaut à celui des estimateurs pour base de sondage unique, et il disparaît si $\overline{Y}_{ml} = \overline{Y}_{mc}$ ou que les taux de réponse dans l'échantillon de numéros de téléphone fixe sont les mêmes pour les utilisateurs se servant principalement du téléphone fixe et ceux se servant principalement du téléphone mobile. Soulignons que, sous ce plan de sondage, aucun facteur de composition ne peut être utilisé pour contrôler le biais.

Le biais de l'estimateur avec présélection pour la CHIS 2007 est à peu près égal à la moitié de celui de l'estimateur moyen en utilisant $\lambda = 0{,}50$ (le biais de présélection est de 1,3 point de pourcentage comparativement à celui de l'estimateur moyen poststratifié en utilisant $\lambda = 0{,}50$ qui est de -3,3 points). Pour les paramètres des enquêtes Pew, le biais de l'estimateur moyen poststratifié et celui de l'estimateur avec présélection sont presque égaux, le biais de l'estimateur avec présélection étant légèrement plus grand que celui de l'estimateur poststratifié (le biais pour l'estimateur avec présélection est de 1,1 point de pourcentage comparativement à -0,7 point pour le chevauchement poststratifié).

Un problème mentionné plus haut tient au fait que les totaux de domaine pour la poststratification, même si l'on ne considère que la situation concernant le type de téléphone (domaines des utilisateurs du téléphone fixe seulement, du téléphone mobile seulement et des deux types de téléphone) ne sont généralement pas disponibles pour les enquêtes au niveau de l'État ou au niveau local. Des estimations sur petits domaines du pourcentage d'adultes qui utilisent un téléphone mobile seulement ont été publiées au niveau de l'État (Blumberg, Luke, Davidson, Davern, Yu et Soderberg 2009), mais ces auteurs ne fournissent pas d'estimations sur petits domaines pour les trois domaines d'utilisateurs. Pour les totaux de contrôle de l'usage du téléphone, la situation est encore plus limitée, seules les estimations nationales de la NHIS étant publiées. Puisque, pour la base de sondage des numéros de téléphone mobile, les taux de réponse varient habituellement selon l'usage, certaines hypothèses au sujet des taux de réponse dans l'échantillon de numéros de téléphone mobile peuvent être utiles pour éviter une surreprésentation importante des adultes utilisant seulement ou utilisant principalement un téléphone mobile dans l'échantillon tiré de la base de sondage de numéros de téléphone mobile quand on utilise le plan de sondage avec chevauchement.

3.2 Effets des erreurs de mesure

En plus de la non-réponse, certaines différences entre les distributions observées dans les tableaux 1 et 2 pourraient être dues à une erreur de mesure. Avant de parler des hypothèses associées à l'erreur de mesure, nous discutons de certaines procédures importantes des enquêtes qui pourraient être reliées à cette erreur. Il s'agit de différences fondamentales entre les enquêtes, telles que le mode de collecte et le thème. La NHIS est une enquête avec interview sur place, tandis que la CHIS et les enquêtes Pew sont des enquêtes téléphoniques. La NHIS ainsi que la CHIS sont des enquêtes sur la santé, tandis que les enquêtes Pew couvrent une grande gamme de sujets.

Ces enquêtes s'appuient aussi sur des méthodes différentes pour recueillir l'information sur la situation en ce qui concerne le type de téléphone ainsi que sur l'usage du téléphone. Dans la NHIS, on demande à un membre adulte de la famille de répondre à des questions sur la situation concernant le type de téléphone et sur l'usage du téléphone pour la famille complète durant une partie de l'interview réservée aux caractéristiques de la famille. Dans le cas de l'échantillon de numéros de téléphone mobile de la CHIS 2007, des questions sur la situation concernant le type de téléphone sont posées durant l'interview de présélection du ménage, mais les questions sur l'usage du téléphone sont posées durant l'interview de l'adulte sélectionné pour participer à l'enquête. Dans le cas de l'échantillon de numéros de téléphone fixe de la CHIS et des enquêtes Pew, les questions sur la situation concernant le type de téléphone et sur l'usage du téléphone sont toutes regroupées dans l'une des dernières parties de l'interview de la personne adulte sélectionnée. Ce placement des questions à un moment plus tardif est possible parce qu'aucune présélection n'est effectuée.

L'échantillonnage d'un adulte est une autre procédure qui peut interagir avec le processus de mesure. Dans la CHIS 2007, un adulte est sélectionné parmi tous ceux partageant le même numéro de téléphone mobile. Dans les enquêtes Pew, et dans la plupart des autres enquêtes par téléphone mobile, ce dernier est considéré comme un appareil personnel et la personne qui y répond est interviewée. Dans les ménages utilisant les deux types de téléphone, les

méthodes appliquées pour la CHIS et les enquêtes Pew peuvent donner lieu à différents échantillons d'adultes.

La source potentielle d'erreur de mesure la plus importante est peut-être liée aux différences entre les questions sur la situation concernant le type de téléphone et l'usage du téléphone utilisées dans les enquêtes. Les questions posées dans chaque enquête sont présentées en annexe. Les approches varient assez bien. Au moins une partie des différences observées dans les études est due au fait que la CHIS et les enquêtes Pew sont menées par téléphone et que l'on possède des renseignements préalables sur la situation concernant le type de téléphone utilisé.

Les questions employées dans les trois enquêtes sont dérivées de questions utilisées dans un supplément à la Current Population Survey (CPS) de 2004. Comme l'exposent Tucker, Brick et Meekins (2007), des essais cognitifs et de codage du comportement effectués pour le supplément ont suscité certaines réserves concernant les questions de la CPS, surtout la question sur l'usage. Les essais ont révélé que l'absence d'une période de référence précise, l'absence d'un code pour « la moitié du temps » (*half the time*) et la difficulté à fournir la réponse pour d'autres membres du ménage faisaient que la question sur l'usage risquait d'être sujette à une erreur de mesure. Tucker et coll. (2007) soulignent aussi la difficulté qu'avaient les répondants à indiquer la situation concernant le type de téléphone et l'usage pour tous les membres du ménage au moyen d'une seule question. En outre, les répondants avaient de la difficulté à comprendre la signification de « téléphone fixe » (*landline*), « ordinaire » (*regular*), téléphone mobile « qui fonctionne » ("*working*" *cell phone*), ainsi que la différence entre utiliser un téléphone mobile et répondre à un téléphone mobile.

Ces questions pourraient avoir une incidence sur la classification par domaine et donc biaiser les estimations. Par exemple, un jeune de 23 ans vivant avec ses parents pourrait déclarer utiliser uniquement un téléphone mobile, tandis que les parents pourraient déclarer utiliser les deux types de téléphone. Les effets de ces types d'erreur de mesure sur les estimations sont difficiles à prédire dans le cas de la NHIS et des enquêtes téléphoniques, mais il n'est pas inattendu d'observer des déclarations non concordantes dans les enquêtes par téléphone et sur place.

Un autre problème de mesure éventuel tient à la relation entre la déclaration concernant l'usage du téléphone et la base de sondage dont proviennent les répondants. L'erreur hypothétique survient si, quand on lui demande quel appareil il utilise pour la plupart de ses appels, le répondant est plus enclin à choisir l'appareil qu'il utilise pour répondre à l'interview. Autant que nous sachions, cette hypothèse n'a pas été testée, mais tout effet d'appareil de cette nature devrait en principe se manifester dans le même sens que l'effet de non-réponse. Un utilisateur des deux types de téléphone devrait avoir une plus forte probabilité de déclarer qu'il se sert principalement du téléphone mobile s'il est échantillonné dans la base de numéros de téléphone mobile et être plus susceptible de déclarer qu'il utilise principalement le téléphone fixe s'il est échantillonné dans la base de numéros de téléphone fixe. Donc, le biais dont nous avons discuté plus haut dans le contexte de la non-réponse pourrait être le résultat de l'effet combiné de la non-réponse et de l'effet d'appareil. S'il n'est pas possible de déterminer l'importance de ces sources de biais, le choix de méthodes pour réduire ce biais n'est pas clair.

4. Approches d'élaboration du plan de sondage et d'estimation en présence d'erreurs non dues à l'échantillonnage

Étant donné les problèmes supplémentaires qui entrent en jeu dans les enquêtes à base de sondage double, les méthodes d'échantillonnage et d'estimation doivent être conçues de manière à tenir compte des sources les plus importantes d'erreur au lieu de se concentrer uniquement sur l'erreur d'échantillonnage. À la présente section, nous examinons les options concernant le plan de sondage et l'estimation pour les enquêtes téléphoniques à base de sondage double dans ces conditions générales de structure d'erreur.

4.1 Approches concernant le plan de sondage

Une décision essentielle concernant le plan de sondage dans le cas d'une enquête téléphonique à base de sondage double est celle de savoir si l'on doit utiliser un plan de sondage avec présélection ou avec chevauchement complet. Nous commençons par examiner la répartition optimale de l'échantillon pour les plans avec chevauchement et avec présélection appropriés pour les enquêtes téléphoniques à base de sondage double quand des échantillons aléatoires simples sont tirés indépendamment des deux bases de sondage et que $N_a > 0$, $N_b > 0$, et $N_{ab} > 0$. Nous supposons tout au long de l'exposé que les tailles d'échantillon sont suffisamment grandes pour que l'on puisse ignorer les facteurs de correction pour population finie.

Nous utilisons une fonction de coût prévu linéaire $E(C) = c_A(n_A + n_B c_B c_A^{-1})$, où c_A est le coût d'une interview par téléphone fixe, c_B est le coût d'une interview par téléphone mobile, et n_A et n_B sont les nombres d'unités échantillonnées dans les bases de sondage *A* et *B*, respectivement. Si nous supposons que la variance élémentaire, σ^2, est constante, la variance de l'estimateur avec chevauchement est $v_{ov}^2 = \sigma^2(N_A(N_a + \lambda^2 N_{ab})n_A^{-1} + N_B(N_b + (1-\lambda)^2 N_{ab})n_B^{-1})$. La répartition de l'échantillon qui minimise la variance en utilisant cette fonction de coût peut être trouvée en se servant de méthodes Lagrangiennes classiques et est

$$n_{o,A} = E(C)\tau^{-1}\sqrt{c_A^{-1}N_A(N_a + \lambda^2 N_{ab})}$$

$$n_{o,B} = E(C)\tau^{-1}\sqrt{c_B^{-1}N_B(N_b + (1-\lambda)^2 N_{ab})}, \quad (6)$$

où

$$\tau = \sqrt{c_A N_A(N_a + \lambda^2 N_{ab})} + \sqrt{c_B N_B(N_b + (1-\lambda)^2 N_{ab})}.$$

Pour un plan de sondage avec présélection, une fonction de coût linéaire appropriée pour les enquêtes téléphoniques à base de sondage double est donnée par $E(C) = c_A n_A + n_b c_b$, où $c_b = c_B + N_B N_b^{-1} c_s$, n_b est le nombre échantillonné d'utilisateurs d'un téléphone mobile seulement, et c_s est le coût de la présélection. La variance de l'estimateur avec présélection est $v_{sc}^2 = \sigma^2(N_A^2 n_A^{-1} + N_B N_b n_B^{-1})$. La répartition optimale est simplement la répartition stratifiée donnée par $n_{s,A} = E(C)N_A(c_A N_A + \sqrt{c_A c_b}N_b)^{-1}$ et

$$n_B = \frac{E(C)N_B}{\sqrt{c_A c_b}N_A + c_b N_b},$$

qui produit

$$n_b = \frac{E(C)N_b}{\sqrt{c_A c_b}N_A + c_b N_b}$$

interviews d'utilisateur d'un téléphone mobile seulement.

S'il n'existe pas d'erreur non due à l'échantillonnage et que le coût prévu est fixe, la variance pour le plan avec chevauchement sous répartition optimale est plus faible que la variance pour le plan avec présélection sous répartition optimale quand le coût de la présélection est suffisamment grand pour que $\sqrt{c_b} > N_b^{-1}(\tau - N_A\sqrt{c_A})$. Si l'on inclut le biais, le plan de sondage avec présélection peut donner une plus petite erreur quadratique moyenne que le plan avec chevauchement, même si cette condition est vérifiée. Dans l'analyse qui suit, nous considérons le biais, mais nous ne tenons pas compte de tous les effets de l'erreur non due à l'échantillonnage. Par exemple, une réponse différentielle influe sur le rendement de la base de sondage dont ont été tirées les unités, ce qui affecte la répartition de l'échantillon et la variance de l'estimation.

Nous comparons les erreurs quadratiques moyennes des plans de sondage avec présélection et avec chevauchement sous les paramètres de la CHIS 2007 donnés précédemment. L'erreur quadratique moyenne est égale à la somme de la variance et du biais au carré. La variance est calculée pour l'estimation globale, mais le biais a seulement pour origine le chevauchement des deux bases de sondage sous nos hypothèses. Les paramètres de coût d'interview et de présélection des unités possédant un téléphone mobile ne sont pas encore bien connus, mais nous utilisons ($c_A = 1$, $c_B = 3$, $c_s = 2$) en nous basant sur l'information donnée par Keeter et coll. (2008) et par Edwards, Brick et Grant (2008). Les autres paramètres nécessaires pour la comparaison sont la répartition de la population par domaine selon le type de téléphone utilisé, et nous approximons les valeurs nationales provenant de la NHIS de 2008 ($N_a = 0,2N$, $N_b = 0,2N$ et $N_{ab} = 0,6N$). Dans cette situation, la variance fondée sur un plan de sondage avec chevauchement sous répartition optimale de l'échantillon avec $\lambda = 0,5$ est légèrement plus faible que la variance pour le plan avec présélection sous répartition optimale (le ratio des variances est égal à 0,976). Les variances des deux plans de sondage sont presque les mêmes quand les paramètres de coût sont tels que la présélection des unités provenant de la base de sondage *B* est un peu moins coûteuse ($c_A = 1$, $c_B = 3$, $c_s = 1,85$).

L'approche de la présélection produit une plus petite erreur quadratique moyenne que l'approche du plan de sondage avec chevauchement sous ces conditions, parce que la première réduit le biais des estimations pour passer de -3,3 points de pourcentage à 1,3 point. Même un biais relativement faible domine la comparaison des erreurs quadratiques moyennes entre les deux plans de sondage, en supposant que le biais sous l'approche de présélection est égal à la moitié du biais sous l'approche du chevauchement. Il en est ainsi parce que les variances des plans avec chevauchement et avec présélection sont très similaires. Cependant, si nous utilisons les paramètres provenant des enquêtes Pew, l'erreur quadratique moyenne pour le plan avec chevauchement est plus faible, parce que le biais sous ce plan est plus faible que le biais sous le plan avec présélection.

La répartition entre les bases de sondage sous l'approche avec chevauchement donnée par l'expression (6) en supposant la présence d'une erreur d'échantillonnage seulement est déterminée par les paramètres de population, les paramètres de coût et le facteur de composition. Bien qu'il ne s'agisse pas de la répartition optimale si l'on admet qu'il existe des taux de réponse différentiels, son examen est néanmoins utile puisqu'il est probable qu'elle soit observée fréquemment en pratique. Dans cette situation, le biais de $\hat{y}_{ps,ab}$ dû à la non-réponse différentielle peut être éliminé en choisissant λ de manière à satisfaire (4). En se fondant sur les paramètres de la CHIS, la valeur qui élimine ce biais est $\lambda \doteq 0,84$. Si nous poursuivons maintenant les hypothèses de coût et de population susmentionnées, mais en fixant $\lambda = 0,84$, la répartition optimale donnée par (6) donnera alors lieu à la sélection d'environ 75 % de l'échantillon dans la base de numéros de téléphone fixe. Comparativement, si l'on considère la répartition avec $\lambda = 0,5$, 63 % seulement de l'échantillon provient de cette base de sondage. Le choix du facteur de composition est critique. Si l'on utilise $\lambda = 0,84$ conjugué à la répartition optimale pour les paramètres de la CHIS, l'estimateur est sans biais et possède une variance qui est environ 5 % plus faible que celle de l'estimateur sous le plan avec présélection optimal.

4.2 Approches d'estimation

Une approche proposée par Brick et coll. (2006) consiste à utiliser un plan de sondage avec chevauchement complet ainsi qu'un estimateur moyen pour le chevauchement post-stratifié selon les totaux de domaine d'usage du téléphone, comme dans les enquêtes Pew. Cet estimateur est sans biais et convergent si les estimations dans les domaines sont sans biais et que les tailles d'échantillon de domaine sont suffisamment grandes.

Les données auxiliaires nécessaires pour cette post-stratification pour l'ensemble des États-Unis sont maintenant publiées régulièrement d'après la NHIS. Comme nous l'avons mentionné plus haut, certaines réserves quant à l'utilisation de ces données comme totaux de contrôle méritent d'être examinées plus en détail. Les totaux de contrôle nécessaires pour cet estimateur sont le nombre d'adultes se servant d'un téléphone fixe seulement, le nombre d'adultes se servant d'un téléphone mobile seulement, ainsi que les nombres d'adultes qui se servent principalement d'un téléphone fixe et de ceux qui se servent principalement d'un téléphone mobile (N_{ml} et N_{mc}, respectivement). Cela répartit les utilisateurs des deux types de téléphones en deux composantes.

Un autre estimateur du total du domaine de chevauchement en utilisant les mêmes données auxiliaires est donné par

$$\begin{aligned}\hat{y}_{sep} &= \frac{N_a}{\hat{N}_a}\hat{y}_a + \frac{N_b}{\hat{N}_b}\hat{y}_b + \lambda_1 g_{ml}^A \hat{y}_{ab}^A(ml) \\ &\quad + (1-\lambda_1) g_{ml}^B \hat{y}_{ab}^B(ml) \\ &\quad + \lambda_2 g_{mc}^A \hat{y}_{ab}^A(mc) + (1-\lambda_2) g_{mc}^B \hat{y}_{ab}^B(mc), \end{aligned} \quad (7)$$

où les facteurs de poststratification détaillés sont $g_{ml}^A = N_{ml}/\hat{N}_{ml}^A$, $g_{mc}^A = N_{mc}/\hat{N}_{mc}^A$, $g_{ml}^B = N_{ml}/\hat{N}_{ml}^B$, $g_{mc}^B = N_{mc}/\hat{N}_{mc}^B$, $0 \le \lambda_1 \le 1$ et $0 \le \lambda_2 \le 1$. Cet estimateur, comme les autres considérés jusqu'à présent, est sans biais et convergent en l'absence d'erreurs non dues à l'échantillonnage. Comme pour l'estimateur (1), les estimations provenant de chaque base de sondage sont poststratifiées avant que l'on calcule leur moyenne. La différence principale entre (1) et (7) est que, dans (7), le domaine des utilisateurs des deux types de téléphone est subdivisé et poststratifié selon l'usage ; ce dernier estimateur introduit aussi des facteurs de composition différents dans le domaine de chevauchement.

L'estimateur $\hat{y}_{sep}$ peut être utile quand (1) présente un biais et que les totaux de contrôle de l'usage sont disponibles pour la poststratification. Si les moyennes prévues dans les domaines d'usage sont approximativement égales ($E\bar{y}_{ab}^A(ml) = E\bar{y}_{ab}^B(ml) = \bar{Y}_{ml}$ et $E\bar{y}_{ab}^A(mc) = E\bar{y}_{ab}^B(mc) = \bar{Y}_{mc}$), alors (7) est sans biais pour tout choix de $0 \le \lambda_1 \le 1$ et $0 \le \lambda_2 \le 1$. Puisque le biais n'est pas affecté par le choix, différents facteurs de composition peuvent être utilisés pour réduire la variance des estimations comme il est habituellement proposé de le faire dans la littérature sur les bases de sondage doubles. Le tableau 3 montre que la proportion de répondants dans les domaines d'usage détaillés varie considérablement selon la base de sondage, ce qui pourrait rendre utile l'utilisation de divers facteurs de composition.

Les totaux de contrôle de l'usage du téléphone n'étant souvent pas disponibles, nous avons étudié la modification de (2) afin d'utiliser divers facteurs de composition semblables à ceux employés dans le chevauchement pour (7). Dans ce cas, l'objectif serait de réduire le biais plutôt que la variance. Un estimateur modifié du total dans le domaine de chevauchement est donné par

$$\begin{aligned}\hat{y}_{\mathrm{mod},ab} &= \lambda_1 g^A \hat{y}_{ab}^A(ml) + (1-\lambda_1) g^B \hat{y}_{ab}^B(ml) \\ &\quad + \lambda_2 g^A \hat{y}_{ab}^A(mc) + (1-\lambda_2) g^B \hat{y}_{ab}^B(mc). \end{aligned} \quad (8)$$

Cependant, cet estimateur pourrait ne pas être utile pour la réduction du biais. Précédemment, nous avons montré que le biais de $\hat{y}_{ps,ab}$ disparaît quand $\lambda_0 = r_l(r_c - r_{c1})(r_c r_{l1} - r_l r_{c1})^{-1}$. Le choix de $\lambda_1 = \lambda_2 = \lambda_0$ dans (8) élimine le biais à la fois pour les estimations produites pour les personnes se servant principalement d'un téléphone fixe et celles produites pour les personnes se servant principalement d'un téléphone mobile, de sorte que l'emploi de facteurs de composition différents n'aide pas à réduire le biais. Le biais de l'estimateur modifié est donné par

$$\begin{aligned} b(\hat{y}_{\mathrm{mod},ab}) = WN_{ab}(&\bar{Y}_{ml}(\lambda_1 r_{l1} r_l^{-1} + (1-\lambda_1) r_{c1} r_c^{-1} - 1) \\ &- \bar{Y}_{mc}(\lambda_2 r_{l1} r_l^{-1} + (1-\lambda_2) r_{c1} r_c^{-1} - 1)), \end{aligned} \quad (9)$$

où nous formulons les mêmes hypothèses que celles utilisées plus haut pour approximer le biais de $\hat{y}_{ps,ab}$.

Une autre raison d'étudier un estimateur pour le domaine de chevauchement tel que (8) tient au fait qu'il convient pour les plans de sondage comportant l'élimination par présélection des adultes se servant principalement d'un téléphone fixe provenant de la base de numéros de téléphone mobile. Cette approche a été examinée parce que le nombre de répondants sélectionnés dans la base de numéros de téléphone mobile qui sont classés comme utilisant principalement un téléphone fixe peut être faible et que l'hypothèse selon laquelle $E\hat{y}_{ab}^B(ml) = \bar{Y}_{ml}$ risque de ne pas être vérifiée, ce qui pourrait produire un biais.

En posant que $\lambda_1 = 1$, (8) se réduit à

$$\begin{aligned}\hat{y}_{\mathrm{mod}\,\lambda=1,ab} &= g^A \hat{y}_{ab}^A(ml) + \lambda_2 g^A \hat{y}_{ab}^A(mc) \\ &\quad + (1-\lambda_2) g^B \hat{y}_{ab}^B(mc). \end{aligned} \quad (10)$$

Dans ce plan de sondage, seul l'échantillon de numéros de téléphone fixe est utilisé pour estimer les totaux pour les personnes utilisant un téléphone fixe seulement et celles utilisant principalement un téléphone fixe. Les échantillons provenant des deux bases de sondage sont utilisés pour

estimer les totaux pour les répondants se servant principalement d'un téléphone mobile. Si nous supposons que $E\bar{y}_{ab}^{A}(ml) = \bar{Y}_{ml}$ et $E\bar{y}_{ab}^{A}(mc) = E\bar{y}_{ab}^{B}(mc) = \bar{Y}_{mc}$, il n'est plus nécessaire que $E\bar{y}_{ab}^{B}(ml) = \bar{Y}_{ml}$ pour que (10) soit sans biais. Comme auparavant, poser que $\lambda_2 = r_l(r_c - r_{c1})(r_c r_{l1} - r_l r_{c1})^{-1}$ élimine le biais dans l'estimation pour les répondants utilisant principalement un téléphone mobile.

5. Discussion

Le présent examen des erreurs de non-réponse et de mesure dans les enquêtes téléphoniques à base de sondage double donne à penser que les effets de ces erreurs peuvent être très importants. Nous en venons donc à penser que la recherche sur les erreurs non dues à l'échantillonnage en vue de réduire les biais peut être plus importante que celle donnant lieu à des réductions progressives de l'erreur d'échantillonnage.

Les présents travaux de recherche révèlent aussi les lacunes de notre savoir sur les erreurs non dues à l'échantillonnage dans ces enquêtes. La direction et la grandeur des effets de l'erreur de mesure sont particulièrement peu claires. Les incohérences de certains résultats de la CHIS 2007 et des enquêtes Pew pourraient fort bien être dues à des erreurs de mesure associées aux approches différentes de collecte des données suivies dans ces enquêtes ou à des interactions dues aux procédures. Un examen approfondi des sources d'erreur dans les enquêtes téléphoniques à base de sondage double est essentiel si l'on veut améliorer la qualité de ces enquêtes et, à notre avis, des expériences destinées à évaluer les effets des erreurs de mesure seraient particulièrement utiles.

Nous avons constaté que la CHIS 2007 et les enquêtes Pew donnaient lieu à une surreprésentation systématique des utilisateurs d'un téléphone mobile seulement et d'un téléphone mobile principalement dans les échantillons provenant de la base de numéros de téléphone mobile, ainsi qu'une légère surreprésentation des utilisateurs d'un téléphone fixe seulement et d'un téléphone fixe principalement provenant de la base de numéros de téléphone fixe. Cependant, le degré de surreprésentation des domaines variait selon l'enquête. Dans la CHIS, la surreprésentation aurait pu produire des biais importants dans les estimations si un plan de sondage avec chevauchement et un estimateur moyen simple avaient été adoptés. Une approche de plan de sondage avec présélection a été utilisée pour la CHIS en vue de réduire le biais éventuel, stratégie qui semble avoir été en grande partie couronnée de succès. Dans les enquêtes Pew, la représentation variait moins selon la base de sondage et le risque de biais était plus faible. Dans ces conditions, l'approche avec chevauchement pourrait produire une erreur quadratique moyenne plus petite que l'approche avec présélection.

Étant donné la possibilité d'un biais dans les enquêtes téléphoniques à base de sondage double pour lesquelles les profils de réponse ressemblent à ceux de la CHIS 2007, nous avons examiné des méthodes d'échantillonnage et d'estimation susceptibles d'être appliquées pour traiter ces biais. Nous avons constaté que les approches avec présélection peuvent être compétitives, et même préférables, pour les enquêtes téléphoniques à base de sondage double quand le biais dû à une non-réponse différentielle ou à l'erreur de mesure est grand. Si le biais n'est pas négligeable, cette constatation est même vraie pour de petites tailles d'échantillon. Toutefois, ces résultats dépendent du choix du facteur de composition et il conviendrait donc de réexaminer la pratique courante consistant à choisir $\lambda = 0{,}5$. Une autre option consiste à choisir le facteur de composition de manière à éliminer le biais de l'estimateur moyen. Dans de nombreux cas, cette approche peut non seulement éliminer le biais, mais aussi être plus efficace.

Nous avons examiné trois estimateurs en vue de traiter le biais dû à la différence de non-réponse dans le domaine de chevauchement. Le premier est $\hat{y}_{ps}$, qui utilise la situation concernant le type de téléphone comme variable pour les totaux de contrôle de domaine. Cet estimateur élimine le biais dû à la différence de non-réponse quand λ_0 est utilisé comme facteur de composition, car celui-ci emploie indirectement l'information sur les totaux des domaines d'utilisateurs d'un téléphone fixe principalement et d'un téléphone mobile principalement dans le calcul des taux de réponse selon le domaine et la base de sondage. Un deuxième estimateur, $\hat{y}_{sep}$, élimine cette source de biais plus directement par poststratification sur les totaux de contrôle pour la situation concernant le type de téléphone et pour l'usage. Cet estimateur permet aussi d'utiliser des facteurs de composition différents dans le domaine de chevauchement afin de réduire la variance des estimations. Le troisième estimateur qui pourrait être utilisé pour réduire le biais est $\hat{y}_{\mathrm{mod}}$, mais cet estimateur est plus pertinent dans le cas d'un plan de sondage qui prévoit l'interview des répondants utilisant un téléphone mobile seulement et ceux utilisant un téléphone mobile principalement sélectionnés dans la base de numéros de téléphone mobile, ainsi que tous les répondants provenant de l'échantillon tiré de la base de numéros de téléphone fixe. Ce plan de sondage et cet estimateur avec présélection modifiés pourraient être particulièrement intéressants quant on craint que la moyenne pour les répondants utilisant principalement un téléphone fixe compris dans l'échantillon provenant de la base de numéros de téléphone mobile soit sujette à un biais de non-réponse. Tous ces estimateurs pourraient également être calés par

ratissage sur des totaux de contrôle démographiques supplémentaires après avoir combiné les deux échantillons.

Étant donné l'état actuel de nos connaissances, nous pensons qu'il existe d'importants avantages à adopter le plan de sondage avec chevauchement complet et $\hat{y}_{ps}$ avec λ_0 choisi en se basant sur d'autres enquêtes similaires. Il convient de souligner que, même si la CHIS et les enquêtes Pew possédaient des profils de réponse très différents, choisir une valeur de $\lambda_0 = 0{,}75$ aurait réduit considérablement le biais pour les deux enquêtes. L'un des avantages de cet estimateur par rapport à $\hat{y}_{sep}$ est, de manière générale, que $\hat{y}_{ps}$ n'est pas poststratifié en fonction des totaux de domaine d'usage. Nous avons le sentiment que les totaux de domaine d'usage estimés d'après une enquête avec interview sur place (NHIS) peuvent être sujets à des erreurs fort différentes de celles affectant les estimations d'après des enquêtes téléphoniques. Ces différences pourraient donner lieu à des estimations basées sur les enquêtes téléphoniques qui sont biaisées et dont la variance est sous-estimée. Dans le cas des enquêtes au niveau de l'État et au niveau local, pour lesquelles même les totaux pour la situation concernant le type de téléphone utilisé ne sont pas bien connus, les totaux de contrôle pour l'usage du téléphone sont vraisemblablement fort douteux.

Un plan de sondage avec présélection et l'utilisation de $\hat{y}_{scr}$ comme estimateur a l'avantage de ne nécessiter des totaux de contrôle que pour l'ensemble de la population et pour la composante des utilisateurs d'un téléphone mobile seulement comparables à ceux estimés d'après les données de la NHIS. Un inconvénient est dû au fait que, contrairement aux estimateurs avec chevauchement, il n'existe aucun paramètre de composition pouvant être utilisé pour réduire directement le biais. Le plan avec présélection plus élaboré qui consiste à interviewer les répondants utilisant un téléphone mobile seulement et ceux utilisant un téléphone mobile principalement provenant de la base de numéros de téléphone mobile et à utiliser $\hat{y}_{\mathrm{mod}}$ est valable, mais aucune étude n'a eu pour objectif d'examiner les conditions qui favoriseraient le choix de cet estimateur.

Une analyse plus complexe des effets de l'erreur non due à l'échantillonnage comprendrait d'autres facteurs, tels que l'effet des différences de taux de réponse selon la base de sondage. Par exemple, nous avons noté que les échantillons provenant de la base de numéros de téléphone mobile produit un plus grand nombre que prévu de ménages utilisant uniquement un téléphone mobile. Ces différences de taux de réponse peuvent être prises en compte dans la répartition de l'échantillon, mais nous ne l'avons pas fait ici. Notre examen de la question révèle que cette prise en compte aboutit à l'affectation d'une plus grande part de l'échantillon à la base de sondage de numéros de téléphone fixe, à l'augmentation de la valeur du facteur de composition et à un accroissement de l'efficacité des plans avec présélection par rapport aux plans avec chevauchement. Le plan de sondage et l'estimateur avec présélection restent sujets au biais susmentionné.

Bien que la présente étude soit axée sur les erreurs non dues à l'échantillonnage dans les enquêtes téléphoniques à base de sondage double, nous soupçonnons que des problèmes semblables se posent dans le cas de nombreuses autres enquêtes à base de sondage double, mais qu'ils ne sont peut-être pas reconnus. Lohr (2009) mentionne des erreurs non dues à l'échantillonnage dans des enquêtes à base de sondage double générales et suggère de comparer les estimations pour le domaine de chevauchement d'après chaque base de sondage comme simple test diagnostique. Nous pensons qu'il s'agit d'un excellent moyen de commencer l'étude des problèmes associés au chevauchement.

Comme nous l'avons mentionné plus haut, le traitement du chevauchement est un problème important dans les enquêtes à base de sondage double, parce que l'erreur non due à l'échantillonnage peut être liée à la base de sondage. Notre étude montre que les erreurs de non-réponse et de mesure sont reliées à la base de sondage dans les enquêtes téléphoniques à base de sondage double. Il est fort probable que les enquêtes téléphoniques à base de sondage double utilisant des modes différents pourraient être sujettes à des effets analogues. Par exemple, considérons le cas d'une enquête-ménage à base de sondage double conçue pour étudier les membres d'une population rare. Supposons que l'on utilise une liste incomplète des membres de la population avec les numéros de téléphone pour le groupe rare comme base de sondage *A* et un échantillon probabiliste aréolaire de ménages comme base de sondage *B*. On pourrait s'attendre que les taux de réponse dans le domaine de chevauchement diffèrent selon la base de sondage et que la différence soit associée aux caractéristiques des répondants, ce qui causera des biais. Même dans le domaine de chevauchement, il pourrait exister des différences telles que celles concernant la durée de l'adhésion de la personne à l'organisation utilisée pour créer la base de sondage *A*, ce qui pourrait être relié à des caractéristiques comme l'âge. Ce type de situation pourrait présenter un parallèle avec certains problèmes cernés dans le domaine de chevauchement dans le cas des enquêtes téléphoniques. Des erreurs de mesure différentes selon le mode de collecte sont également possibles.

Étant donné la possibilité d'un biais dans les enquêtes à base de sondage double, l'une des constatations importantes de notre étude est que le facteur de composition, λ, influence le biais et a un effet sur la variance. Même si le choix de la valeur de λ n'a habituellement qu'un léger effet sur la variance si cette valeur est proche de la valeur optimale, le biais peut être plus sensible à ce choix. Donc,

dans les enquêtes à base de sondage double, il est important de prendre en considération la façon dont le choix de λ affecte le biais et l'erreur quadratique moyenne des estimations. Les autres méthodes d'échantillonnage et d'estimation discutées dans le présent article peuvent également s'appliquer à d'autres enquêtes à base de sondage double. L'utilité de ces méthodes dépend de la compréhension de la nature des erreurs non dues à l'échantillonnage ainsi que de la disponibilité des données auxiliaires qui pourraient être utilisées dans le calage.

Remerciements

Nous remercions Scott Keeter, Stephen Blumberg et Julian Luke d'avoir fourni les données sur lesquelles s'appuie le présent article. Nous remercions également de nombreuses personnes de leurs commentaires constructifs concernant des versions antérieures du manuscrit, y compris Sherm Edwards, Ralph DiGaetano, David Grant, David Hubble, Paul Lavrakas, Graham Kalton, Scott Keeter et Courtney Kennedy.

Annexe

Questions sur l'usage du téléphone

National Health Interview Survey

[TRADUCTION]

N1. *Existe-t-il dans votre foyer au moins un téléphone qui fonctionne à l'heure actuelle et qui n'est pas un téléphone mobile* ?

N2. *Est-ce qu'un membre de votre famille possède un téléphone mobile qui fonctionne* ?

N3. *Combien de téléphones mobiles qui fonctionnent les membres de votre famille possèdent-ils* ?

[Si la réponse est « oui » à N1 ainsi que N2, posez N4.]

N4. *De tous les appels téléphoniques que reçoit votre famille, est-ce que...*

Tous ou presque tous sont reçus sur des téléphones mobiles ?

Certains sont reçus sur des téléphones mobiles et certains sur des téléphones ordinaires ?

Très peu sont reçus sur des téléphones mobiles ou aucun n'est reçu sur des téléphones mobiles ?

California Health Interview Survey – Téléphone mobile

[TRADUCTION]

CC1. *Ce téléphone mobile est-il votre seul téléphone ou possédez-vous aussi un téléphone ordinaire à la maison* ?

[Si le téléphone est un téléphone mobile et que le répondant possède un téléphone ordinaire, posez la question CC2.]

CC2. *De tous les appels téléphoniques que vous recevez, est-ce que...*

Tous ou presque tous sont reçus sur des téléphones mobiles ?

Certains sont reçus sur des téléphones mobiles et certains sur des téléphones ordinaires ?

Très peu sont reçus sur des téléphones mobiles ou aucun n'est reçu sur des téléphones mobiles ?

[Si le répondant répond environ la moitié, inscrire sa réponse.]

California Health Interview Survey – Téléphone fixe

[TRADUCTION]

CL1. *Avez-vous un téléphone mobile qui fonctionne* ?

[Si la réponse est oui ou que le répondant partage un téléphone mobile, posez CL2.]

CL2. *De tous les appels téléphoniques que vous recevez, est-ce que...*

Tous ou presque tous sont reçus sur des téléphones mobiles ?

Certains sont reçus sur des téléphones mobiles et certains sur des téléphones ordinaires ?

Très peu sont reçus sur des téléphones mobiles ou aucun n'est reçu sur des téléphones mobiles ?

[Si le répondant répond « environ la moitié », inscrire sa réponse.]

Pew Research Center for the People & The Press – Téléphone mobile

[TRADUCTION]

PC1. *Maintenant, si vous pensez à votre utilisation du téléphone... existe-t-il au moins un téléphone À L'INTÉRIEUR de votre foyer qui fonctionne à l'heure actuelle et qui n'est pas un téléphone mobile* ?

[Si la réponse est oui, posez PC2.]

PC2. *De tous les appels téléphoniques que vous recevez, est-ce que ...*

[Alternez les options en gardant CERTAINS au milieu.]

Tous ou presque tous sont reçus sur un téléphone mobile ?

Certains sont reçus sur un téléphone mobile et certains, sur un téléphone ordinaire à la maison ?

Tous ou presque tous sont reçus sur un téléphone ordinaire à la maison ?

Pew Research Center for the People & The Press – Téléphone fixe

[TRADUCTION]

PL1. *Maintenant, en pensant à votre utilisation du téléphone... Possédez-vous un téléphone mobile qui fonctionne* ?

[Si la réponse est oui, posez PL2.]

PL2. *De tous les appels téléphoniques que vous recevez, est-ce que ...*

[Alternez les options en gardant CERTAINS au milieu]

Tous ou presque tous sont reçus sur un téléphone mobile ?

Certains sont reçus sur un téléphone mobile et certains, sur un téléphone ordinaire à la maison ?

Tous ou presque tous sont reçus sur un téléphone ordinaire à la maison ?

Bibliographie

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.

Blumberg, S.J., et Luke, J.V. (2009). Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, juillet-décembre 2008. National Center for Health Statistics. Disponible au http://www.cdc.gov/nchs/nhis.htm.

Blumberg, S.J., Luke, J.V., Davidson, G., Davern, M.E., Yu, T. et Soderberg, K. (2009). Wireless substitution: State-level estimates from the National Health Interview Survey, janvier-décembre 2007. Hyattsville, MD : National Center for Health Statistics. *National Health Statistics Reports*, 14.

Brick, J.M., Brick, P.D., Dipko, S., Presser, S., Tucker, C. et Yuan, Y. (2007). Cell phone survey feasibility in the U.S.: Sampling and calling cell numbers versus landline numbers. *Public Opinion Quarterly*, 71, 29-33.

Brick, J.M., Dipko, S., Presser, S., Tucker, C. et Yuan, Y. (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *Public Opinion Quarterly*, 70, 780-793.

Brick, J.M., Edwards, W.S. et Lee, S. (2007). Sampling telephone numbers and adults, interview length, and weighting in the California Health Interview Survey cell phone pilot study. *Public Opinion Quarterly*, 71, 793-813.

California Health Interview Survey (2009). CHIS 2007 Methodology Series: Rapport 4 – Response Rates. Los Angeles, CA : UCLA Center for Health Policy Research. Disponible au www.chis.ucla.edu/pdf/CHIS2007_method4.pdf.

Edwards, W.S., Brick, J.M. et Grant, D. (2008). Relative Costs of a Multi-frame, Multi-mode Enhancement to an RDD Survey. Présenté à Annual Conference of the American Association for Public Opinion Research, la Nouvelle-Orléans, LA.

Fleeman, A. (2007). Survey Research Using Cell Phone Sample: Important Operational and Methodological Considerations. Présenté à Annual Conference of the American Association for Public Opinion Research, Anaheim, CA.

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

Hartley, H.O. (1962). Multiple Frame Surveys. *ASA Proceedings of the Social Statistics Section*, 203-206.

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

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

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

Kennedy, C. (2007). Evaluating the effects of screening for telephone service in dual frame RDD Surveys. *Public Opinion Quarterly*, 71, 750-771.

Keeter, S., Dimock, M. et Christian, L. (2008). Calling Cell Phones in '08 Pre-election Polls. News Release from The Pew Research Center for the People & the Press. Disponible au www.pewresearch.org/pubs/1061/cell-phones-election-polling.

Kuusela, V., Callegaro, M. et Vehovar, V. (2008) The influence of mobile telephones on telephone surveys. Dans *Advances in Telephone Survey Methodology*, (Éds., J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link et R.L. Sangster), New York : John Wiley & Sons, Inc., Chapitre 4, 87-112.

Lohr, S. (2009). Multiple frame surveys. Dans *Handbook of Statistics*: *Sample Surveys_Design, Methods and Applications*, (Éd., D. Pfeffermann). Elsevier, Amsterdam, Chapitre 4, Vol. 29A.

Lohr, S., et Rao, J.N.K. (2000). Inference in dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

Lohr, S., et Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.

Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.

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

Steeh, C. (2004). A New Era for Telephone Surveys. Présenté à Annual Conference of the American Association for Public Opinion Research, Phoenix, AZ.

Tucker, C., Brick, J.M. et Meekins, B. (2007). Household telephone service and usage patterns in the U.S. in 2004: Implications for telephone samples. *Public Opinion Quarterly*, 71, 3-22.

Vicente, P., et Reis, E. (2009). The mobile-only population in Portugal and its impact in a dual frame telephone survey. *Survey Research Methods*, 3, 105-111.

Techniques d'enquête, juin 2011
Vol. 37, N° 1, pp. 17-30
Statistique Canada, N° 12-001-X au catalogue

Estimation du maximum de vraisemblance pour les tableaux de contingence et la régression logistique en présence de données incorrectement appariées

James O. Chipperfield, Glenys R. Bishop et Paul Campbell [1]

Résumé

L'appariement des données consiste à jumeler des enregistrements issus de deux fichiers ou plus que l'on pense appartenir à une même unité (par exemple une personne ou une entreprise). Il s'agit d'un moyen très courant de renforcer la dimension temporelle ou des aspects tels que la portée ou la profondeur des détails. Souvent, le processus d'appariement des données n'est pas exempt d'erreur et peut aboutir à la formation d'une paire d'enregistrements qui n'appartiennent pas à la même unité. Alors que le nombre d'applications d'appariement d'enregistrements croît exponentiellement, peu de travaux ont porté sur la qualité des analyses effectuées en se servant des fichiers de données ainsi appariées. Traiter naïvement ces fichiers comme s'ils ne contenaient pas d'erreurs mène, en général, à des estimations biaisées. Le présent article décrit l'élaboration d'un estimateur du maximum de vraisemblance pour les tableaux de contingence et la régression logistique en présence de données incorrectement appariées. Simple, cette méthode d'estimation est appliquée en utilisant l'algorithme EM bien connu. Dans le contexte qui nous occupe, l'appariement probabiliste des données est une méthode reconnue. Le présent article démontre l'efficacité des estimateurs proposés au moyen d'une étude empirique s'appuyant sur cet appariement probabiliste.

Mots clés : Appariement des données , appariement probabiliste ; maximum de vraisemblance ; tableaux de contingence ; régression logistique.

1. Introduction

L'appariement des données, également appelé appariement ou couplage d'enregistrements, est la tâche consistant à jumeler des enregistrements que l'on pense appartenir à une même unité (par exemple une personne ou une entreprise) et qui sont tirés de deux fichiers ou plus. L'appariement des données est une technique indiquée pour jumeler des ensembles de données en vue de renforcer la dimension temporelle, ou des aspects tels que la portée ou la profondeur des détails. Dans des conditions idéales, l'appariement serait parfait, autrement dit seuls les enregistrements appartenant à la même unité seraient appariés et tous les appariements possibles seraient faits. Malheureusement, très souvent, il n'en est pas ainsi, surtout si l'on se sert pour apparier les enregistrements de champs pouvant contenir des valeurs incorrectes, des valeurs manquantes ou des valeurs légitimement différentes pour une unité particulière.

On recourt souvent à l'appariement probabiliste quand les fichiers contiennent un ensemble commun de variables ou de champs qui fournissent des renseignements d'identification partiels, mais ne constituent pas un identificateur d'unité unique. Dans l'appariement probabiliste (Fellegi et Sunter 1969), un score est attribué à chacun des appariements possibles en se basant sur la probabilité que les enregistrements appartiennent à la même unité. Ce score est calculé en comparant les valeurs des variables d'appariement qui sont communes aux deux fichiers. Un appariement est alors déclaré si le score d'appariement est supérieur à un seuil donné. Un algorithme d'optimisation peut être utilisé pour s'assurer que chaque enregistrement d'un fichier ne soit pas apparié à plus d'un enregistrement d'un autre fichier. Les méthodes probabilistes d'appariement de fichiers sont bien établies aujourd'hui (voir Herzog, Scheuren et Winkler 2007, Winkler 2001 et Winkler 2005), et il existe toute une gamme de progiciels pour les mettre en œuvre.

Cette situation découle de l'importance continue de l'appariement dans divers domaines, particulièrement ceux touchant aux politiques en matière de santé et aux politiques sociales. Les exemples récents d'appariements probabilistes de données effectués par l'Australian Bureau of Statistics (ABS) comprennent l'appariement d'enregistrements provenant du Recensement de la population et du logement de l'Australie de 2006 avec ceux d'un certain nombre d'ensembles de données, dont les enregistrements des décès survenus en Australie (Australian Bureau of Statistics 2008), le Census Dress Rehearsal (Répétition générale du recensement) de 2006 (Solon et Bishop 2009) et l'Australian Migrants Settlements Database (Wright, Bishop et Ayre 2009). Dans le secteur australien de la santé, les méthodes d'appariement probabiliste sont employées par la Western Australian Data Linkage Unit (Holman, Bass, Rouse et Hobbs 1999) et par le New South Wales Centre for Heath Record Linkage. Sur la scène internationale, les méthodes probabilistes sont utilisées par Statistique Canada (Fair 2004), le Census Bureau des États-Unis (voir Winkler

1. James O. Chipperfield, Australian Bureau of Statistics. Courriel : james.chipperfield@abs.gov.au ; Glenys R. Bishop, The Australian National University ; Paul Campbell, Australian Bureau of Statistics.

2001), le National Center for Health Statistics des États-Unis (National Center for Health Statistics 2009) et l'Office fédéral de la statistique de la Suisse dans le cadre de son étude longitudinale des personnes vivant en Suisse.

L'appariement des données rend possibles de nouveaux produits et analyses statistiques. En général, traiter naïvement le fichier de données appariées de manière probabiliste comme s'il s'agissait d'un appariement parfait produit des estimations biaisées. Lahiri et Larsen (2005), ainsi que Scheuren et Winkler (1993) ont proposé des méthodes permettant d'estimer sans biais les coefficients d'un modèle de régression linéaire sous appariement probabiliste d'enregistrements. Plus récemment, Chambers et coll. (2009) et Chambers (2008) ont élargi la portée de ces travaux à un vaste ensemble de modèles en utilisant des équations d'estimation généralisées et, dans le cas de l'appariement de deux fichiers, en permettant que l'un des fichiers soit un sous-ensemble de l'autre.

Le présent article décrit l'élaboration d'une approche d'estimation du maximum de vraisemblance (MV) pour analyser les données appariées de manière probabiliste. La technique d'estimation est simple et mise en œuvre en se servant de l'algorithme EM bien connu. L'approche consiste à remplacer les statistiques qui seraient observées dans le cas de données parfaitement appariées par leur espérance conditionnellement aux données appariées. En supposant que l'espérance est spécifiée correctement, cette approche permet d'éviter les deux limites des travaux antérieurs décrites ci-après.

Premièrement, alors que les méthodes antérieures reposaient sur un appariement exécuté en un seul passage, l'appariement probabiliste comporte habituellement des passages multiples. Dans ce cas, seuls les enregistrements non appariés au moment du premier passage peuvent faire l'objet d'un appariement durant le deuxième passage, puis seuls les enregistrements non appariés au cours des deux premiers passages peuvent être appariés lors du troisième, et ainsi de suite. Chaque passage est conçu en vue d'apparier les enregistrements présentant un ensemble commun particulier de caractéristiques. Par exemple, le premier passage peut être conçu de manière à apparier les enregistrements appartenant aux personnes qui n'ont pas changé d'adresse entre les dates de référence des deux fichiers. Le deuxième passage peut être conçu pour tenir compte des changements d'adresse. Un exemple de ce genre d'approche est donné au tableau 1 à la section 5.

Deuxièmement, les méthodes antérieures reposaient sur l'hypothèse que les deux fichiers contenaient des enregistrements appartenant exactement aux mêmes unités ou que l'ensemble d'unités figurant dans un fichier était un sous-ensemble de celles présentes dans l'autre fichier. L'approche que nous proposons ne requiert pas que l'un des fichiers à apparier soit un sous-ensemble de l'autre fichier. Fréquente en pratique, cette situation s'est présentée dans tous les exemples susmentionnés d'appariements à l'ABS. Il convient aussi de mentionner que les fichiers à apparier ne doivent pas nécessairement être reliés par un mécanisme d'échantillonnage, comme dans le cas où le fichier le plus petit est un sous-échantillon aléatoire des unités comprises dans le plus grand fichier. L'élimination de cette restriction signifie que les deux fichiers peuvent être des ensembles de données administratives.

Considérons l'appariement de deux fichiers désignés par X et Y. Le fichier Y contient la variable y de la population d'individus U_y comprenant n_y enregistrements. Le fichier X contient un vecteur, $\mathbf{x}$, de variables sur la population d'individus U_x comprenant n_x enregistrements. L'inférence a pour cible la population de n_{xy} individus, désignée par $U_{xy} = U_x \cap U_y$, qui sont communs au fichier X et au fichier Y. Les fichiers X et Y contiennent également un vecteur de champs, désigné par $\mathbf{z}$, qui sont utilisés pour apparier les fichiers en utilisant un algorithme d'appariement probabiliste. Naturellement, puisque nous considérons un appariement probabiliste ici, la variable $\mathbf{z}$ ne constitue pas un identificateur d'unité unique.

L'appariement des fichiers X et Y permet d'analyser la distribution conjointe de $\mathbf{x}$ et de y. Deux sources d'erreur peuvent avoir une incidence sur l'analyse de la distribution conjointe en se servant du fichier de données appariées. Ces erreurs correspondent aux *appariements incorrects* et aux *enregistrements non appariés*.

Un appariement est correct si les deux enregistrements appariés appartiennent à une même personne. Un appariement est incorrect si les deux enregistrements appariés n'appartiennent pas à la même personne. Les appariements incorrects peuvent accroître ou réduire artificiellement la corrélation entre $\mathbf{x}$ et y. Un exemple du second cas est l'appariement aléatoire, dans lequel les enregistrements du fichier X sont appariés aléatoirement aux enregistrements du fichier Y.

Le i^{e} enregistrement du fichier X est défini comme un *enregistrement non apparié* si $i \in U_{xy}$ *et* que l'enregistrement i n'a pas été apparié à un enregistrement du fichier Y. Autrement dit, un enregistrement non apparié est un enregistrement du fichier X qui pourrait être apparié correctement, mais n'a pas été apparié du tout (tout au long de l'exposé, nous adoptons la convention de définir les enregistrements non appariés en fonction du fichier X, mais la définition pourrait également être formulée en fonction des enregistrements du fichier Y). Il se pourrait que l'on ne puisse pas toujours apparier un enregistrement particulier du fichier X en ayant la certitude que l'appariement est correct. Cette situation peut se présenter s'il manque dans un enregistrement des champs qui sont utiles pour établir

l'appariement correct. De manière plus générale, des cas d'enregistrements non appariés peuvent avoir lieu quand certaines sous-populations sont relativement difficiles à apparier. Par exemple, des champs tels que l'état matrimonial, la qualification, le domaine d'études et le plus haut niveau de scolarité ne seraient en général pas aussi puissants lorsque l'appariement concerne des enfants que lorsqu'il s'agit d'adultes ayant atteint la maturité. Dans cette situation, l'apparieur des données doit décider s'il doit ou non apparier ce genre d'enregistrements. Nous définissons l'ensemble d'enregistrements appariés par U_I de taille n^* de sorte que $n^* \leq n_x$ et $n^* \leq n_y$.

Le problème que pose l'analyse quand des enregistrements sont non appariés présente des points communs manifestes avec le problème de la non-réponse totale d'une unité. Dans les deux cas, un seul sous-ensemble d'enregistrements légitimes est disponible pour l'analyse. Le mécanisme de non-réponse dans les enquêtes par sondage dépend, en réalité, d'un ensemble inconnu de variables. En revanche, ici, nous avons le léger avantage de savoir que la probabilité qu'un enregistrement demeure non apparié ne peut être qu'une fonction de **z**. Le problème de la non-réponse est souvent résolu par pondération ou au moyen d'un certain argument de conditionnement. Dans le présent article, nous envisageons les deux approches pour résoudre le problème des enregistrements non appariés.

Il existe un compromis naturel entre le nombre d'enregistrements non appariés et le nombre d'appariements incorrects (et par conséquent le biais qu'ils introduisent). Considérons le cas où le fichier X est un sous-échantillon du fichier Y de sorte que $U_{xy} = U_x$. L'appariement de tous les enregistrements du fichier X ne donnera lieu, par définition, à aucun enregistrement non apparié, mais maximisera le nombre d'appariements incorrects. Si nous décidons plutôt de ne former que des appariements dont nous sommes convaincus qu'ils sont corrects, le nombre d'appariements incorrects diminuera, mais le nombre d'enregistrements non appariés augmentera. En pratique, trouver l'équilibre optimal entre les biais dus aux enregistrements non appariés et aux appariements incorrects dépend de l'analyse qui doit être effectuée, de la méthode d'appariement et de leur interaction. Pour une discussion pratique approfondie de cette question, voir Bishop (2009).

Il convient de mentionner que le problème de l'inférence en présence d'appariements incorrects d'enregistrements est semblable au problème de l'inférence en présence de classifications incorrectes de la variable dépendante, qui est une forme d'erreur de mesure (voir Fuller 1987). Dans ce dernier cas, des hypothèses d'identification font la distinction entre le mécanisme d'erreur de classification et les mécanismes du modèle, et sont nécessaires puisqu'habituellement, on ne dispose d'aucune mesure exempte d'erreur. Par exemple, Hausman et coll. (1998) considèrent le cas d'une erreur de classification dans la variable dépendante d'un modèle de régression logistique. Leur hypothèse d'identification est que la valeur de la variable dépendante éventuellement classée incorrectement est une fonction particulière des variables explicatives du modèle. La méthode que nous proposons ne nécessite pas les hypothèses d'identification fortes que requièrent les problèmes d'erreur de mesure, essentiellement parce qu'une mesure sans erreur peut être obtenue au moyen d'un échantillon pour examen manuel qui permet de déterminer les appariements corrects. Les hypothèses que nous faisons dans le présent article sont énoncées à la section 3.

À la section 2, nous résumons l'approche du MV pour les tableaux de contingence et l'analyse de régression sous appariement parfait. À la section 3, nous considérons l'approche du MV en présence d'appariements incorrects. À la section 4, nous examinons l'approche du MV en présence d'appariements incorrects et d'enregistrements non appariés. À la section 5, au moyen d'une étude empirique, nous démontrons l'efficacité de bon nombre des estimateurs proposés. Enfin à la section 6, nous résumons les résultats.

2. Appariement parfait

En vue de présenter la notation, nous discutons dans cette section du cas où l'appariement est parfait. Par conséquent, l'approche d'estimations est classique, puisque, de toute évidence, aucun ajustement spécial n'est nécessaire pour tenir compte des appariements incorrects. À la section 2.1, nous discutons de l'estimation des probabilités dans les cellules d'un tableau de contingence et à la section 2.2, de l'estimation des coefficients de régression d'une régression logistique.

2.1 Tableaux de contingence

En ce qui concerne la notation, il est commode, lorsque l'on envisage l'analyse d'un tableau de contingence, de transformer $\mathbf{x}_i$ en une variable catégorique unique x de sorte que $x = 1, 2, ..., g, ..., G$. Soit y une variable catégorique du fichier Y, où $y = 1, ..., c, ..., C$.

Considérons la factorisation qui suit de la distribution de x et de y

$$p(y, x) = p_1(y \mid x; \boldsymbol{\Pi})\, p_2(x),$$

où $\boldsymbol{\Pi} = (\boldsymbol{\pi}_1', ..., \boldsymbol{\pi}_g', ..., \boldsymbol{\pi}_G')'$, $\boldsymbol{\pi}_g = (\pi_{1|g}, ..., \pi_{c|g}, ..., \pi_{C|g})'$, $\pi_{c|g}$ est la probabilité que $y = c$ sachant $x = g$. Nous supposons que, pour chaque valeur de x, il existe C valeurs possibles de y, ce qui implique que la dimension de $\boldsymbol{\Pi}$ est CG.

Considérons maintenant l'estimation du maximum de vraisemblance du paramètre $\boldsymbol{\Pi}$, caractérisant p_1, sous

appariement parfait. Un appariement parfait signifie que chacun des enregistrements du fichier X est correctement apparié à l'enregistrement correspondant du fichier Y (c'est-à-dire qu'il n'y a pas d'appariement incorrect ni d'enregistrement non apparié). Sous appariement parfait, $n_{xy} = n_x$ et l'ensemble d'enregistrements appariés est désigné par $\mathbf{d} = \{(y_i,\ x_i) : i = 1, ..., n_{xy}\}$. Sous appariement parfait, la fonction de score pour $\boldsymbol{\pi}_x = (\pi_{1|x}, ..., \pi_{c|x}, ..., \pi_{C|x})'$ caractérisée par la distribution multinomiale est

$$\text{Score}(\boldsymbol{\pi}_x ; \mathbf{d}) =$$

$$(\text{Score}(\pi_{1|x} ; \mathbf{d}),..., \text{Score}(\pi_{c|x} ; \mathbf{d}),..., \text{Score}(\pi_{C-1|x} ; \mathbf{d}))' \quad (1)$$

où

$$\text{Score}(\pi_{c|x} ; d) = \Sigma_i (w_{ic|x}\pi_{ic|x}^{-1} - w_{iC|x}\pi_{iC|x}^{-1})$$

$$= n_{c|x}\pi_{c|x}^{-1} - n_{C|x}\pi_{C|x}^{-1},$$

pour $c = 1, ..., C-1$, où $n_{c|x} = \Sigma_i w_{ic|x}$, $w_{ic|x} = 1$ si $y_i = c$ et $x_i = x$ et $w_{ic|x} = 0$ autrement, et la catégorie correspondant à $y = C$ est choisie arbitrairement comme catégorie de référence. La résolution de $\text{Score}(\boldsymbol{\pi}_x ; \mathbf{d}) = \mathbf{0}_{C-1}$ pour trouver $\boldsymbol{\pi}_x$, où $\mathbf{0}_{C-1}$ est un vecteur colonne de zéros de dimension $C-1$, donne l'estimateur du maximum de vraisemblance (MV)

$$\hat{\pi}_{c|x} = n_{c|x} / n_x, \quad (2)$$

où

$$n_x = \Sigma_c \Sigma_i w_{ic|x}$$

et

$$\hat{\pi}_{C|x} = 1 - \sum_{c=1}^{C-1} \hat{\pi}_{c|x}.$$

2.2 Régression logistique

Considérons le modèle de régression logistique

$$E(y_i) = \upsilon_i \quad (3)$$

$$\upsilon_i = 1 / [1 + \exp(\boldsymbol{\beta}'\mathbf{x}_i)]. \quad (4)$$

Pour (4), les K éléments de $\mathbf{x}_i$ sont des variables dichotomiques et y_i est maintenant une variable dichotomique disponible dans le fichier Y. Si nous définissons $\mathbf{x} = (\mathbf{x}_1,...,\mathbf{x}_i,...,\mathbf{x}_{n_{xy}})'$, $\mathbf{y} = (y_1,...y_i,...,y_{n_{xy}})'$ et $\boldsymbol{\upsilon} = (\upsilon_1,...\upsilon_i,..., \upsilon_{n_{xy}})'$, la matrice de score pour $\boldsymbol{\beta}$ basée sur des données parfaitement appariées, $\mathbf{d}$, est donnée par

$$\text{Score}(\boldsymbol{\beta} ; \mathbf{d}) = \mathbf{x}'(\mathbf{y} - \boldsymbol{\upsilon}). \quad (5)$$

La résolution de $\text{Score}(\boldsymbol{\beta} ; \mathbf{d}) = \mathbf{0}_K$ pour trouver $\boldsymbol{\beta}$ donne l'estimation du MV, $\hat{\boldsymbol{\beta}}$, que l'on peut obtenir en appliquant la méthode bien connue de Newton-Raphson.

3. Analyse en présence d'appariements incorrects

À la présente section, nous considérons la situation où le fichier d'enregistrements appariés contient des appariements incorrects, mais aucun enregistrement non apparié. Cette situation se produit quand chacun des enregistrements du fichier X est apparié à un enregistrement du fichier Y (d'où $n_x \le n_y$). Définissons le fichier d'enregistrements appariés par $\mathbf{d}^* = \{\mathbf{d}_i^* = (y_i^*, \mathbf{x}_i) : i = 1, ..., n_x\}$, où y_i^* est la valeur de y qui est *appariée* à l'enregistrement i du fichier X. Pour clarifier, y_i est la valeur vraie de y pour l'enregistrement i dans le fichier X, de sorte que $y_i^* = y_i$ si l'enregistrement i est apparié correctement.

L'estimateur donné par (2), associé à l'hypothèse que $y_i^* = y_i$ pour $i = 1, ..., n_x$, est naïf, puisqu'il traite le fichier d'enregistrements appariés de manière probabiliste comme s'ils étaient parfaitement appariés. En général, l'estimateur naïf contient un biais. À la présente section, nous dérivons les estimateurs du MV qui tiennent compte du fait que les données ont été appariées de manière probabiliste ou appariées imparfaitement d'une certaine façon.

Il est courant, en pratique, de tirer du fichier d'enregistrements appariés un sous-échantillon, désigné par s_c, puis de l'examiner manuellement. Durant l'examen manuel, un appariement, $\mathbf{d}_i$, est classé comme étant correct ou incorrect. Soit $\delta_i = 1$ si l'enregistrement i dans le fichier X est correctement apparié et $\delta_i = 0$ autrement.

La conception du sous-échantillon pour examen manuel est un problème important, surtout parce que l'examen manuel est souvent coûteux. Les utilisations possibles d'un échantillon pour examen manuel comprennent l'estimation de la proportion d'enregistrements correctement appariés et d'enregistrements non appariés, pour pouvoir décider quels enregistrements devraient être appariés et lesquels devraient rester non appariés, pour s'assurer que l'inférence en se servant de $\mathbf{d}^*$ est correcte (c'est-à-dire l'objectif du présent article) ou pour déterminer comment pourraient être améliorée la façon dont les enregistrements sont appariés. (Dans les applications de l'ABS susmentionnées, les échantillons pour examen manuel ont été conçus de manière à s'assurer que chaque appariement ait au moins une probabilité spécifiée d'être correct.) Si l'on veut s'assurer que l'inférence en se servant de $\mathbf{d}^*$ sera correcte, la sélection de l'échantillon pour examen manuel par échantillonnage aléatoire simple est une approche raisonnable. Un sous-échantillon pour examen manuel plus efficace pourrait peut-être être conçu, mais il n'existe aucun moyen évident de le faire, parce que les paramètres que nous devons estimer pour appliquer la méthode du MV décrite dans le présent article dépend de l'analyse particulière (par exemple choix de y et $\mathbf{x}$). Concevoir un échantillon pour examen

manuel convenant pour toutes les analyses possibles serait difficile.

Nous factorisons la distribution conjointe $p(y_i, \mathbf{x}_i, \delta_i)$ par

$$p(y_i \mid \mathbf{x}_i\ ; \boldsymbol{\theta})\, p(\mathbf{x}_i)\, p(\delta_i \mid \mathbf{x}_i), \qquad (6)$$

où $\boldsymbol{\theta} = \boldsymbol{\beta}$ dans le cas de la régression, et $\boldsymbol{\theta} = \boldsymbol{\Pi}$ dans le cas du tableau de contingence. La factorisation (6) signifie que les appariements sont incorrects au hasard (IAH), autrement dit, que les distributions $y_i \mid \mathbf{x}_i$ et $\delta_i \mid \mathbf{x}_i$ sont indépendantes. Sous cette hypothèse, il suffit de maximiser la vraisemblance associée au facteur $p(y_i \mid \mathbf{x}_i\ ; \boldsymbol{\theta})$. Tout au long de la présente section, nous faisons l'hypothèse (6). Il importe de souligner que (6) et le développement qui suit ne s'appuient sur aucune hypothèse nécessitant que le fichier X soit un sous-ensemble du fichier Y (par exemple quand les unités du fichier X sont un sous-échantillon des unités du fichier Y) ou que le processus d'appariement ne comporte qu'un seul passage. Nous supposons également que l'exactitude de l'appariement, δ_i, est indépendante d'un enregistrement à l'autre.

Comme nous l'avons mentionné dans l'introduction, un score est attribué à chaque enregistrement apparié en se basant sur la probabilité que les enregistrements appartiennent à la même unité. Désignons le score par r_i. Un examinateur a suggéré d'utiliser r_i pour paramétriser plus exactement la distribution de δ_i. Techniquement, cette suggestion nécessiterait le remplacement de $p(\delta_i \mid \mathbf{x}_i)$ par $p(\delta_i \mid \mathbf{x}_i, r_i)$ dans (6) et réduirait vraisemblablement la variabilité des estimateurs du MV discutés à la section 3. Il serait intéressant d'explorer cette piste dans le cadre de futurs travaux.

3.1 Tableaux de contingence

Soit $w^*_{ic|x} = 1$ si $y^*_i = c$ et $x_i = x$, et $w^*_{ic|x} = 0$ autrement. L'espérance de $w_{ic|x}$ sachant $\mathbf{d}^*_i$ est

$$\begin{aligned} E_{d|d^*}(w_{ic|x} \mid x_i = x, y^*_i = y^*) = \\ w^*_{ic|x} p_{xy^*} + (1 - p_{xy^*})\pi_{c|x} \quad &\text{si } i \notin s_c \\ = w^*_{ic|x} \quad &\text{si } i \in s_c \text{ et } \delta_i = 1 \\ = \pi_{c|x} \quad &\text{si } i \in s_c \text{ et } \delta_i = 0 \end{aligned}$$

et p_{xy^*} est la probabilité que le i^{e} appariement soit correct sachant que $x_i = x$ et $y^*_i = y^*$. L'estimateur du MV de $\pi_{c|x}$ en se servant des données appariées de manière probabiliste, $\mathbf{d}^*_i$, est alors

$$\tilde{\pi}_{c|x} = \tilde{n}_{c|x} \left(\sum_c \tilde{n}_{c|x} \right)^{-1} \qquad (7)$$

où

$$\tilde{n}_{c|x} = \Sigma_i \tilde{w}_{ic|x}, \qquad (8)$$

$$\begin{aligned} \tilde{w}_{ic|x} &= w^*_{ic|x} \hat{p}_{xy^*} + (1 - \hat{p}_{xy^*})\tilde{\pi}_{c|x} \quad \text{si } i \notin s_c \\ &= w^*_{ic|x} \quad \text{si } i \in s_c \\ &= \tilde{\pi}_{c|x} \quad \text{si } i \in s_c \text{ et } \delta_i = 0 \end{aligned} \qquad (9)$$

et

$$\hat{p}_{xy^*} = \left(\sum_{i \in s_c} w^*_{ic|x} \delta_i \right) \left(\sum_{i \in s_c} w^*_{ic|x} \right)^{-1}. \qquad (10)$$

La procédure d'estimation consiste à itérer (7), (8) et (9) jusqu'à la convergence. Plus précisément, l'algorithme est :

1. Calculer $\hat{p}_{xy^*}$ d'après (10).
2. Initialiser $\tilde{\pi}^{(0)}_{c|x}$ puis calculer $\tilde{w}^{(0)}_{c|x}$ d'après (9) et ensuite $\tilde{n}^{(0)}_{c|x}$ d'après (8).
3. Calculer $\tilde{\pi}^{(t)}_{c|x}$ d'après (7) en utilisant $\tilde{n}^{(t-1)}_{c|x}$.
4. Calculer $\tilde{w}^{(t)}_{c|x}$ d'après (9) en utilisant $\tilde{\pi}^{(t)}_{c|x}$ puis calculer $\tilde{n}^{(t)}_{c|x}$ d'après (8) en utilisant $\tilde{w}^{(t)}_{c|x}$.
5. Itérer 3 et 4 jusqu'à la convergence.

La valeur initiale $\tilde{\pi}^{(0)}_{c|x}$ pourrait être choisie comme étant l'estimation naïve de $\pi_{c|x}$, décrite plus haut à la section 3. Cependant, nous avons constaté que le choix de la valeur initiale n'est pas important.

3.2 Régression logistique

Nous décrivons ci-après deux méthodes du MV (méthodes 1 et 2) pour estimer $\boldsymbol{\beta}$ en se servant de données appariées de manière probabiliste, $\mathbf{d}^*$. Les deux méthodes donnent des estimations sans biais sous l'hypothèse que les appariements sont incorrects au hasard. Elles se distinguent par le niveau d'agrégation auquel sont estimées les probabilités qu'un appariement soit correct. La méthode 1 requiert l'obtention de ces probabilités à un niveau plus fin d'agrégation, ce qui pourrait signifier qu'elle produit des estimations plus variables que la méthode 2.

3.2.1 Méthode 1

L'espérance de y conditionnellement aux données appariées est

$$\begin{aligned} E_{\text{d}|\text{d}^*}(y_i \mid \mathbf{x}_i = \mathbf{x}, y^*_i = y^*) = \\ y^*_i p_{\mathbf{x}y^*} + (1 - p_{\mathbf{x}y^*})\upsilon_i \quad &\text{si } i \notin s_c \\ = y^*_i \quad &\text{si } i \in s_c \text{ et } \delta_i = 1 \\ = \upsilon_i \quad &\text{si } i \in s_c \text{ et } \delta_i = 0 \end{aligned}$$

et p_{xy^*} est la probabilité que le i^{e} appariement soit correct sachant que $x = x_i$ et $y^*_i = y^*$.

L'estimateur du MV est alors obtenu par itération en vue de trouver la solution, désignée par $\tilde{\boldsymbol{\beta}}$, pour $\boldsymbol{\beta}$ dans (5) avec y_i remplacé par $\tilde{y}_i$, où

$$\begin{aligned} \tilde{y}_i &= y_i^* \hat{p}_{\mathbf{x}y^*} + (1 - \hat{p}_{\mathbf{x}y^*})\tilde{\upsilon}_i \quad \text{si } i \notin s_c \\ &= y_i^* \quad \text{si } i \in s_c \text{ et } \delta_i = 1 \\ &= \tilde{\upsilon}_i \quad \text{si } i \in s_c \text{ et } \delta_i = 0, \end{aligned} \tag{11}$$

$\tilde{\upsilon}_i$ a la même forme que υ_i excepté que $\boldsymbol{\beta}$ est remplacé par $\tilde{\boldsymbol{\beta}}$, et $\hat{p}_{\mathbf{x}y^*}$ est la proportion estimée d'appariements corrects dans l'échantillon pour examen manuel pour chaque combinaison de $\mathbf{x}$ et y^*.

3.2.2 Méthode 2

Posons que $\mathbf{x}'\mathbf{y}$ dans (5) possède le k^{e} élément

$$r_k = \mathbf{x}_k'\mathbf{y} = \sum_i^n y_i x_{ik} = \sum_i^n r_{ik},$$

où $r_{ik} = y_i x_{ik}$. L'espérance de r_{ik} conditionnellement à $\mathbf{d}^*$ est

$$\begin{aligned} &E_{\mathrm{d}|\mathrm{d}^*}(r_{ik} \mid \mathbf{x}_i = \mathbf{x},\, y_i^* = y_i) = \\ &\quad [y_i^* p_{ky^*} + (1 - p_{ky^*})\upsilon_i]x_{ik} \quad \text{si } i \notin s_c \\ &\quad = y_i^* x_{ik} \quad \text{si } i \in s_c \text{ et } \delta_i = 1 \\ &\quad = \upsilon_i x_{ik} \quad \text{si } i \in s_c \text{ et } \delta_i = 0 \end{aligned} \tag{12}$$

et p_{ky^*} est la probabilité qu'un appariement avec $x_{ik} = 1$ soit correct sachant que $y_i^* = y^*$. L'estimateur du MV est alors obtenu par itération en vue de trouver la solution, désignée par $\tilde{\tilde{\boldsymbol{\beta}}}$, pour $\boldsymbol{\beta}$ dans (5) avec r_{ik} remplacé par $\tilde{r}_{ik}$, où

$$\begin{aligned} \tilde{r}_{ik} &= [y_i^* \hat{p}_{ky^*} + (1 - \hat{p}_{ky^*})\tilde{\tilde{\upsilon}}_i]x_{ik} \quad \text{si } i \notin s_c \\ &= y_i^* x_{ik} \quad \text{si } i \in s_c \text{ et } \delta_i = 1 \\ &= \tilde{\tilde{\upsilon}}_i x_{ik} \quad \text{si } i \in s_c \text{ et } \delta_i = 0, \end{aligned} \tag{13}$$

$\tilde{\tilde{\upsilon}}_i$ a la même forme que υ_i excepté que $\boldsymbol{\beta}$ est remplacé par $\tilde{\tilde{\boldsymbol{\beta}}}$, et $\hat{p}_{ky^*}$ est la proportion estimée d'appariements corrects dans l'échantillon pour examen manuel pour chaque combinaison de $\mathbf{x}$ et y^*. Autrement dit, si $y_i^* = 1$,

$$p_{ky^*} = \left(\sum_{i \in s_c}^n y_i^* x_{ik} \delta_i\right)\left(\sum_{i \in s_c}^n y_i^* x_{ik}\right)^{-1}$$

et si $y^* = 0$,

$$p_{ky^*} = \left(\sum_{i \in s_c}^n (1 - y_i^*) x_{ik} \delta_i\right)\left(\sum_{i \in s_c}^n (1 - y_i^*) x_{ik}\right)^{-1}.$$

Cette approche ne nécessite le calcul que de $2K$ probabilités d'après l'échantillon pour examen manuel et, en ce sens, pourrait être préférable à l'approche décrite à la section 3.2.1 qui requiert le calcul d'un plus grand nombre de probabilités.

3.3 Estimation de la variance par le bootstrap

À la présente section, nous décrivons comment calculer la variance des estimations du MV de la section 3. Désignons le paramètre d'intérêt par $\boldsymbol{\theta}$, présenté plus haut, et l'estimation de son MV par $\tilde{\boldsymbol{\theta}}$. L'estimation par le bootstrap (Rubin et Little 2003) de la variance de $\tilde{\boldsymbol{\theta}}$, dénotée par $\hat{v}_{\text{boot}}(\tilde{\boldsymbol{\theta}})$, s'obtient comme il suit :

1. Tirer un échantillon répété de taille n_x du ficher de données appariées, $\mathbf{d}^*$, par échantillonnage aléatoire simple avec remise. Désigner le r^{e} échantillon répété par $\mathbf{d}^*(r)$. Le r^{e} échantillon pour examen manuel répété est $s_c(r) = s_c \cap \mathbf{d}^*(r)$.
2. Calculer $\tilde{\boldsymbol{\theta}}(r)$ qui a la même forme que $\tilde{\boldsymbol{\theta}}$ excepté que $\mathbf{d}^*(r)$ est utilisé au lieu de $\mathbf{d}^*$ et que $s_c(r)$ est utilisé au lieu de s_c.
3. Répéter les étapes 1 et 2 R fois, où R est le nombre de répliques.
4. Calculer

$$\hat{v}_{\text{boot}}(\tilde{\boldsymbol{\theta}}) = \frac{1}{R}\sum\nolimits_{b=1}^{R}(\tilde{\boldsymbol{\theta}}(b) - \tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}(b) - \tilde{\boldsymbol{\theta}})'.$$

4. Analyse en présence d'appariements incorrects et d'enregistrements non appariés

À la présente section, nous discutons de deux moyens d'analyser les données appariées en présence d'appariements incorrects et d'enregistrements non appariés. Comme nous l'avons mentionné dans l'introduction, le problème de l'analyse en présence d'enregistrements non appariés possède des similitudes évidentes avec le problème de la non-réponse d'une unité, ou non-réponse totale. L'existence d'enregistrements non appariés peut entraîner la sur ou la sous-représentation de certaines caractéristiques dans le fichier de données appariées, ce qui peut biaiser l'analyse. Comme nous l'exposons de manière plus détaillée plus bas, nous tirons parti du fait que le mécanisme donnant lieu aux enregistrements non appariés ne peut dépendre que de $\mathbf{z}$.

Nous considérons ici deux méthodes d'inférence en présence d'appariements incorrects et d'enregistrements non appariés, où les enregistrements appariés sont désignés au moyen de l'indice $i = 1, ..., n^*$ (rappelons que le i^{e} enregistrement du fichier X est un *enregistrement non apparié* si $i \in U_{xy}$ *et* que l'enregistrement i n'a été apparié à aucun enregistrement du fichier Y). La méthode consiste à modéliser de façon indépendante les processus qui déterminent quels enregistrements sont appariés incorrectement et lesquels sont non appariés (voir la section 5 pour un exemple). Ces modèles requièrent qu'un sous-échantillon, désigné par s_{xc}, des enregistrements du fichier X soit soumis à un examen manuel. Les enregistrements compris dans le

sous-échantillon seront soit appariés, soit non appariés à des enregistrements du fichier Y. Les enregistrements appariés du sous-échantillon doivent être catégorisés comme étant correctement ou incorrectement appariés durant le processus d'examen manuel. Un enregistrement du sous-échantillon qui n'est pas apparié doit être classé comme étant *non apparié* ou *autre*. *Non apparié* signifie que l'enregistrement correspondant a été trouvé dans le fichier Y, mais qu'il n'y a pas eu d'appariement, tandis que *autre* indique que l'enregistrement correspondant n'a pas été découvert dans le ficher Y et est par conséquent considéré comme non existant. Cette seconde classification pourrait être beaucoup plus difficile et plus longue que la première, parce qu'elle suppose qu'un autre processus exempt d'erreur existe pour vérifier qu'il est correct que certains appariements n'aient pas été faits. De par leur nature, les enregistrements non appariés contiennent peu d'information permettant de déterminer quel est l'appariement correct, même durant l'examen manuel. Ce genre de processus peut ne pas exister, auquel cas la correction pour tenir compte des enregistrements non appariés semble impossible. Cependant, il pourrait comprendre un examen manuel des noms qui figurent dans les deux fichiers à apparier. Par exemple, la personne chargée de l'examen manuel pourrait se rendre compte que les noms *John O. Smith* et *Joh O. Smith* qui figurent dans deux enregistrements distincts pourraient, en fait, faire référence à la même personne (avec un « n » manquant dans le deuxième cas, peut-être à cause d'erreur de lecture optique), tandis que le processus automatisé d'appariement pourrait traiter les deux noms comme étant entièrement différents. L'examinateur peut alors décider que les deux enregistrements susmentionnés correspondent à la même personne et devraient donc être appariés. Bishop (2009) et Wright (2009) discutent des avantages de l'examen manuel.

La première méthode consiste à conditionner l'analyse sur une variable $\zeta_i = \zeta_i(\mathbf{z}_i)$. La variable ζ est définie de façon qu'en présence d'enregistrements non appariés, l'inférence soit sans biais conditionnellement à ζ. Le terme ζ est introduit parce que, dans de nombreux cas, il serait peu commode ou inutile de conditionner sur toute l'information contenue dans $\mathbf{z}$. Il est possible de donner à ζ_i une valeur non manquante, même quand $\mathbf{z}_i$ contient des valeurs manquantes. La forme exacte de la fonction $\zeta(\mathbf{z})$ devrait être justifiée après l'analyse du sous-échantillon, s_{xc}. Par exemple, si les personnes de moins de 20 ans sont sous-représentées dans le fichier de données appariées, ζ indiquerait si une personne a moins de 20 ans. Un moyen d'approcher l'analyse serait d'inclure ζ comme covariable dans le modèle de régression. La méthode décrite à la section 3 s'appliquerait alors directement. Cependant, les analystes pourraient souhaiter procéder à l'intégration sur ζ afin que cette variable ne figure pas dans le modèle logistique ou dans le tableau de contingence. À la section 4.2, nous expliquons comment le faire pour les tableaux de contingence. À la section 4.3, nous discutons d'une approche fondée sur la pseudo-vraisemblance qui consiste à attribuer aux enregistrements non appariés des pondérations visant à tenir compte de toute sur ou sous-représentation de certaines sous-populations dans les données appariées. De nouveau, le choix des pondérations devrait être justifié après l'analyse du sous-échantillon, s_{xc}, qui indique quels sont les enregistrements non appariés. Cet aspect est examiné plus en détail dans le contexte de l'étude empirique.

4.1 Pouvons-nous ignorer les enregistrements non appariés ?

Définissons la variable $\gamma_i = 1$ si l'enregistrement i du fichier X est non apparié et $\gamma_i = 0$ autrement. Soit aussi ζ_i une variable telle que $\zeta_i = 1, 2, ..., h, ...H$, où H est le nombre de catégories pour ζ. Nous pouvons ignorer le fait que des enregistrements sont non appariés si nous sommes prêts à supposer que, conditionnellement à $\mathbf{x}_i$, les distributions de y_i, γ_i et δ_i sont indépendantes. Techniquement, cette hypothèse mène à la factorisation suivante

$$p(y_i, \mathbf{x}_i, \delta_i, \gamma_i, \zeta_i) \propto$$
$$p(y_i \mid \mathbf{x}_i ; \boldsymbol{\theta})\, p(\delta_i \mid \mathbf{x}_i)\, p(\gamma_i \mid \mathbf{x}_i)\, p(\zeta_i)$$

où, de nouveau, $\boldsymbol{\theta} = \boldsymbol{\beta}$ ou $\boldsymbol{\Pi}$. Il vaut la peine de vérifier si cette hypothèse est valide dans le cas du sous-échantillon pour examen manuel. Si l'hypothèse est raisonnable, il n'est pas nécessaire d'appliquer les méthodes décrites aux sections 4.2 et 4.3, et les méthodes de la section 3 suffisent.

Nous pourrions ne pas vouloir émettre l'hypothèse susmentionnée, mais être prêts à supposer que, conditionnellement à $\mathbf{x}$ et ζ, les distributions de y_i, γ_i et δ_i sont indépendantes. Dans ce cas, nous disons que les enregistrements non appariés sont ignorables. Techniquement, cette hypothèse mène à la factorisation suivante,

$$p(y_i, \mathbf{x}_i, \delta_i, \gamma_i, \zeta_i) \propto$$
$$p(y_i \mid \mathbf{x}_i, \zeta_i ; \boldsymbol{\Lambda})\, p(\delta_i \mid \mathbf{x}_i ; \boldsymbol{\tau})\, p(\gamma_i \mid \mathbf{x}_i, \zeta_i)\, p(\zeta_i)$$

où $\boldsymbol{\Lambda}$ est le paramètre pour la distribution de $y_i \mid \mathbf{x}_i, \zeta_i$. Si nous nous intéressons à $p(y_i \mid \mathbf{x}_i ; \boldsymbol{\theta})$, mais non à $p(y_i \mid \mathbf{x}_i, \zeta_i ; \boldsymbol{\Lambda})$, une approche consiste à éliminer ζ_i de cette dernière par intégration (c'est-à-dire en prenant la moyenne sur toutes les valeurs possibles).

4.2 Maximum de vraisemblance conditionnel (MVC) pour les tableaux de contingence

Premièrement, paramétrisons la distribution conjointe de y_i, x_i et ζ_i par la distribution multinomiale de paramètre

$\boldsymbol{\Lambda}$. Définissons $\boldsymbol{\Lambda} = (\boldsymbol{\Pi}_1', ... \boldsymbol{\Pi}_h', ..., \boldsymbol{\Pi}_H')'$, où $\boldsymbol{\Pi}_h = (\boldsymbol{\pi}_{1h}', ..., \boldsymbol{\pi}_{gh}', ..., \boldsymbol{\pi}_{Gh}')'$, $\pi_{gh} = (\pi_{1|gh}, ..., \pi_{c|gh}, ..., \pi_{C|gh})'$ et $\pi_{c|gh}$ est la probabilité que $y_i = c$, $x_i = g$ et $\zeta_i = h$. L'estimateur du MV de $\boldsymbol{\Pi} = (\pi_{c|x})$ de la section 2.1 quand les erreurs d'appariement ne peuvent pas être ignorées est $\tilde{\tilde{\boldsymbol{\Pi}}} = (\tilde{\tilde{\pi}}_{c|x})$, où

$$\tilde{\tilde{\pi}}_{c|x} = \sum_{h=1}^{H} \tilde{\pi}_{c|xh} \hat{\pi}_{h|x}, \qquad (14)$$

où

$$\tilde{\pi}_{c|xh} = \tilde{n}_{c|xh} \left(\sum\nolimits_c \tilde{n}_{c|xh} \right)^{-1}, \qquad (15)$$

$\tilde{n}_{c|xh} = \Sigma_{i \in U_l} \tilde{w}_{ic|xh}$, $\Sigma_{i \in U_l}$ est la somme sur les n^* enregistrements appariés et $\hat{\pi}_{h|x}$ pour $h = 1, ..., H$ est l'estimation classique de la distribution marginale de $\boldsymbol{\zeta}$ sachant x dans le fichier X. En outre, si $i \notin s_c$,

$$\tilde{w}_{ic|xh} = w^*_{ic|xh} \hat{p}_{xy^*h} + (1 - \hat{p}_{xy^*h}) \tilde{\tilde{\pi}}_{c|xh}, \qquad (16)$$

$\hat{p}_{xy^*h}$ est la probabilité que le i^{e} appariement soit correct sachant que $x_i = x$, $\zeta_i = h$ et $y_i^* = y^*$, $w^*_{ic|xh} = 1$ si $x_i = x$, $\zeta_i = h$ et $y_i^* = y^*$, et $w^*_{ic|xh} = 0$ autrement. Si $i \in s_c$, alors $\tilde{w}_{ic|xh} = w^*_{ic|xh} = w_{ic|xh}$ s'il est déterminé que l'appariement est correct et $\tilde{w}_{ic|xh} = \tilde{\tilde{\pi}}_{c|xh}$ s'il est déterminé qu'il est incorrect.

L'estimateur du MV $\tilde{\tilde{\pi}}_{c|x}$ s'obtient par itération entre (14), (15) et (16) jusqu'à la convergence.

4.3 Pseudo-maximum de vraisemblance (PMV)

À la présente section, nous discutons d'une alternative au MVC décrit à la section 4.2, que nous appelons pseudo-maximum de vraisemblance (voir Chambers et Skinner 2003). Il s'agit essentiellement d'une approche de pondération, qui pourrait être plus facile à mettre en œuvre que le MVC et qui s'appuie sur la factorisation donnée à la section 4.2. Elle comprend la résolution de versions pondérées des fonctions de score, $\text{Score}(\boldsymbol{\pi}_x ; \mathbf{d}) = \mathbf{0}_{C-1}$ et $\text{Score}(\boldsymbol{\beta} ; \mathbf{d}) = \mathbf{0}_{\text{K}}$ pour trouver $\boldsymbol{\pi}_x$ et $\boldsymbol{\beta}$ respectivement, où le poids d'un enregistrement est égal à l'inverse de la probabilité que l'enregistrement demeure non apparié. Nous désignons la probabilité que l'enregistrement i ne restera pas non apparié par $t_i = E(\gamma_i)$, de sorte que les poids unitaires sont donnés par $q_i = t_i^{-1}$, où ici $i = 1, ..., n^*$. Par conséquent, l'estimateur du PMV pour $\pi_{c|x}$ est

$$\tilde{\pi}^{\text{PMV}}_{c|x} = \breve{n}_{c|x} \left(\sum\nolimits_c \breve{n}_{c|x} \right)^{-1}, \qquad (17)$$

où $\breve{n}_{c|xy} = \Sigma_{i \in U_l} q_i \tilde{w}_{ic|x}$. L'estimation de $\tilde{\pi}^{\text{PMV}}_{c|x}$ s'obtient par itération entre la mise à jour de $\tilde{w}_{ic|x}$, donné par (7), et (17) jusqu'à la convergence. L'estimateur du PMV de $\boldsymbol{\beta}$ est le même que l'estimateur du MV, excepté que, maintenant, dans l'équation d'estimation (5) les poids unitaires correspondaient à q_i. Une approche possible pour estimer l'exactitude des estimations du PMV sous appariement parfait consiste à utiliser la méthode du bootstrap décrite plus haut, mais en introduisant maintenant le poids q_i.

Pour illustrer la situation où les enregistrements non appariés ne peuvent pas être ignorés, considérons l'appariement d'une base de données contenant des renseignements personnels sur la situation d'emploi à une autre contenant des renseignements sur le niveau de scolarité. En outre, supposons que l'âge et le sexe, qui sont des variables corrélées à l'emploi et à la scolarité, sont disponibles dans l'une des bases de données. Après avoir effectué un examen manuel, nous pourrions constater que les enregistrements pour les jeunes hommes ont 50 % de chances de plus de rester non appariés que les enregistrements pour les femmes. Cela pourrait tenir au fait que les hommes sont moins susceptibles que les femmes de fournir des renseignements personnels qui sont utiles pour l'appariement. Clairement, dans le fichier de données appariées, il convient de donner aux enregistrements des hommes une pondération double de celle des enregistrements des femmes afin que l'analyse conjointe de la situation d'emploi et du niveau de scolarité soit sans biais.

5. Étude empirique

L'Australian Bureau of Statistics a réalisé une étude de qualité comportant l'appariement des enregistrements du Recensement de la population et du logement de 2006 aux enregistrements de la répétition générale de ce recensement (RGR). Durant la répétition générale du recensement, des renseignements ont été recueillis auprès de 78 349 personnes, un an avant le recensement. Durant le Recensement de 2006, des renseignements ont été recueillis auprès de plus de 19 millions de personnes.

Pendant une brève période, durant laquelle les données du Recensement de 2006 ont été traitées, les noms et adresses étaient disponibles pour le recensement ainsi que pour la répétition générale du recensement. Durant cette période, les deux fichiers d'enregistrements au niveau de la personne ont été appariés en utilisant des normes d'information différentes :

- la *norme d'or* (NO) consistait à utiliser le nom, l'adresse, l'îlot (*mesh block*) et certains éléments de données du recensement. L'îlot est une zone géographique contenant habituellement 50 logements. Tous les noms et adresses ont été détruits à la fin de la période de traitement des données du recensement.

- la *norme de bronze* (NB) consistait à utiliser l'îlot et certains éléments de données du recensement (c'est-à-dire à ne pas se servir du nom et de l'adresse). Il s'agit d'une méthode qu'il est proposé d'utiliser pour les futurs travaux d'appariement de l'ABS.

Une description détaillée de l'étude de la qualité et de la méthode d'appariement figure dans Solon et Bishop (2009). Le rôle de la norme d'or (NO) dans l'étude de la qualité est essentiel. Elle fournit une référence par rapport à laquelle peut être évaluée la fiabilité de la norme de bronze (NB). L'utilité de la NO comme référence est due au fait que le nom et l'adresse sont des variables puissantes pour repérer les personnes figurant dans le recensement et dans la répétition générale du recensement et qu'elles ont été soumises à un examen manuel approfondi. Nous supposons donc que la NO correspond à l'appariement parfait. Par conséquent, les différences entre les estimations fondées sur la NO et sur la NB sont interprétées comme étant des erreurs. Autrement dit, nous nous intéressons à la fiabilité de la NB *relativement* à la NO.

5.1 Méthodologie d'appariement

5.1.1 Variables de groupage et d'appariement et algorithme d'affectation 1 – 1

À la présente sous-section, nous donnons un aperçu de l'appariement des enregistrements de la RGR à ceux du recensement en appliquant la norme de bronze (NB). La méthode d'appariement comprend une série de passages, où chacun est défini par un ensemble de variables de groupage et d'appariement, ainsi qu'un algorithme d'affectation 1-1. Dans le cas de passages multiples, seuls les enregistrements non appariés au premier passage peuvent être appariés au deuxième, seuls les enregistrements non appariés au deuxième passage peuvent être appariés au troisième, et ainsi de suite.

Le tableau 1 donne les variables de groupage, désignées par « G » pour la NB. Par exemple, durant le passage 1, un enregistrement du recensement et un enregistrement de la RGR ne sont considérés comme un appariement possible que s'ils contiennent la même valeur pour l'îlot.

Les variables d'appariement sont utilisées pour mesurer le degré de concordance entre une paire d'enregistrements. Un haut niveau de concordance donne à penser que la probabilité que la paire d'enregistrements constitue un appariement correct est élevée. Le tableau 1 donne les variables d'appariement, désignées par « A », pour la NB. Par exemple, durant le passage 1 pour la NB, une gamme de variables, dont le jour, le mois et l'année de la naissance, le pays de naissance et le plus haut niveau de qualification sont utilisés comme variables de couplage.

Tableau 1
Exemple de variables de groupage (G) et d'appariement (A) utilisées pour apparier les données du Recensement de 2006 avec celles de la répétition générale du recensement. Différentes variables de groupage ont été utilisées lors de chacun des deux passages

Variable	Passage 1	Passage 2
Jour de la naissance	A	G
Mois de la naissance	A	G
Année de la naissance	A	G
Sexe	A	G
Statut d'Autochtone	A	A
Pays de naissance	A	A
Langue parlée	A	A
Année de l'arrivée	A	A
État matrimonial	A	A
Religion	A	A
Domaine d'études pour la qualification la plus élevée	A	A
Niveau de la qualification la plus élevée	A	A
Plus haut niveau de scolarité	A	A
Îlot (*Mesh block*)	G	A

À chaque passage, la sortie comprend un score pour chaque paire d'enregistrements. Le score est une mesure du degré de concordance entre les enregistrements formant la paire. Nous remettons à plus tard la définition formelle du score (pour des détails, voir (3.6), Conn et Bishop 2006), mais nous illustrons plus bas comment il peut être interprété. Considérons la NB au passage 2, pour lequel les enregistrements d'une paire contiennent la même date de naissance complète et le même sexe ; une paire d'enregistrements recevra un score de 23,5 s'il y a concordance pour l'îlot (+17) et l'année d'arrivée (+8) et désaccord pour la religion (-1,5) (dans cet exemple, la situation de concordance pour les autres variables d'appariement contribuerait également au score, mais nous les ignorons pour simplifier l'illustration). La contribution de la concordance pour l'îlot (+17) est plus grande celle de la concordance concernant l'année d'arrivée (+8) parce que la première est moins susceptible que la seconde d'avoir lieu par chance uniquement.

Afin de formaliser la cible de l'algorithme d'appariement, désignons le score pour l'enregistrement *i* de la RGR et l'enregistrement *j* du recensement durant le passage *p* en appliquant la NB par r_{pij}. L'ensemble des scores des paires d'enregistrements r_{pij} et le seuil d'exclusion f_p sont utilisés par le progiciel d'appariement *Febrl* (voir Christen et Churches 2005) pour déterminer l'ensemble optimal d'appariements pour le passage *p*. Le terme f_p est la valeur minimale du score pour qu'une paire d'enregistrements soit considérée comme un appariement durant le passage *p*. L'algorithme *Febrl* cherche à maximiser $\sum_i r_{pij}$, sous la contrainte $r_{pij} > f_p$. De toute évidence, le nombre d'appariements dépend de f_p.

Dans la suite, nous évaluons la NB au moyen de deux ensembles différents de seuils, où un ensemble de seuils est défini par les seuils des passages 1 et 2. Le premier seuil,

que nous disons très faible (TF), est considéré comme un seuil optimal, puisque, pour une gamme de seuils, les estimations naïves obtenues étaient les estimations « les plus proches » des estimations correspondantes produites en appliquant la NO (voir Bishop 2009). Le deuxième seuil, que nous disons extrêmement faible (EF), a effectivement pour objectif de maximiser le nombre d'enregistrements de la RGR qui sont appariés. Ci-après, nous désignons les deux fichiers d'enregistrements appariés sous la NB par les noms de leur seuil, TF et EF.

5.1.2 Résultats d'appariement

Sous la norme d'or (NO), 70 274 des 78 349 enregistrements de la RGR ont été appariés. Sous l'hypothèse que la NO correspond à l'appariement parfait, 8 075 personnes possédaient un enregistrement dans la RGR, mais non dans le recensement. En réalité, la NO n'est pas parfaite. Pour une discussion à ce sujet, voir Bishop, 2009.

Sous la norme de bronze avec seuil très faible (TF), 57 790 enregistrements de la RGR ont été appariés. Des 70 274 enregistrements de la RGR qui ont été appariés en appliquant la NO, 13 784 sont demeurés non appariés, 700 ont été appariés incorrectement et 55 790 ont été appariés correctement an appliquant le seuil TF. En outre, 1 300 enregistrements ont été appariés en appliquant le seuil TF, mais non en appliquant la NO et sont également des appariements incorrects. Donc, il y a eu, en tout, 2 000 (= 700 + 1 300) appariements incorrects.

Sous la norme de bronze avec seuil extrêmement faible (EF), 74 350 enregistrements de la RGR ont été appariés. Des 70 274 enregistrements de la RGR qui ont été appariés en appliquant la NO, 2 811 sont demeurés non appariés, 9 793 ont été appariés incorrectement et 57 670 ont été appariés correctement en appliquant la norme de bronze avec le seuil EF. En outre, 6 887 enregistrement de la RGR ont été appariés sous la norme de bronze EF, mais non sous la NO.

En résumé, 97 % des appariements du fichier TF sont corrects et 20 % (=13 784/70 274) des enregistrements de la RGR appariés en appliquant la NO demeurent non appariés. Les chiffres correspondants pour le fichier EF sont de 78 % et 4 % (=2 811/70 274).

5.1.3 Modélisation de la probabilité qu'un appariement soit correct

Pour *tous* les appariements dans les conditions EF et TF, nous savions s'il s'agissait d'un appariement correct ou incorrect (par exemple, si un appariement EF est également produit par la NO, cet appariement est correct. Sinon, l'appariement EF est incorrect). Par conséquent, p_{xy^*} de la section 3.1 était connue pour la NO. Toutefois, pour simuler la réalité, nous avons estimé p_{xy^*} d'après un échantillon pour examen manuel de taille 1 000 qui a été tiré des fichiers de données appariées par échantillonnage aléatoire simple.

5.1.4 Modélisation de la probabilité qu'un enregistrement demeure non apparié

À chaque enregistrement de la RGR apparié en appliquant la NO, nous avons attribué une variable indiquant si l'enregistrement était non apparié en appliquant la NB. Autrement dit, si l'enregistrement demeurait non apparié sous la NB, la variable indicatrice prenait la valeur « 1 » et autrement, la valeur « 0 ». Nous avons ajusté un modèle logistique en utilisant la NO, où la variable dépendante était la variable indicatrice susmentionnée et les variables explicatives provenaient de la RGR. Le modèle contenait plus de 20 variables explicatives qui ont été choisies selon la méthode classique de sélection ascendante/descendante. Les variables explicatives comprennent le niveau de scolarité, la langue, la naissance outre-mer, le statut d'autochtone, et les indicateurs de variable clé manquante, tels que l'*îlot* (*meshblock*). La prédiction résultante a donné t_i qui est utilisé plus bas pour mettre en œuvre la méthode du pseudo-MV pour les tableaux de contingence ainsi que la régression logistique.

5.2 Résultats de l'analyse tabulaire

Le tableau 2 donne les résultats du recoupement des situations d'emploi des personnes autochtones déclarées à la RGR et au recensement. Le tableau 2a montre que en appliquant la NO, l'estimation de la proportion d'Autochtones occupés au recensement, sachant qu'ils étaient occupés au moment de la RGR, est de 78,3 %. L'estimation naïve correspondante dans les conditions TF, qui reposent sur l'hypothèse que les données sont parfaitement appariées, est de 86,7 %. Même après avoir remplacé chacun des 700 appariements TF incorrects par l'appariement correct correspondant et écarté les 1 300 enregistrements non appariés pour lesquels aucun appariement correct n'existe, l'estimation naïve demeure pratiquement inchangée, soit 86,0 % (*appariements or* dans le tableau 2a). nous voyons donc que la différence entre les estimations TF et NO ne résultent pas tellement d'appariements incorrects, et qu'elle est due principalement à des enregistrements non appariés. Cela explique partiellement pourquoi l'estimation du MV (86,4 %) dans les conditions EF (voir la section 3.1), qui comporte une correction pour les appariements incorrects seulement, ne produit qu'une faible amélioration. Le MV conditionnel (MVC) (voir la section 4) a été examiné pour essayer de réduire l'erreur due aux enregistrements non appariés risquant de donner lieu à une représentation incorrecte, pour ce qui est de l'âge et du sexe, dans le fichier de données appariées. L'estimation de l'emploi par le MVC

était de 86,6 %. L'amélioration produite par le MVC n'était pas importante, ce qui indique que le mécanisme sous-jacent qui donne des enregistrements non appariés ne dépendait pas de l'âge ni du sexe. Les estimations du PMV (voir la section 4) n'ont pas non plus amélioré beaucoup la situation, ce qui indique que le modèle logistique décrit à la section 5.1.4 n'expliquait pas le mécanisme générant les enregistrements non appariés. Curieusement, l'estimation du MV en utilisant le fichier EF était de 81,8 %, c'est-à-dire l'estimation de loin la plus proche de l'estimation de 78,3 % sous la NO. Or, dans le cas du seuil EF, les appariements incorrects sont la principale source d'erreur, c'est-à-dire le type d'erreur d'appariement que corrige l'estimateur du MV. Par conséquent, la correction des erreurs dues à des appariements incorrects a été beaucoup plus fructueuse que celle des erreurs dues à des enregistrements non appariés.

Les erreurs types des estimations NO, naïves et MV sont indiquées entre parenthèses dans le tableau 2a. Dans le cas de TF et de EF, les erreurs-types de l'estimation MV sont environ 25 % et 75 % plus grandes, respectivement, que les erreurs-types de l'estimation naïve correspondantes. En outre, les erreurs-types de l'estimation MV pour EF sont légèrement plus faibles que pour TF, ce qui signifie que les appariements supplémentaires effectués en appliquant le seuil EF valaient la peine. Clairement, l'inférence naïve dans les conditions EF surestime le niveau de confiance des estimations. Pour les conditions TF, les estimations naïve et MV et leurs erreurs-types sont très proches.

Quel que soit le seuil utilisé, les estimations MV des tableaux 2a, b et c sont systématiquement plus proches de l'estimation NO que l'estimation naïve correspondante. Par exemple, dans le tableau 2b, l'estimation MV pour le seuil TF est de 36,9 %, c'est-à-dire appréciablement plus proche de l'estimation NO de 37,9 % que l'estimation naïve de 33,3 %. En fonction des estimations du tableau 2, nous pourrions soutenir que la décision d'utiliser le seuil TF ou EF n'est pas tellement importante, à condition d'utiliser l'estimateur du MV.

Le tableau 3 ressemble au tableau 2, à part le fait qu'il décrit les analyses des enregistrements appariés pour toutes les personnes de 15 ans et plus, plutôt que pour les Autochtones seulement. De nouveau, l'estimateur du MV produit systématiquement une amélioration dans le cas du seuil EF, mais non dans le cas du seuil TF. Le tableau 4 donne la situation d'étudiant en 2006 pour les personnes qui étaient étudiantes en 2005. Encore une fois, le MV donne généralement des estimations plus proches de l'estimation or correspondante, en particulier pour le seuil EF.

Tableau 2
Pourcentages d'Autochtones dans les diverses catégories d'emploi en 2006 sachant leur catégorie d'emploi en 2005. Pour chaque ensemble de données appariées, soit Seuil très faible ou extrêmement faible, les méthodes d'estimation peuvent être comparées à la norme d'or

	Estimations pour divers ensembles de donnée appariées et méthodes							
a : Autochtones occupés en 2005								
Situation en 2006	**or**			**Seuil très faible**			**Seuil extrêmement faible**	
		Naïve	**Appariements or**	**MV**	**PMV**	**MVC**	**Naïve**	**MV**
Occupé	78,3	86,7	86,0	86,4	86,6	86,1	71,9	81,8
	(1,7)	(2,4)		(3,0)			(1,7)	(2,9)
En chômage	3,7	4,2	4,3	4,1	4,1	4,2	6,3	3,3
	(0,84)	(1,2)		(2,5)			(0,82)	(2,1)
Inactif	17,8	9,0	9,6	9,3	9,1	9,6	21,6	14,7
	(1,6)	(2,4)		(3,1)			(1,6)	(2,8)

b : Autochtones en chômage en 2005					
Situation en 2006	**or**	**Très faible**		**Extrêmement faible**	
		Naive	**ML**	**Naive**	**ML**
Occupé	27,5	27,7	27,2	35,2	23,8
En chômage	34,4	38,9	36,4	32,3	38,0
Inactif	37,9	33,3	36,3	32,3	38,0
c : Autochtones inactifs en 2005					
Occupé	13,7	10,8	10,7	24,3	10,5
En chômage	5,8	7,6	7,4	6,3	5,8
Inactif	80,4	81,5	81,8	69,2	83,5

Tableau 3
Pourcentages de l'ensemble des personnes de plus de 15 ans dans les diverses catégories d'emploi en 2006 sachant leur catégorie d'emploi en 2005. Pour chaque ensemble de données appariées, soit seuil Très Faible ou Extrêmement Faible, les méthodes d'estimation peuvent être comparées à la norme d'or

	Estimations pour divers ensembles de donnée appariées et méthodes				
Situation en 2006	**or**	**Très faible**		**Extrêmement faible**	
		Naïve	**MV**	**Naïve**	**MV**
a : Personnes occupées en 2005					
Occupé	91,8	92,2	92,6	89,7	92,4
En chômage	1,8	1,7	1,6	1,9	1,6
Inactif	6,2	6,1	5,6	8,3	5,8
b : Personnes en chômage en 2005					
Occupé	44,5	44,3	44,0	49,4	43,8
En chômage	26,8	26,6	27,5	22,8	27,6
Inactif	28,6	28,7	28,4	27,6	28,5
c : Personnes inactives en 2005					
Occupé	12,1	12,3	11,1	16,8	11,0
En chômage	3,1	3,1	3,0	3,0	3,0
Inactif	84,7	84,5	85,7	80,1	85,9

Tableau 4
Situation d'étudiant en 2006 pour les étudiants du secondaire en 2005

Situation d'étudiant en 2006	**Or**	**Très faible**		**Extrêmement faible**	
		Naïve	**MV**	**Naïve**	**MV**
Élève du secondaire	79,3	79,3	79,6	77,4	79,6
Études secondaires terminées	14,0	14,3	13,7	14,7	14,1
Études secondaires non terminées	6,6	6,3	6,6	7,8	6,2

5.3 Simulation

L'étude par simulation qui suit illustre les problèmes que pose l'analyse naïve et les avantages de l'utilisation de la méthode décrite dans le présent article. Dans la simulation, les fichiers X et Y, contenant chacun 2 000 enregistrements, sont générés indépendamment 400 fois, chaque fichier généré étant désigné par X(r) et Y(r), respectivement, avec $r = 1, ..., 400$. En particulier, sur X(r), x_i est tiré aléatoirement de la loi de Bernoulli de paramètre 0,5. Sur Y(r), y_i est tiré aléatoirement de la loi de Bernoulli de paramètre υ_i, où $\upsilon_i = 1/[1+\exp(\beta_0 + \beta_1 x_i)]$, $\boldsymbol{\beta} = (\beta_0, \beta_1)', \beta_0 = -0,5$, $\beta_1 = 1,5$. Le r^{e} ensemble de données imparfaitement appariées, $\mathbf{d}^*(r)$, est généré en appariant correctement chaque enregistrement du fichier Y(r) à un enregistrement du fichier X(r) avec la probabilité $p = 0,8, 0,90, 0,95$ et 1. Pour chaque r^{e} ensemble de données appariées, un échantillon pour examen manuel de 300 appariements est sélectionné. Chaque appariement contenu dans l'échantillon pour examen manuel est classé comme étant correct ou incorrect. Nous résumons la performance de l'estimateur du MV décrit à la section 3.2.2 et de la méthode naïve, qui repose sur l'hypothèse qu'il n'existe aucune erreur d'appariement, par leur taux de couverture à 95 % et leur erreur quadratique moyenne (EQM). Les taux de couverture sont fondés sur les erreurs-types calculées par la méthode du bootstrap décrite à la section 3.3 avec $R = 40$ répliques. L'EQM de $\tilde{\boldsymbol{\beta}}$ est calculée par

$$\text{EQM}(\tilde{\boldsymbol{\beta}}) = \frac{1}{400}\sum_{r=1}^{400}(\tilde{\boldsymbol{\beta}}_r - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}}_r - \boldsymbol{\beta})'$$

où $\tilde{\boldsymbol{\beta}}_r$ est l'estimation du MV de $\boldsymbol{\beta}$ d'après $\mathbf{d}^*(r)$.

Le tableau 5 montre que l'approche naïve produit de mauvais taux de couverture, à cause de son biais important en présence d'erreurs d'appariement et, par conséquent, son EQM relativement élevée. Pour le MV-méthode 1, les taux de couverture sont très proches des taux nominaux. Les résultats montrent que, si le pourcentage d'appariements corrects passe de 100 % à 80 %, l'EQM pour l'estimation MV augmente d'un facteur 3 environ pour β_0 et β_1 (les taux de couverture et l'EQM des méthodes 1 et 2 d'estimation du MV étant fort semblables, seuls les résultats pour la première sont présentés ici).

Tableau 5
Erreur quadratique moyenne et taux de couverture pour les données simulées appariées, avec la probabilité p qu'un appariement soit correct

		Erreur quadratique moyenne				**Taux de couverture à 95 %**		
		0,8	0,9	0,95	1	0,8	0,9	0,95
Naïve	β_0	0,024	0,010	0,0056	0,0043*	0,35	0,80	0,93
	β_1	0,11	0,038	0,016	0,011*	0,05	0,62	0,88
MV-Méthode 1	β_0	0,013	0,0078	0,0055	0,0043*	93,0	94,25	93,5
	β_1	0,031	0,018	0,013	0,011*	96,0	94,5	96,25

*Quand $p = 1$, les estimateurs naïf et MV sont identiques par définition.

6. Discussion

L'appariement des données est une méthode appropriée quand des ensembles de données doivent être jumelés en vue de renforcer la dimension temporelle ou des aspects tels que la portée ou la profondeur des renseignements. L'appariement des données est utilisé de plus en plus fréquemment par les organismes statistiques partout dans le monde. Il est bien connu que des erreurs peuvent se produire durant l'appariement des fichiers, par exemple quand des méthodes probabilistes sont appliquées. Toutefois, peu d'études traitant de la façon de faire des inférences valides en présence de ce genre d'erreurs ont été publiées. Le présent article offre des conseils méthodologiques et pratiques en vue d'aider les analystes dans ce domaine.

En général, traiter naïvement un fichier de données appariées comme si l'appariement était parfait donne lieu à des estimations biaisées. L'analyste ne devrait adopter l'approche naïve que si le nombre d'enregistrements non appariés, définis comme étant des enregistrements qui pourraient être appariés correctement, mais qui ne l'ont pas été du tout, ainsi que le nombre d'appariements incorrects sont négligeables. Le présent article décrit une approche fondée sur le maximum de vraisemblance en vue de faire de inférences valides en présence des deux sources d'erreur. Cette approche s'appuie sur l'algorithme EM bien connu et est facile à appliquer en pratique. La méthode peut être utilisée quand l'un des fichiers n'est pas nécessairement un sous-ensemble de l'autre et que l'appariement comporte des passages multiples. Ces situations se présentent fréquemment en pratique, comme en témoignent de nombreux exemples récents à l'Australian Bureau of Statistics. L'étude empirique montre que l'approche du MV améliore de manière significative les estimations fondées sur des données appariées.

Dans le cas particulier où le fichier X est obtenu par tirage d'un échantillon aléatoire du fichier Y, la procédure d'estimation décrite n'est pas « complètement » celle du maximum de vraisemblance, parce qu'elle ne s'appuie pas sur le fait que les totaux de population pour le fichier Y sont connus. Bien que l'inférence selon la méthode décrite ici demeure valide dans ce cas, il y aurait peut-être moyen de la rendre plus efficace (voir Scott et Wild 1997).

Remerciements

Les auteurs remercient Raymond Chambers et deux examinateurs de *Techniques d'enquête* de leur contribution au présent article.

Bibliographie

Australian Bureau of Statistics (2008). Census Data Enhancement - Indigenous Mortality Quality Study, 2006-07. Document d'information n° de catalogue 4723.0.

Bishop, G. (2009). Assessing the Likely Quality of the Statistical Longitudinal Census Dataset. Travaux de recherche de méthodologie, n° de catalogue 1351.0.55.026, Australian Bureau of Statistics, Canberra.

Chambers, R., Chipperfield, J.O., Davis, W. et Kovačević, M. (2009). Regression Inference Based on Estimating Equations and Probability-Linked Data. Soumis pour publication.

Chambers, R.L., et Skinner, C.J. (2003). *Analysis of Survey Data*. New York : John Wiley & Sons, Inc.

Chambers, R. (2008). Regression analysis of probability-linked data. *Statisphere*, Volume 4, http://www.statisphere.govt.nz/official-statistics-research/series/vol-4.htm.

Christen, P., et Churches, T. (2005). Febrl – Freely extensible biomedical record linkage. Version 0.3.1, vue le 17 novembre 2008, http://cs.anu.edu.au/~Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/contents.html.

Conn, L., et Bishop, G. (2006). Exploring Methods for Creating a Longitudinal Census Dataset. Articles du comité consultative sur la méthodologie, n° de catalogue 1352.0.55.076, Australian Bureau of Statistics, Canberra.

Fair, M. (2004). Generalized record linkage system-Statistics Canada's record linkage software. *Austrian Journal of Statistics*, 33(1 et 2), 37-53.

Fellegi, I.P., et Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Fuller, W.A. (1987). *Measurement Error Models*. New York : John Wiley & Sons, Inc.

Hausman, J.A., Abrevaya, J. et Scott-Morton, F.M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87, 239-269.

Herzog, T.N., Scheuren, F.J. et Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques*. New York : Springer.

Holman, C.D.J., Bass, A.J., Rouse, I.L. et Hobbs, M.S.T. (1999). Population-based linkage of health records in Western Australia: Development of a health services research linked database. *Australian and New Zealand Journal of Public Health*, 23(5), 453-459.

Lahiri, P., et Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 222-230.

National Center for Health Statistics (2009). Linkages between Survey Data from the National Center for Health Statistics and Program Data from the Social Security Administration. Rapport de méthodologie, http://www.cdc.gov/nchs/data/datalinkage/ssa_methods_report_2009.pdf.

Rubin, D.B., et Little, R.J.A. (2003). *Statistical analysis of missing data*, 2e Édition. New York : John Wiley & Sons, Inc.

Scheuren, F., et Winkler, W.E. (1993). Analyse de régression de fichiers de données couplés par ordinateur. *Techniques d'enquête*, 19, 45-65.

Scott, A.J., et Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84, 57-71.

Solon, R., et Bishop, G. (2009). A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset. Travaux de recherche de méthodologie, n° de catalogue 1351.0.55.025, Australian Bureau of Statistics, Canberra.

Winkler, W.E. (2001). Record Linkage Software and Methods for Merging Administrative Lists. Collection de rapports de recherche statistique, n° RR2001/03, Bureau of the Census.

Winkler, W.E. (2005). Approximate String Comparator Search Strategies for Very Large Administrative Lists. Collection de rapports de recherche statistique, n° RRS2005/02, Bureau of the Census.

Wright, J., Bishop, G. et Ayre, T. (2009). Assessing the Quality of Linking Migrant Settlement Records to Census Data. Travaux de recherche de méthodologie, n° de catalogue 1351.0.55.027, Australian Bureau of Statistics, Canberra.

Techniques d'enquête, juin 2011
Vol. 37, N° 1, pp. 31-44
Statistique Canada, N° 12-001-X au catalogue

Estimation sur petits domaines hiérarchique bayésienne sous un modèle spatial avec application à des données d'enquête sur la santé

Yong You et Qian M. Zhou [1]

Résumé

Dans le présent article, nous étudions l'estimation sur petits domaines en nous servant de modèles au niveau du domaine. Nous considérons d'abord le modèle de Fay-Herriot (Fay et Herriot 1979) pour le cas d'une variance d'échantillonnage connue lissée et le modèle de You-Chapman (You et Chapman 2006) pour le cas de la modélisation de la variance d'échantillonnage. Ensuite, nous considérons des modèles spatiaux hiérarchiques bayésiens (HB) qui étendent les modèles de Fay-Herriot et de You-Chapman en tenant compte à la fois de l'hétérogénéité géographiquement non structurée et des effets de corrélation spatiale entre les domaines pour le lissage local. Les modèles proposés sont mis en œuvre en utilisant la méthode d'échantillonnage de Gibbs pour une inférence entièrement bayésienne. Nous appliquons les modèles proposés à l'analyse de données d'enquête sur la santé et comparons les estimations fondées sur le modèle HB aux estimations directes fondées sur le plan. Nos résultats montrent que les estimations fondées sur le modèle HB ont de meilleures propriétés que les estimations directes. En outre, les modèles spatiaux au niveau du domaine proposés produisent des CV plus petits que les modèles de Fay-Herriot et de You-Chapman, particulièrement pour les domaines ayant trois domaines voisins ou plus. Nous présentons aussi une comparaison des modèles bayésiens et une analyse de l'adéquation du modèle.

Mots clés : Modèle au niveau du domaine ; comparaison de modèles bayésiens ; taux de prévalence de la maladie ; échantillonnage de Gibbs ; modèle spatial hiérarchique ; vérification du modèle prédictif *a posteriori* ; variance d'échantillonnage.

1. Introduction

Les méthodes d'estimation sur petits domaines fondée sur un modèle ont été utilisées à grande échelle en pratique en raison de la demande croissante d'estimations précises pour des régions locales et divers petits domaines. En général, les enquêtes par sondage sont conçues pour fournir des estimations fiables pour de grandes régions ou des agrégats de petits domaines, tel que le pays dans son ensemble et les provinces. Des estimations par sondage directes fondées uniquement sur des données d'échantillon propres au domaine fournissent habituellement des estimations fiables du paramètre d'intérêt pour les grands domaines. Pour les petits domaines, particulièrement certaines petites régions géographiques ou des petits domaines particuliers, les estimations directes sont susceptibles de produire de grandes erreurs-types, à cause de la petite taille des échantillons dans ces petits domaines. Par conséquent, en inférence pour les petits domaines, il est nécessaire d'emprunter de l'information à des domaines connexes pour produire des estimations indirectes qui augmentent les tailles effectives d'échantillons et donc, la précision des estimations. Il est généralement reconnu aujourd'hui que ces estimations indirectes doivent être fondées sur des modèles explicites qui fournissent des liens avec les domaines connexes grâce à l'utilisation de données supplémentaires, telles que des chiffres de recensement ou des données de dossiers administratifs ; voir, par exemple, Rao (2003) et Jiang et Lahiri (2006) pour une discussion plus approfondie des méthodes d'estimation sur petits domaines fondée sur un modèle. Les estimations fondées sur un modèle sont obtenues en vue d'améliorer les estimations directes fondées sur le plan en ce qui a trait à la précision et à la fiabilité, c'est-à-dire réduire les coefficients de variation (CV). Ces modèles d'estimation sur petits domaines se répartissent en deux grandes catégories, à savoir les modèles au niveau du domaine et les modèles au niveau de l'unité. Les modèles au niveau du domaine sont fondés sur des estimations directes au niveau du domaine et les modèles au niveau de l'unité, sur des observations individuelles faites dans les petits domaines. Dans le présent article, nous nous concentrons sur les modèles au niveau du domaine qui empruntent de l'information à diverses régions pour améliorer les estimations directes sous le plan.

Parmi les modèles au niveau du domaine, celui de Fay-Herriot (Fay et Herriot 1979) est un modèle élémentaire très utilisé en pratique pour obtenir des estimations fondées sur un modèle fiables pour de petits domaines. Fondamentalement, le modèle de Fay-Herriot possède deux composantes, à savoir un modèle d'échantillonnage pour les estimations directes et un modèle de lien pour les paramètres d'intérêt. Le modèle d'échantillonnage comprend l'estimation par sondage direct et la variance d'échantillonnage correspondante. Le modèle de Fay-Herriot repose sur l'hypothèse que la variance d'échantillonnage est connue dans le modèle. Habituellement, on obtient un estimateur lissé de la variance

1. Yong You, Division de la recherche et de l'innovation en statistiques, Statistique Canada. Courriel : yongyou@statcan.gc.ca ; Qian M. Zhou, Département de biostatistique, Université Harvard.

d'échantillonnage que l'on traite ensuite comme étant connue dans le modèle. Wang et Fuller (2003), ainsi que You et Chapman (2006) ont étudié la situation où les variances d'échantillonnage sont inconnues et modélisées séparément par des estimateurs directs. Dans le présent article, nous examinerons les méthodes de lissage et de modélisation pour les variances d'échantillonnage dans le modèle d'échantillonnage.

Le modèle de lien relie le paramètre d'intérêt à un modèle de régression contenant des effets aléatoires propres au domaine. Dans le modèle de Fay-Herriot, on suppose habituellement que les effets aléatoires de domaine sont des variables aléatoires indépendantes et identiquement distribuées (*iid*) qui suivent une loi normale pour refléter les variations géographiquement non structurées entre les domaines. Cependant, dans certaines applications aux petits domaines, particulièrement dans les problèmes d'estimation de la santé publique, la variation géographique de la prévalence d'une maladie est un sujet d'intérêt, et l'estimation des profils spatiaux globaux de risque et l'emprunt d'information à diverses régions pour réduire les variances des estimations finales sont deux éléments importants. Donc, il pourrait être plus raisonnable de construire des modèles spatiaux sur les effets aléatoires propres aux domaines pour traduire l'interdépendance spatiale de ces effets. Les modèles spatiaux sont généralement utilisés pour l'estimation sur petits domaines ayant trait à la santé, et plusieurs de ces modèles ont été proposés (par exemple, Cressie 1990 ; Ghosh, Natarajan, Stroud et Carling 1998 ; Maiti 1998 ; Ghosh, Natarajan, Walter et Kim 1999 ; He et Sun 2000 ; Moura et Migon 2002 ; Singh, Shukla et Kundu 2005 ; Souza, Moura et Migon 2009). Best, Richardson et Thomson (2005) ont présenté une revue complète des modèles spatiaux pour la cartographie des maladies. Rao (2003) a également discuté de plusieurs modèles spatiaux pour petits domaines.

L'objectif du présent article est d'examiner des modèles à corrélation spatiale pour l'estimation sur petits domaines et d'illustrer leur utilité en les appliquant à des données d'enquête sur la santé. La présentation de l'article est la suivante. À la section 2, nous commençons par étudier les modèles au niveau du domaine, y compris le modèle de Fay-Herriot et les modèles de lien à corrélation spatiale. Ensuite, à la section 3, nous proposons des modèles d'estimation sur petits domaines hiérarchiques bayésiens (HB) avec corrélation spatiale et obtenons l'inférence HB pour les paramètres de petit domaine grâce à la méthode d'échantillonnage de Gibbs. À la section 4, nous appliquons les modèles proposés à l'analyse de données sur des petits domaines provenant de l'Enquête sur la santé dans les collectivités canadienne. Nous comparons la performance des estimations fondées sur un modèle aux estimations directes fondées sur le plan et, en outre, nous comparons les modèles proposés au modèle de Fay-Herriot et au modèle de You-Chapman (You et Chapman 2006) afin d'étudier les effets de la prise en compte de la structure spatiale sur les effets aléatoires de domaine. Nous présentons aussi une comparaison des modèles bayésiens et une analyse de l'adéquation du modèle. Enfin, à la section 5, nous tirons certaines conclusions.

2. Modèles et inférence pour petits domaines

2.1 Modèle de Fay-Herriot

Soit θ_i le paramètre d'intérêt pour le i^{e} domaine, où $i = 1, ..., m$, et m est le nombre total de domaines. Le modèle de Fay-Herriot repose sur l'hypothèse que les θ_i sont reliés à des données auxiliaires particulières au domaine $\mathbf{x}_i = (x_{i1}, ..., x_{ip})'$ au moyen du modèle de régression linéaire qui suit :

$$\theta_i = \mathbf{x}_i'\beta + v_i,\ i = 1, ..., m \qquad (1)$$

où $\beta = (\beta_1, ..., \beta_p)'$ est le vecteur de dimension $p \times 1$ des coefficients de régression, et les v_i sont les effets aléatoires propres au domaine que l'on suppose être *iid* avec $\mathrm{E}(v_i) = 0$ et $\mathrm{Var}(v_i) = \sigma_v^2$. L'hypothèse de normalité peut également être incluse. Ce modèle est appelé modèle de lien pour θ_i. Le modèle de Fay-Herriot repose aussi sur l'hypothèse qu'un estimateur direct fondé sur le plan y_i, qui est habituellement sans biais sous le plan pour le paramètre d'intérêt θ_i, existe quand la taille d'échantillon du domaine $n_i > 1$. On suppose habituellement que

$$y_i = \theta_i + e_i,\ i = 1, ..., m \qquad (2)$$

où les e_i sont les erreurs d'échantillonnage associées à l'estimateur direct y_i. Nous supposons en outre que les e_i sont des variables aléatoires normales indépendantes de moyenne $E(e_i \mid \theta_i) = 0$ et de variance d'échantillonnage $\mathrm{Var}(e_i \mid \theta_i) = \sigma_i^2$. Le modèle (2) est appelé modèle d'échantillonnage pour l'estimateur direct y_i. La combinaison des deux composantes (1) et (2) donne un modèle linéaire à effets mixtes, le modèle de Fay-Herriot, de la forme

$$y_i = \mathbf{x}_i'\beta + v_i + e_i,\ i = 1, ..., m. \qquad (3)$$

Dans le modèle de Fay-Herriot élémentaire (3), on suppose habituellement que les variances d'échantillonnage σ_i^2 sont connues, ce qui est une hypothèse très forte. En général, nous pouvons utiliser les estimations directes des variances d'échantillonnage d'après les données d'enquête, mais ces estimations sont instables si les tailles d'échantillon sont faibles. Par conséquent, en pratique, on introduit dans le modèle un estimateur lissé de σ_i^2 que l'on traite alors comme étant connue. La fonction de variance généralisée est habituellement appliquée en pratique pour obtenir un estimateur lissé de la variance d'échantillonnage (par exemple, Dick 1995). Ces dernières

années, une méthode de lissage des effets de plan a été élaborée et utilisée en pratique pour obtenir des estimateurs de variance lissés (par exemple, Singh, Folsom et Vaish 2005 ; You 2008a ; Liu, Lahiri et Kalton 2008). En particulier, You (2008a) a appliqué une approche de modélisation à effets de plan égaux pour obtenir des estimations lisses des variances d'échantillonnage. L'effet de plan pour le i^{e} domaine peut s'écrire approximativement sous la forme

$$\text{deff}_i = \frac{s_i^2}{s_{ri}^2}, \text{ pour } i = 1, ..., m,$$

où s_i^2 est l'estimation directe sans biais de la variance d'échantillonnage fondée sur le plan d'échantillonnage complexe, et s_{ri}^2 est l'estimation de la variance d'échantillonnage sous un plan d'échantillonnage aléatoire simple. Pour chaque domaine, en émettant l'hypothèse d'un effet de plan commun, un facteur deff lissé peut être obtenu en appliquant $\text{deff} = \sum_{i=1}^{m} \text{deff}_i/m$. Ensuite, une estimation lissée de la variance d'échantillonnage $\tilde{\sigma}_i^2$ peut être obtenue sous la forme $\tilde{\sigma}_i^2 = s_{ri}^2 \cdot \text{deff}$.

Au lieu d'introduire les estimations lissées des variances d'échantillonnage dans le modèle, nous pouvons modéliser directement ces variances. Dans Wang et Fuller (2003) et You et Chapman (2006), ces auteurs supposent que la variance d'échantillonnage σ_i^2 est inconnue et estiment σ_i^2 au moyen d'un estimateur direct sans biais s_i^2, qui est indépendant de l'estimateur direct sous le plan y_i. Ils supposent aussi que $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$, où $d_i = n_i - 1$ et n_i est la taille d'échantillon pour le i^{e} domaine. You et Chapman (2006) ont examiné l'approche HB complète avec la méthode d'échantillonnage de Gibbs qui tient compte automatiquement de l'incertitude supplémentaire associée à l'estimation de σ_i^2. Dans le présent article, nous considérons à la fois les approches de lissage et de modélisation pour les variances d'échantillonnage.

2.2 Modèles spatiaux

En vue d'intégrer les effets aléatoires spatialement corrélés dans le modèle de lien, un moyen simple et évident consiste à ajouter un effet aléatoire spatial u_i dans le modèle de lien indépendant (1) comme il suit :

$$\theta_i = \mathbf{x}_i' \beta + v_i + u_i, \qquad (4)$$

où les u_i suivent le modèle conditionnel autorégressif (CAR) intrinsèque bien connu donné par

$$u_i \mid u_{-i} \sim \text{N}\left(\frac{\sum_{j \neq i} w_{ij} u_j}{\sum_{j \neq i} w_{ij}}, \frac{\sigma_u^2}{\sum_{j \neq i} w_{ij}} \right), \qquad (5)$$

où u_{-i} désigne les valeurs des effets aléatoires spatiaux u_j dans tous les autres domaines avec $j \neq i$, les poids w_{ij} sont des constantes fixées et σ_u^2 est une composante de variance inconnue. En pratique, un choix fréquent pour w_{ij} consiste à poser que $w_{ij} = 0$ à moins que les domaines i et j soient voisins (c'est-à-dire aient une limite commune), auquel cas $w_{ij} = 1$. Le modèle (4) est proposé par Besag, York et Mollie (1991) pour séparer les effets spatiaux de l'hétérogénéité globale dans les domaines. Dans le modèle (4), les effets aléatoires indépendants v_i traduisent l'hétérogénéité géographiquement non structurée entre les domaines, et les effets aléatoires spatiaux u_i traduisent la dépendance spatiale entre les domaines. De cette façon, le degré de dépendance spatiale globale peut être exprimé en se basant sur la proportion de la variation totale dans les $v_i + u_i$ reflétées par chaque composante.

En pratique, le choix entre un modèle non structuré (par exemple, le modèle de lien élémentaire) donné par (1) et un modèle purement structuré spatialement (par exemple, le modèle autorégressif intrinsèque) donné par (5) n'est souvent pas clair. Pour le modèle (4), l'inférence *a posteriori* au sujet de la dépendance spatiale est fondée sur la proportion de la variation totale de la somme de $v_i + u_i$ reflétée par chaque composante. Cependant, bien que les lois conditionnelles univariées de la composante spatiale (5) soient bien définies, la loi conjointe correspondante est impropre (de moyenne non définie et de variance infinie). De surcroît, le modèle (4) pose un problème éventuel d'identifiabilité où seule la somme des effets aléatoires $v_i + u_i$ est bien définie par les données ; voir, par exemple, Best et coll. (2005), pour une discussion plus détaillée.

Nous pouvons également considérer une autre paramétrisation spatiale étudiée par Leroux, Lei, et Breslow (1999) et par MacNab (2003), qui permet d'éviter le problème d'identifiabilité qui se pose avec le modèle (4). Soit $\theta_i = \mathbf{x}_i' \beta + b_i$, et $\mathbf{b} = (b_1, ..., b_m)'$. À l'instar de Leroux et coll. (1999) et de MacNab (2003), nous appliquons le modèle conditionnel autorégressif (CAR) qui suit aux effets spatiaux de domaine $\mathbf{b} = (b_1, ..., b_m)'$:

$$\mathbf{b} \sim \text{MVN}(\mathbf{0}, \Sigma(\sigma_b^2, \lambda)) \qquad (6)$$

$$\Sigma(\sigma_b^2, \lambda) = \sigma_b^2 \mathbf{D}^{-1}, \ \mathbf{D} = \lambda \mathbf{R} + (1 - \lambda)\mathbf{I} \qquad (7)$$

où σ_b^2 est un paramètre de dispersion spatiale et λ est un paramètre d'autocorrélation spatiale, $0 \leq \lambda \leq 1$; $\mathbf{I}$ est une matrice identité de dimension m ; $\mathbf{R}$ est une matrice, habituellement appelée matrice de voisinage (*neighbourhood matrix*), dont le i^{e} élément diagonal est égal au nombre de voisins du domaine i, et dont les éléments hors diagonale dans chaque ligne sont égaux à -1 si les domaines correspondants sont voisins et à 0 autrement. Le modèle CAR donné par les expressions (6) et (7) correspond à la loi conditionnelle de b_i suivante :

$$b_i \mid b_{-i} \sim \text{N}\left(\frac{\lambda}{1 - \lambda + \lambda w_{i+}} \sum_{j \neq i} w_{ij} v_j, \frac{\sigma_b^2}{1 - \lambda + \lambda w_{i+}} \right),$$

où $w_{i+} = \sum_{j \neq i} w_{ij}$. Le modèle CAR (6) - (7) devient le modèle autorégressif intrinsèque (5) si $\lambda = 1$. Par ailleurs, si $\lambda = 0$, il se réduit au modèle de lien indépendant (1) dans lequel les effets aléatoires propres au domaine v_i sont supposés indépendants. Il convient de souligner que la moyenne et les variances conditionnelles de $b_i \mid b_{-i}$ sont les sommes pondérées des moments lissés globaux provenant du modèle de lien élémentaire (1) et des moments lissés locaux provenant du modèle autorégressif intrinsèque :

$$\mathrm{E}(b_i \mid b_{-i}) = \frac{1-\lambda}{1-\lambda+\lambda w_{i+}} \times 0 + \frac{\lambda w_{i+}}{1-\lambda+\lambda w_{i+}} \left(\sum_{j \neq} w_{ij} b_j / w_{i+} \right)$$

$$\mathrm{Var}(b_i \mid b_{-i}) = \frac{1-\lambda}{1-\lambda+\lambda w_{i+}} \times \sigma_b^2 + \frac{\lambda w_{i+}}{1-\lambda+\lambda w_{i+}} (\sigma_b^2 / w_{i+}).$$

Donc, le modèle (6) - (7) représente un équilibre entre le modèle de lien indépendant (1) et le modèle CAR intrinsèque (5). Le paramètre de corrélation spatiale λ mesure l'importance des effets spatiaux pour le lissage local des domaines voisins. La structure de modélisation (6) traduit à la fois l'hétérogénéité non structurée entre les domaines et les effets de corrélation spatiale du domaine voisin.

2.3 Modèles hiérarchiques bayésiens et inférence

Afin d'estimer θ_i, le paramètre d'intérêt, nous appliquons une approche hiérarchique bayésienne (HB) en utilisant la méthode d'échantillonnage de Gibbs. Comparativement à d'autres approches, telles que l'EBLUP et l'approche empirique bayésienne (EB), l'approche HB est simple et les inférences concernant θ_i sont exactes contrairement aux inférences EB ou EBLUP. En outre, l'approche HB donne le moyen de traiter des modèles pour petits domaines complexes en utilisant la méthode Monte Carlo par chaîne de Markov (MCMC), qui permet de surmonter en grande partie les difficultés que posent le calcul des intégrales multidimensionnelles des quantités *a posteriori*.

Soit $\mathbf{y} = (y_1, ..., y_m)'$, $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)'$, et $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_m)'$. Nous commençons par construire deux modèles HB sans et avec structure spatiale sous l'hypothèse que les variances d'échantillonnage sont connues et remplacées par l'estimation lissée σ_i^2.

Modèle 1 : Modèle de Fay-Herriot, désigné par MFH (Fay et Herriot 1979 ; Rao 2003).

- $y_i \mid \theta_i \sim \mathrm{N}(\theta_i, \sigma_i^2 = \tilde{\sigma}_i^2)$, pour $i = 1, ..., m$;
- $\theta_i \mid \beta, \sigma_v^2 \sim \mathrm{N}(\mathbf{x}_i'\beta, \sigma_v^2)$, pour $i = 1, ..., m$;
- Priors pour les paramètres $(\beta, \sigma_v^2) : \pi(\beta) \propto 1$; $\pi(\sigma_v^2) \sim \mathrm{GI}(a_0, b_0)$, où a_0, b_0 sont des constantes connues choisies très petites pour refléter les connaissances vagues au sujet de σ_v^2. N désigne la loi normale et GI, la loi gamma inverse.

Modèle 2 : Modèle CAR au niveau du domaine proposé, en tant qu'extension du modèle de Fay-Herriot, désigné par CAR-MFH.

- $\mathbf{y} \mid \boldsymbol{\theta} \sim \mathrm{MVN}(\boldsymbol{\theta}, \mathbf{E})$, où $\mathbf{E}$ est une matrice diagonale dont le i^{e} élément diagonal est $\sigma_i^2 = \tilde{\sigma}_i^2$;
- $\boldsymbol{\theta} \mid \beta, \sigma_v^2 \sim \mathrm{MVN}(\mathbf{X}\beta, \sigma_v^2 \mathbf{D}^{-1})$, où $\mathbf{D} = \lambda \mathbf{R} + (1-\lambda)\mathbf{I}$, avec $\mathbf{I}$ désignant une matrice identité de dimension m et $\mathbf{R}$, la matrice de voisinage ;
- Priors pour les paramètres $(\beta, \lambda, \sigma_v^2) : \pi(\beta) \propto 1$; $\pi(\lambda) \sim \mathrm{Uniform}(0,1)$, où $0 \leq \lambda \leq 1$; $\pi(\sigma_v^2) \sim \mathrm{GI}(a_0, b_0)$, où a_0, b_0 sont des constantes connues choisies très petites. MVN désigne la loi normale multivariée.

Il convient de souligner que le modèle CAR-MFH proposé se réduit au modèle MFH lorsque le paramètre d'autocorrélation spatial est $\lambda = 0$.

Nous considérons aussi deux modèles HB dont la variance d'échantillonnage σ_i^2 est inconnue et modélisée par l'estimateur sans biais direct s_i^2.

Modèle 3 : Modèle de You-Chapman désigné par MYC (You et Chapman 2006).

- $y_i \mid \theta_i, \sigma_i^2 \sim \mathrm{N}(\theta_i, \sigma_i^2)$, pour $i = 1, ..., m$;
- $d_i s_i^2 \mid \sigma_i^2 \overset{\text{ind}}{\sim} \sigma_i^2 \chi_{d_i}^2$ où $d_i = n_i - 1$, pour $i = 1, ..., m$;
- $\theta_i \mid \beta, \sigma_v^2 \sim \mathrm{N}(\mathbf{x}_i'\beta, \sigma_v^2)$, pour $i = 1, ..., m$;
- Priors pour les paramètres inconnus $(\beta, \sigma_v^2, \sigma_i^2, i = 1, ..., m) : \pi(\beta) \propto 1$; $\pi(\sigma_v^2) \sim \mathrm{GI}(a_0, b_0)$, $\pi(\sigma_i^2) \sim \mathrm{GI}(a_i, b_i)$ pour $i = 1, ..., m$, où a_i, b_i $(0 \leq i \leq m)$ sont des constantes connues choisies très petites pour refléter les connaissances vagues au sujet de σ_i^2 et σ_v^2.

Modèle 4 : Modèle CAR au niveau du domaine proposé avec variances d'échantillonnage inconnues, en tant qu'extension du modèle de You-Chapman, désigné par CAR-MYC.

- $\mathbf{y} \mid \boldsymbol{\theta}, \sigma_1^2, ..., \sigma_m^2 \sim \mathrm{MVN}(\boldsymbol{\theta}, \mathbf{E})$, où la matrice $\mathbf{E}$ contient les éléments diagonaux σ_i^2 ;
- $d_i s_i^2 \mid \sigma_i^2 \overset{\text{ind}}{\sim} \sigma_i^2 \chi_{d_i}^2$, où $d_i = n_i - 1$, pour $i = 1, ..., m$;
- $\boldsymbol{\theta} \mid \beta, \sigma_v^2 \sim \mathrm{MVN}(\mathbf{X}\beta, \sigma_v^2 \mathbf{D}^{-1})$, où $\mathbf{D} = \lambda \mathbf{R} + (1-\lambda)\mathbf{I}$;
- priors pour les paramètres $(\beta, \lambda, \sigma_v^2, \sigma_i^2, i = 1, ..., m)$: $\pi(\beta) \propto 1$; $\pi(\lambda) \sim \mathrm{Uniform}(0, 1)$, où $0 \leq \lambda \leq 1$; $\pi(\sigma_v^2) \sim \mathrm{GI}(a_0, b_0)$; $\pi(\sigma_i^2) \sim \mathrm{GI}(a_i, b_i)$ pour $i = 1, ..., m$, où a_i, b_i $(0 \leq i \leq m)$ sont des constantes connues choisies très petites.

De nouveau, soulignons que le modèle proposé CAR-MYC se réduit au modèle de You-Chapman quand $\lambda = 0$. Le modèle 3 ainsi que le modèle 4 comportent l'hypothèse implicite que la taille d'échantillon de domaine $n_i \geq 2$. Si des priors plats (*flat*) sont utilisés pour σ_i^2, il faudrait que $n_i \geq 4$ pour être certains que les lois *a posteriori* soient appropriées (You et Chapman 2006).

Nous appliquons la méthode d'échantillonnage de Gibbs pour estimer la moyenne *a posteriori* $\mathrm{E}(\theta_i \mid \mathbf{y})$ et la variance *a posteriori* correspondante $\mathrm{Var}(\theta_i \mid \mathbf{y})$. Les lois conditionnelles complètes des paramètres requises sous les divers modèles sont données à l'annexe A. Pour les modèles de Fay-Herriot et de You-Chapman, toutes les lois conditionnelles complètes possèdent des formes explicites et le tirage d'échantillons à partir de ces lois est simple. Pour les modèles spatiaux au niveau du domaine proposés CAR-MFH et CAR-MYC, la loi conditionnelle du paramètre de corrélation spatiale λ ne possède pas de forme explicite. Nous utilisons l'algorithme de Metropolis-Hastings avec l'échantillonneur de Gibbs (Chip et Greenberg 1995) pour mettre à jour λ. Sous le modèle CAR-MFH, la loi conditionnelle complète de λ dans l'échantillonneur de Gibbs peut s'écrire sous la forme

$$[\lambda \mid \boldsymbol{\theta}, \beta, \sigma_v^2] \propto \mathrm{h}(\lambda)\,\mathrm{f}(\lambda)$$

où $\mathrm{f}(\lambda)$ est une densité de probabilité de la loi uniforme Uniform (0, 1) donnée par

$$\mathrm{f}(\lambda) \propto 1, \text{ où } 0 \leq \lambda \leq 1$$

et $\mathrm{h}(\lambda)$ est une fonction donnée par

$$\mathrm{h}(\lambda) \propto \left|\left[\lambda \mathbf{R} + (1-\lambda)\mathbf{I}\right]^{-1}\right|^{-1/2} \times \exp\left\{-\frac{1}{2\sigma_v^2}(\boldsymbol{\theta} - \mathbf{X}\beta)'\left[\lambda \mathbf{R} + (1-\lambda)\mathbf{I}\right](\boldsymbol{\theta} - \mathbf{X}\beta)\right\}.$$

Nous utilisons $\mathrm{f}(\lambda)$ comme densité de probabilité génératrice « candidate » dans l'étape de mise à jour de l'algorithme de Metropolis-Hastings. Pour mettre λ à jour à partir des valeurs courantes de $(\boldsymbol{\theta}^{(k)}, \beta^{(k)}, \sigma_v^{2(k)})$, la procédure est la suivante :

1. tirer λ^* d'une loi uniforme ;
2. calculer la probabilité d'acceptation $\alpha(\lambda^*, \lambda^{(k)}) = \min\{h(\lambda^*)/h(\lambda^{(k)}), 1\}$;
3. générer u à partir d'une loi uniforme ; si $u < \alpha(\lambda^*, \lambda^{(k)})$, la valeur candidate λ^* est acceptée, c'est-à-dire $\lambda^{(k+1)} = \lambda^*$; sinon, λ^* est rejetée, et fixée à $\lambda^{(k+1)} = \lambda^{(k)}$.

Pour le modèle CAR-MYC, une procédure similaire peut être appliquée au moment du tirage des échantillons à partir de la loi conditionnelle de λ.

3. Analyse des données

3.1 Description des données et mise en œuvre

L'Enquête sur la santé dans les collectivités canadiennes (ESCC) est une enquête fédérale réalisée par Statistique Canada. L'objectif principal de l'ESCC est de fournir des estimations à jour et fiables des déterminants de la santé, de l'état de santé et de l'utilisation du système de santé au Canada. Il s'agit d'une enquête transversale dont le cycle de collecte est de deux ans. La première année de l'enquête, le cycle « x.1 » a pour cible les membres de 12 ans et plus de la population à domicile et est une enquête générale sur la santé de la population réalisée auprès d'un grand échantillon (130 000 personnes) conçu en vue de fournir des estimations fiables au niveau des régions sociosanitaires, des provinces et du Canada. La deuxième année de l'enquête, le cycle « x.2 » est réalisé auprès d'un échantillon plus petit (30 000 personnes), réparti en fonction des achats d'unités d'échantillonnage supplémentaires des provinces et conçu pour fournir des résultats au niveau provincial et national sur des thèmes particuliers liés à la santé. Bien que les estimations nationales et provinciales soient très importantes, la demande de données sur la santé à des niveaux plus fins d'agrégation géographique augmente dans un certain nombre de provinces, dont la Colombie-Britannique (C.-B.), l'Île-du-Prince-Édouard (Î.-P.-É.), le Québec et d'autres. Le cycle « x.1 » de l'ESCC permet de recueillir des données pour 136 régions sociosanitaires réparties dans les dix provinces et les trois territoires. Il est réalisé essentiellement à l'aide de deux bases de sondage. La première, qui est la base principale, est fondée sur la base de sondage aréolaire conçue pour l'Enquête sur la population active du Canada. Cette base aréolaire est utilisée pour échantillonner les logements conformément à un plan d'échantillonnage en grappes stratifiées à plusieurs degrés. La deuxième base de sondage consiste en une liste de numéros de téléphone. La méthode de composition aléatoire est utilisée dans certaines régions sociosanitaires pour des raisons de coût. Des renseignements plus détaillés sur le plan de sondage sont fournis dans Béland (2002). Dans le présent article, nous utilisons un petit ensemble de données provenant du cycle 1.1 pour illustrer l'analyse. Nous voulons estimer le taux de prévalence de la maladie dans les régions sociosanitaires à l'intérieur des provinces. En particulier, nous appliquons les quatre modèles décrits à la section 2 pour estimer le taux de prévalence de l'asthme dans 20 régions sociosanitaires dans la province de la Colombie-Britannique en utilisant les données provenant du cycle 1.1. La figure 1 montre la carte des 20 régions sociosanitaires de la Colombie-Britannique. Nous utilisons cette carte pour définir la matrice de corrélation de voisinage utilisée dans les modèles spatiaux. L'annexe B donne la liste des régions sociosanitaires et les structures spatiales connexes.

Soit θ_i le taux réel de prévalence de l'asthme dans la i^e région sociosanitaire en C.-B., $i = 1, ..., 20$. D'après les données du cycle 1.1 de l'enquête, nous obtenons l'estimation directe y_i de θ_i comme étant le ratio du nombre de personnes asthmatiques (estimation directe) divisé par la taille de population correspondante (constante connue). Nous avons également inclus six variables auxiliaires au niveau du domaine utilisées dans le modèle, à savoir la taille totale de population, le nombre de personnes dont l'asthme est l'un des symptômes de maladie chronique, le nombre de personnes dont l'asthme est le symptôme principal de maladie chronique, le nombre de personnes dont le diabète est l'un des symptômes de maladie chronique, le nombre de personnes dont le diabète est le symptôme principal de maladie chronique, et le nombre de visites à l'hôpital. Notons que, dans la littérature traitant de la cartographique des maladies (par exemple, Mollié 1996 ; Maiti 1998 ; MacNab 2003), la loi de Poisson ou la loi binomiale est habituellement celle qui est supposée dans le modèle d'échantillonnage pour l'estimation directe y_i. Cependant, dans l'estimation sur petits domaines, l'estimation directe y_i s'obtient en se fondant sur le plan d'échantillonnage complexe utilisé dans l'enquête. Donc, on émet habituellement l'hypothèse d'un modèle d'échantillonnage normal pour les estimations directes y_i ; voir, par exemple, Datta, Lahiri, Maiti et Lu (1999), Rao (2003), Mohadjer, Rao, Liu, Krenzke et Van de Kerckhove (2007), et You (2008a). Il convient de souligner que nous n'avons pris en considération qu'un seul type de données sur la prévalence de la maladie provenant d'une seule province dans notre étude, et utilisé cet exemple pour illustrer le modèle proposé et évaluer les effets de la modélisation spatiale dans les modèles pour petits domaines.

Pour appliquer l'échantillonnage de Gibbs, nous utilisons $L = 5$ exécutions parallèles ayant chacune une longueur de « rodage » de $B = 2\,000$ et une taille d'échantillonnage de Gibbs de $G = 5\,000$. Pour les modèles CAR-MFH et CAR-MYC proposés, afin de réduire l'autocorrélation qui résulte de l'algorithme d'acceptation–rejet dans l'exécution, nous prenons une itération sur cinq après la période de « rodage ». Par conséquent, pour les modèles MFH et MYC, nous avons $n = 5\,000$ échantillons pour chaque exécution, et pour les modèles CAR-MFH et CAR-MYC, nous avons $n = 1\,000$ échantillons pour chaque exécution. Nous surveillons la convergence de l'échantillonnage de Gibbs pour le paramètre de petits domaines θ_i et d'autres paramètres inconnus dans le modèle en utilisant le facteur de réduction potentiel d'échelle (*potential scale reduction factor*) (Gelman et Rubin 1992 ; Gelman, Carlin, Stern et Rubin 2004, pages 296-297). Nous avons calculé les facteurs de réduction pour tous les paramètres surveillés dans le modèle dans l'échantillonnage de Gibbs. Les valeurs du facteur sont toutes proches de 1 (inférieures à 1,05), ce qui donne à penser que la convergence souhaitée pour ces paramètres est réalisée par l'échantillonneur de Gibbs.

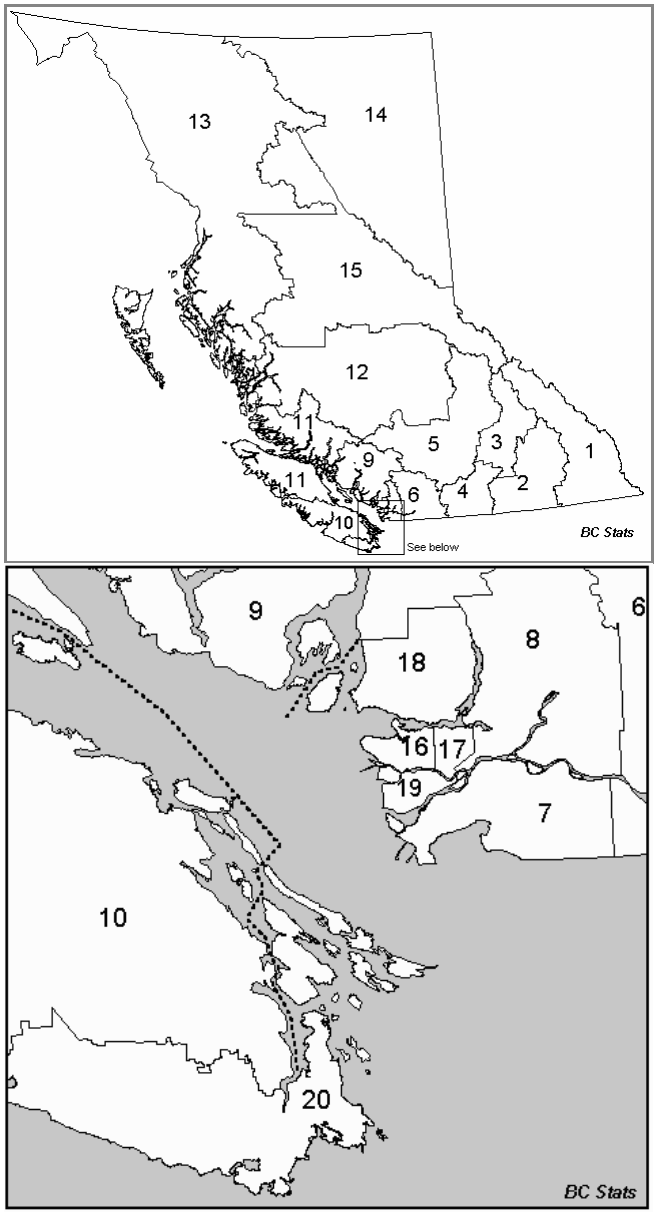

Figure 1 Carte des 20 régions sociosanitaires de la Colombie-Britannique

Nous avons utilisé des priors vagues pour les hyperparamètres du modèle comme il est d'usage en pratique pour l'estimation sur petits domaines de type HB. En particulier, il est fréquent d'utiliser le prior plat pour le paramètre de régression $\pi(\beta) \propto 1$ et les priors gamma inverses appropriés pour les composantes de la variance (par exemple, Arora et Lahiri 1997 ; Ghosh et coll. 1998 ; Datta et coll. 1999 ; You et Rao 2000 ; Rao 2003, page 237 ; Souza et coll. 2009). À l'exemple de MacNab (2003), nous avons

utilisé le prior uniforme $\pi(\lambda) \sim$ pour le paramètre d'autocorrélation. Les priors uniformes sont aussi utilisés fréquemment pour les paramètres d'autocorrélation dans les modèles spatiaux (par exemple, Maiti 1998 ; He et Sun 2000 ; Rao 2003, page 266). Nous avons également essayé plusieurs valeurs différentes pour les priors gamma inverses. Les estimations HB sont assez stables et non sensibles au choix des priors vagues appropriés. Une discussion plus détaillée de l'analyse de sensibilité peut être consultée, par exemple, dans You et Chapman (2006) pour des modèles similaires.

3.2 Comparaison des résultats

Pour commencer, nous présentons les estimations HB du taux de prévalence de l'asthme sous les modèles MFH et CAR-MFH dans lesquels les variances d'échantillonnage σ_i^2 sont supposées connues. Nous avons utilisé l'estimation lissée $\tilde{\sigma}_i^2$ obtenue par la technique de lissage exposée dans You (2008a) comme il est décrit à la section 2. La figure 2 donne les estimations directes et les estimations fondées sur le modèle HB sous les modèles MFH et CAR-MFH pour les 20 régions sociosanitaires de la Colombie-Britannique. Les régions sociosanitaires sont représentées sur l'axe des x, classées par ordre de taille d'échantillon, de la plus petite (Peace Liard) à gauche à la plus grande (South Fraser Valley) à droite. Le modèle 1 (MFH) et le modèle 2 (CAR-MFH) donnent des estimations ponctuelles similaires et les deux estimations fondées sur un modèle mènent à des estimations modérément lisses comparativement aux estimations directes. En outre, les estimations directes et les deux estimations HB du taux de prévalence de la maladie sont très proches pour certaines régions sociosanitaires dont la taille d'échantillon est grande, mais diffèrent dans une certaine mesure pour certaines régions dont la taille d'échantillon est plus petite. Des résultats similaires sont obtenus sous le modèle 3 (MCY) et le modèle 4 (CAR-MCY).

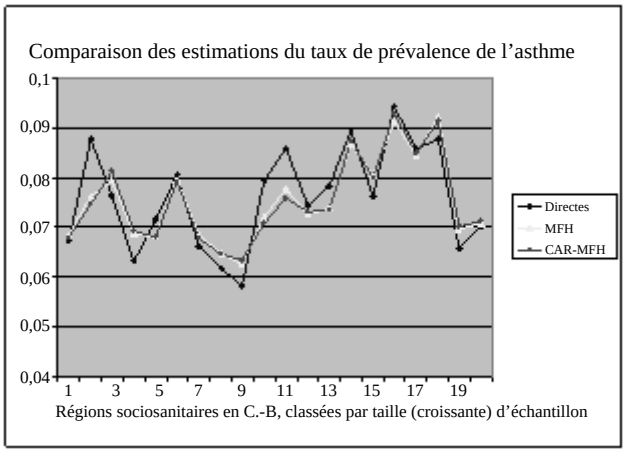

Figure 2 Estimations directes et fondées sur un modèle HB sous les modèles MFH et CAR-MFH

La figure 3 donne les CV des estimations directes et des deux estimations fondées sur un modèle HB pour les régions sociosanitaires classées selon la taille d'échantillon, de la plus petite à la plus grande, comme à la figure 2. Les CV des estimations HB sont obtenus en divisant la racine carrée de la variance *a posteriori* par la moyenne *a posteriori*. Comme prévu, les CV des estimations directes ont nettement tendance à diminuer à mesure que la taille d'échantillon augmente. Par contre, les deux estimations fondées sur un modèle HB donnent des CV plus lisses. En outre, ces deux estimations offrent une amélioration importante par rapport aux estimations directes fondées sur le plan en ce qui a trait à la précision et à la fiabilité, c'est-à-dire la réduction des CV. Comparativement aux estimations directes, la réduction moyenne du CV des estimations HB sous le modèle MFH est de l'ordre de 22,7 %, variant de 7,8 % à 40,5 %, et la réduction moyenne du CV des estimations HB sous le modèle CAR-MFH proposé est de 27,8 %, variant de 12,5 % à 52,1 %. Donc, il est clair que le modèle spatial proposé CAR-MFH est supérieur au modèle de Fay-Herriot. Nous avons également obtenu des résultats similaires pour les modèles MYC et CAR-MYC quand la variance d'échantillonnage est modélisée directement. La réduction moyenne du CV sous MYC est de 23,9 %, tandis qu'elle est de 29,0 % sous le modèle spatial proposé CAR-MYC. Des résultats détaillés, y compris les estimations ponctuelles et les CV correspondants, sont présentés dans un tableau à l'annexe C. Dans notre exemple, la taille de l'échantillon au niveau de la région sociosanitaire est relativement grande. Néanmoins, les estimations fondées sur un modèle révèlent une amélioration importante par rapport aux estimations directes fondées sur le plan de sondage. Nos résultats indiquent que les modèles pour petits domaines proposés peuvent être utilisés afin d'améliorer les estimations directes, même quand la taille d'échantillon est relativement grande. Il convient de souligner que les intervalles de crédibilité bayésiens pour les paramètres de petit domaine peuvent être construits facilement en utilisant la sortie MCMC de l'échantillonneur de Gibbs, si cela est nécessaire en pratique. Il s'agit de l'un des avantages de l'utilisation de l'inférence HB par la voie de l'échantillonnage MCMC. Cependant, ici, nous ne présentons que les estimations ponctuelles fondées sur un modèle et les CV correspondants, car notre objectif principal est de comparer les estimations fondées sur un modèle aux estimations directes et de montrer les gains d'efficacité des modèles. Le gain d'efficacité est clairement appréciable.

Afin d'étudier les effets de l'intégration de la structure spatiale dans le modèle, à la figure 4, nous présentons les CV des estimations directes et HB selon les régions sociosanitaires classées en fonction du nombre de régions voisines, en allant du plus petit (deux voisines) au plus grand

(sept voisines). La figure montre que les estimations HB d'après le modèle CAR-MFH proposé ont un CV plus petit que celles produites au moyen du modèle de Fay-Herriot. En outre, l'amélioration que représente le modèle CAR-MFH par rapport au modèle de Fay-Herriot est nettement plus marquée dans les régions ayant un grand nombre de voisines, alors que les deux modèles donnent des CV très proches dans les régions dont le nombre de régions adjacentes est plus petit. Nous obtenons des résultats très semblables pour le modèle CAR-MYC comparativement au modèle MYC. Le tableau 1 donne la réduction moyenne des CV pour les régions sociosanitaires ayant le même nombre de voisines. Les résultats du tableau 1 représentent la réduction du CV des modèles spatiaux proposés quand la variance d'échantillonnage est connue ou inconnue. Par exemple, pour σ_i^2 connue ($\tilde{\sigma}_i^2$ lissée), pour les régions ne comptant que deux voisines, la réduction moyenne du CV sous le modèle CAR-MFH par rapport au modèle de Fay-Herriot n'est que de 0,9 % environ, tandis que pour les régions comptant sept voisines, la réduction moyenne du CV sous le modèle CAR-MFH par rapport au modèle MFH peut aller jusqu'à environ 20 %. Pour le cas de σ_i^2 inconnue, nous obtenons des résultats similaires pour CAR-MYC par rapport MYC. Les résultats numériques du tableau 1 confirment la tendance nette à une plus grande réduction des CV sous le modèle spatial proposé que sous le modèle MFH ou MYC à mesure que le nombre de régions voisines augmente. Donc, un plus grand nombre de régions voisines peuvent fournir plus de renseignements sur la structure spatiale en vue d'améliorer la précision et la fiabilité des estimations HB.

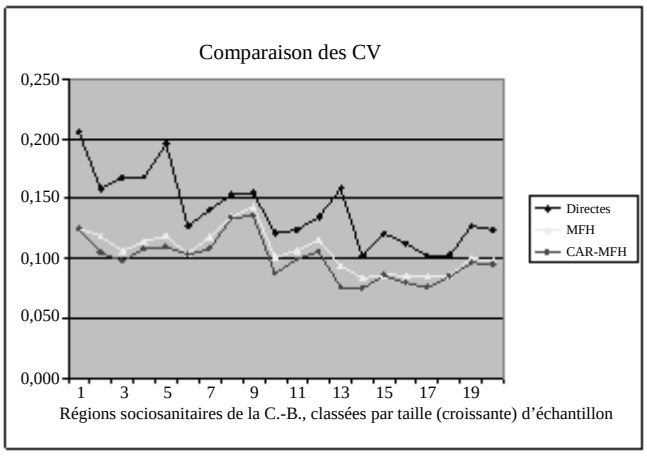

Figure 3 CV des estimations directes et HB sous les modèles MFH et CAR-MFH

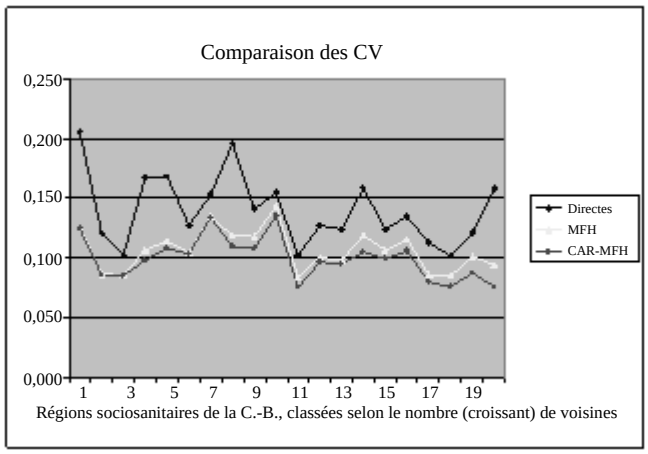

Figure 4 CV des estimations directes et HB sous les modèles MFH et CAR-MFH avec les régions sociosanitaires classées selon le nombre de régions voisines

Tableau 1
Comparaison de la réduction moyenne du CV

Nombre de régions voisines	Réduction moyenne du CV	
	CAR-MFH par rapport à MFH	CAR-MYC par rapport à MYC
2	0,9 %	1.8 %
3	3,7 %	3.5 %
4	6,3 %	6.0 %
5	8,9 %	8.7 %
6	13,7 %	11,0 %
7	19,2 %	20,7 %

3.3 Comparaison des modèles bayésiens

À la présente section, nous comparons les modèles proposés CAR-MFH et CAR-MYC aux modèles MFH et MYC, respectivement. Pour la comparaison de modèles hiérarchiques bayésiens, on peut se servir du critère d'information de déviance (DIC pour *Deviance Information Criterion*) proposé par Spiegelhalter, Best, Carlin et van der Linde (2002). Ce critère a été utilisé fréquemment ces dernières années pour comparer les modèles bayésiens à effets non emboîtés et mixtes. Le DIC est basé sur la déviance du modèle $D(\theta)$, qui est égale à moins deux fois la log-vraisemblance du modèle, et il est habituellement calculé sous la forme $\text{DIC} = D(\hat{\theta}) + 2p_D$, où $D(\hat{\theta})$ est la déviance du modèle, évaluée à la moyenne *a posteriori* des paramètres du modèle, qui résume la qualité de l'ajustement du modèle, et p_D est le nombre effectif de paramètres, qui traduit la complexité du modèle. p_D est défini comme étant $p_D = \bar{D}(\theta) - D(\hat{\theta})$, où $\bar{D}(\theta)$ est la moyenne *a posteriori* de la déviance du modèle. Donc, le DIC est défini comme la somme de l'adéquation du modèle et de la complexité du modèle. L'ajustement du modèle est d'autant meilleur que la valeur du DIC est faible. Le calcul du DIC est relativement

simple à condition qu'il existe une forme explicite pour la déviance, et p_D peut être calculé après l'exécution de l'échantillonnage de Gibbs en prenant la moyenne d'échantillon des valeurs simulées de $D(\theta)$ dont on soustrait l'estimation de la déviance $D(\hat{\theta})$ obtenue par insertion de valeurs (*plug-in*). Pour les quatre modèles présentés à la section 2, nous avons calculé les valeurs du DIC correspondantes, qui sont présentées au tableau 2. Il est clair que les modèles spatiaux proposés CAR-MFH et CAR-MYC ont tous deux un DIC plus petit que les modèles non spatiaux MFH et MYC, respectivement, ce qui signifie que les modèles spatiaux sont meilleurs que les modèles non spatiaux dans notre étude. Les deux modèles spatiaux CAR-MFH et CAR-MYC donnent de bons résultats dans cet exemple. Les résultats de la comparaison des modèles corroborent les résultats d'estimation présentés à la section 3.2.

Tableau 2
Comparaison des valeurs du DIC pour les quatre modèles hiérarchiques

Modèle	Valeur du DIC
MFH	27,1
CAR-MFH	24,6
MYC	26,8
CAR-MYC	24,5

3.4 Test d'adéquatoin du modèle

Afin de vérifier l'adéquation globale des modèles proposés CAR-MFH et CAR-MYC, nous avons utilisé la méthode de la loi prédictive *a posteriori*. Soit y_{rep} l'observation répétée sous le modèle. La loi prédictive *a posteriori* de y_{rep} sachant les données observée y_{obs} est définie comme étant $f(y_{\text{rep}}|y_{\text{obs}}) = \int f(y_{\text{rep}}|\theta)\, f(\theta|y_{\text{obs}})\, d\theta$. Dans cette approche, nous pouvons définir une statistique de test $T(y, \theta)$ qui dépend des données *y* et éventuellement du paramètre θ, et comparer la valeur observée $T(y_{\text{obs}}, \theta \mid y_{\text{obs}})$ à la loi prédictive *a posteriori* de $T(y_{\text{rep}}, \theta \mid y_{\text{obs}})$, tout écart significatif indiquant l'échec du modèle. Le manque d'ajustement aux données par rapport à la loi prédictive *a posteriori* peut être mesuré par la valeur *p* de la statistique de test (Meng 1994 ; Gelman, Meng et Stern 1996). La valeur *p* prédictive *a posteriori* est définie comme étant $p = P(T(y_{\text{rep}}, \theta) \geq T(y_{\text{obs}}, \theta) \mid y_{\text{obs}})$. Si le modèle donné est bien ajusté aux données observées, $T(y_{\text{obs}}, \theta \mid y_{\text{obs}})$ doit être proche de la partie centrale de l'histogramme des valeurs de $T(y_{\text{rep}}, \theta \mid y_{\text{obs}})$ si y_{rep} est générée de manière répétée à partir de la loi prédictive *a posteriori*. Par conséquent, la valeur *p* prédictive *a posteriori* devrait, en principe, être proche de 0,5 si le modèle est adéquatement ajusté aux données. Des valeurs *p* extrêmes (proche de 0 ou de 1) suggèrent un mauvais ajustement. La vérification des modèles au moyen de la valeur *p* prédictive *a posteriori* a été critiquée comme étant trop prudente, à cause de la double utilisation des données observées ; voir, par exemple, Bayarri et Berger (2000). Ces auteurs ont proposé des mesures de la valeur *p* de rechange pour la vérification des modèles, appelées valeur *p* prédictive *a posteriori* partielle et valeur *p* prédictive conditionnelle. Cependant, leurs méthodes sont plus difficiles à mettre en œuvre et à interpréter (Rao 2003 ; Sinharay et Stern 2003). Comme l'ont fait remarquer Sinharay et Stern (2003), la valeur *p* prédictive *a posteriori* est particulièrement utile si nous considérons le modèle courant comme un point final plausible auquel des modifications ne doivent être apportées que s'il s'avère que l'adéquation de l'ajustement laisse beaucoup à désirer.

Pour exécuter la vérification du modèle prédictif *a posteriori*, nous devons spécifier une statistique de test $T(y, \theta)$. You (2008b) a étudié par simulation plusieurs statistiques de test pour la vérification du modèle prédictif *a posteriori* dans le cas des modèles pour petits domaines et à proposé une statistique de test donnée par

$$T(y, \theta) = \left| \max(y_i) - \text{moyen.}(\theta_i) \right| - \left| \min(y_i) - \text{moyen.}(\theta_i) \right|.$$

You (2008b) montre que la statistique de test proposée $T(y, \theta)$ est sensible au choix de la distribution des effets aléatoires et de différentes fonctions de moyenne sous le modèle de Fay-Herriot. Une statistique de test similaire est également proposée dans Gelman et coll. (2004) pour la vérification du modèle prédictif *a posteriori*. Dans notre étude, sous le modèle proposé CAR-MFH, la valeur *p* estimée est de 0,472, et sous le modèle CAR-MYC, elle est de 0,453. Rien n'indique donc un manque d'ajustement du modèle et les deux modèles spatiaux proposés sont assez bien ajustés aux données.

Pour examiner l'ajustement du modèle au niveau d'observation individuel, nous avons également calculé les valeurs des probabilités prédictives individuelles p_i^* sous la forme $p_i^* = P(y_{i(\text{rep})} < y_{i(\text{obs})} \mid y_{\text{obs}})$; voir, par exemple, Gelfand (1996), ainsi que Daniels et Gatsonis (1999). Ces probabilités prédictives individuelles renseignent sur le degré de surestimation ou de sous-estimation systématique des données observées. Pour le modèle CAR-MFH, les p_i^* varient de 0,325 à 0,768 avec une moyenne de 0,517 et une médiane de 0,496 ; pour le modèle CAR-MYC, les p_i^* varient de 0,316 à 0,772 avec une moyenne de 0,511 et une médiane de 0,497. Les deux modèles donnent des résultats fort semblables, et les valeurs moyennes et médianes sont de l'ordre de 0,5. Il n'existe aucune indication d'une surestimation ou d'une sous-estimation systématique des modèles proposés. Les valeurs *p* globales et les probabilités prédictives individuelles montrent que les modèles spatiaux pour petits domaines proposés sont assez bien ajustés aux données.

3.5 Diagnostic du biais

Afin d'évaluer le biais éventuel des estimations fondées sur les modèles proposés par rapport aux estimations directes sous le plan de sondage, comme Brown, Chambers, Heady et Heasman (2001), nous appliquons une simple méthode d'analyse de régression aux estimations directes et aux estimations fondées sur un modèle HB. You (2008a) a également utilisé la méthode d'analyse de régression pour diagnostiquer le biais d'un modèle. Si les estimations fondées sur le modèle sont proches des valeurs réelles des taux de prévalence de la maladie dans les petits domaines, les estimations directes sous le plan de sondage, qui sont supposées fournir des estimations sans biais par rapport aux taux réel de prévalence de la maladie, devraient se comporter comme des variables aléatoires dont les valeurs prédites correspondent aux valeurs des estimations fondées sur un modèle. Autrement dit, les estimations fondées sur un modèle devraient être des prédicteurs sans biais des estimations directes. En ce qui concerne l'analyse de régression, nous ajustons essentiellement le modèle de régression $Y = \alpha + \beta X$ aux données et estimons les coefficients, et nous voyons dans quelle mesure la droite de régression s'approche de $Y = X$. Soit Y les estimations directes et X les estimations fondées sur le modèle. Sous le modèle proposé CAR-MFH, nous obtenons la droite de régression $Y = -0,0021(0,011) + 1,0365(0,1445)X$; sous le modèle proposé CAR-MYC, nous obtenons la droite de régression $Y = -0,0028(0,0108) + 1,0458(0,1427)X$. Donc, les deux droites de régression diffèrent fort peu de $Y = X$. Nous concluons par conséquent que les estimations fondées sur le modèle convergent vers les estimations directes sans biais supplémentaire éventuel induit par les modèles proposés. Les résultats donnent aussi une indication de l'absence de tout biais dû à une erreur éventuelle de spécification du modèle.

4. Conclusion

Dans le présent article, nous avons discuté de deux modèles au niveau du domaine, à savoir le modèle bien connu de Fay-Herriot dans lequel la variance d'échantillonnage est supposée être connue, et le modèle de You-Chapman dans lequel la variance d'échantillonnage est inconnue et modélisée séparément par son estimateur direct. Dans l'un et l'autre modèle, il est supposé que les effets aléatoires de domaine sont des variables aléatoires *iid* normales pour traduire les effets de l'hétérogénéité inexpliquée des domaines. Après avoir comparé diverses formes de modèles CAR gaussiens proposés dans la littérature (par exemple Best et coll. 2005) pour la cartographie des maladies en vue d'intégrer les effets spatialement corrélés, nous avons étendu le modèle avec effets de domaine indépendants à un modèle de corrélation spatiale, et l'avons combiné aux modèles pour petits domaines classiques. Les modèles à corrélation spatiale pour petits domaines CAR-MFH et CAR-MYC que nous proposons comprennent les modèles d'échantillonnage pour petits domaines et un modèle de lien à corrélation spatiale qui reflète l'hétérogénéité non structurée entre les domaines, ainsi que les effets de corrélation spatiale des domaines voisins. Le paramètre d'autocorrélation spatiale ne doit pas être spécifié dans le modèle et est estimé d'après les données.

Dans l'analyse des données, nous avons comparé les modèles spatiaux proposés aux modèles à effets non spatiaux pour estimer le taux de prévalence de l'asthme dans les 20 régions sociosanitaires de la Colombie-Britannique. Nous constatons que, comparativement aux estimations directes, les estimations fondées sur un modèle représentent une amélioration importante, qui se traduit par des estimations ponctuelles modérément lissées et des CV beaucoup plus petits. En particulier, les modèles proposés sont supérieurs au modèle de Fay-Herriot ou à celui de You-Chapman, que les variances d'échantillonnage soient supposées connues ou inconnues. En outre, la réduction des CV produits par les modèles spatiaux proposés comparativement au modèle de Fay-Herriot ou au modèle de You-Chapman est plus importante pour les domaines ayant un grand nombre de voisins. La comparaison des modèles bayésiens et l'analyse d'adéquation du modèle donnent aussi des résultats favorables aux modèles spatiaux pour petits domaines proposés.

Dans de futurs travaux, les modèles spatiaux pour petits domaines proposés pourront être étendus aux modèles d'échantillonnage et de lien non appariés (You et Rao 2002) sous variance d'échantillonnage connue ou inconnue. Nous prévoyons évaluer les effets de divers modèles spatiaux, ainsi que les effets des structures spatiales sur l'estimation. Pour l'analyse des données, nous produirons des estimations de l'état de santé fondées sur les modèles proposés pour les régions sociosanitaires des diverses régions du Canada et étudierons la possibilité d'étendre l'approche fondée sur un modèle à la production d'estimations à un niveau de détail plus fin, tel que les domaines âge-sexe à l'intérieur des régions sociosanitaires. Nous prévoyons également examiner la méthode de clonage des données (Lele, Dennis et Lutscher 2007 ; Lele, Nadeem et Schmuland 2010) pour les modèles spatiaux. Un avantage de cette méthode est que les résultats sont indépendants du choix des priors. Cependant, la demande de ressources informatiques pourrait être considérable.

Remerciements

Nous remercions le rédacteur associé et un examinateur de leurs commentaires et suggestions détaillés. Les travaux

de recherche de Yong You ont été financés par les ressources de financement global de la recherche de la Direction de la méthodologie de Statistique Canada. Les travaux de Qian M. Zhou ont été exécutés sous la supervision de Yong You durant un stage de recherche à Statistique Canada financé par le MITACS/PNSDC. Q.M. Zhou a présenté les modèles proposés et certains résultats décrits dans l'article à l'Assemblée annuelle de 2008 de la Société statistique du Canada (SSC) tenue à Ottawa et a été la lauréate 2008 du prix du meilleur article présenté par un étudiant décerné par la Section des méthodes d'enquête de la SSC.

Annexe A

Lois conditionnelles complètes

A.1. Lois conditionnelles complètes pour l'échantillonnage de Gibbs sous le modèle 1 : MFH.

- $[\theta_i \mid y_i, \beta, \sigma_v^2] \sim \mathrm{N}[\gamma_i y_i + (1-\gamma_i)\mathbf{x}_i'\beta, \tilde{\sigma}_i^2\gamma_i]$, où $\gamma_i = \sigma_v^2/(\sigma_v^2+\tilde{\sigma}_i^2)$, pour $i = 1, ..., m$;
- $[\beta|\boldsymbol{\theta}, \sigma_v^2] \sim \mathrm{N}\left[\left(\sum_{i=1}^m \mathbf{x}_i\mathbf{x}'\right)^{-1}\left(\sum_{i=1}^m \mathbf{x}_i\theta_i\right), \sigma_v^2\left(\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right]$;
- $[\sigma_v^2 \mid \boldsymbol{\theta}, \beta] \sim \mathrm{GI}\left[a_0 + \frac{1}{2}m, b_0 + \frac{1}{2}\sum_{i=1}^m (\theta_i - \mathbf{x}_i'\beta)^2\right]$.

A.2. Lois conditionnelles complètes pour l'échantillonnage de Gibbs sous le modèle 2 : CAR-MFH.

- $[\boldsymbol{\theta} \mid \mathbf{y}, \beta, \lambda, \sigma_v^2] \sim \mathrm{MVN}(\boldsymbol{\Lambda}\mathbf{y} + (\mathbf{I}-\boldsymbol{\Lambda})\mathbf{X}\beta, \boldsymbol{\Lambda}\mathbf{E})$, où $\boldsymbol{\Lambda} = (\mathbf{E}^{-1}+\mathbf{D}/\sigma_v^2)^{-1}\mathbf{E}^{-1}$ avec $\mathbf{E} = \mathrm{diag}\{\tilde{\sigma}_1^2, ..., \tilde{\sigma}_m^2\}$ et $\mathbf{D} = \lambda\mathbf{R} + (1-\lambda)\mathbf{I}$;
- $[\beta|\boldsymbol{\theta}, \lambda, \sigma_v^2] \sim \mathrm{MVN}[(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\boldsymbol{\theta}, \sigma_v^2(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}]$;
- $[\lambda|\boldsymbol{\theta}, \beta, \sigma_v^2] \propto \left|[\lambda\mathbf{R} + (1-\lambda)\mathbf{I}]^{-1}\right|^{-1/2} \times \exp\left\{-\frac{1}{2\sigma_v^2}(\boldsymbol{\theta} - X\beta)'[\lambda\mathbf{R} + (1-\lambda)\mathbf{I}](\boldsymbol{\theta} - X\beta)\right\}$;
- $[\sigma_v^2|\boldsymbol{\theta}, \beta, \lambda] \sim \mathrm{GI}\left[a_0 + \frac{m}{2}, b_0 + \frac{1}{2}(\boldsymbol{\theta} - X\beta)'\mathbf{D}(\boldsymbol{\theta} - X\beta)\right]$.

A.3. Lois conditionnelles complètes pour l'échantillonnage de Gibbs sous le modèle 3 : MYC.

- $[\theta_i \mid y_i, \beta, \sigma_i^2, \sigma_v^2] \sim \mathrm{N}[\gamma_i y_i + (1-\gamma_i)\mathbf{x}_i'\beta, \sigma_i^2\gamma_i]$, où $\gamma_i = \sigma_v^2/(\sigma_v^2+\sigma_i^2)$, pour $i = 1, ..., m$;
- $[\beta|\boldsymbol{\theta}, \sigma_v^2] \propto \mathrm{N}\left[\left(\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^m \mathbf{x}_i\theta_i\right), \sigma_v^2\left(\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right]$;
- $[\sigma_i^2 \mid y_i, \theta_i] \sim \mathrm{GI}\left(a_i + \frac{d_i+1}{2}, b_i + \frac{(y_i-\theta_i)^2 + d_i s_i^2}{2}\right)$, où $d_i = n_i - 1$, pour $i = 1, ..., m$;
- $[\sigma_v^2 \mid \boldsymbol{\theta}, \beta] \sim \mathrm{GI}\left[a_0 + \frac{1}{2}m, b_0 + \frac{1}{2}\sum_{i=1}^m (\theta_i - \mathbf{x}_i'\beta)^2\right]$.

A.4. Lois conditionnelles complètes pour l'échantillonnage de Gibbs sous le modèle 4 : CAR-MYC.

- $[\boldsymbol{\theta} \mid \mathbf{y}, \beta, \lambda, \sigma_v^2, \sigma_i^2] \sim \mathrm{MVN}(\boldsymbol{\Lambda}\mathbf{y} + (\mathbf{I}-\boldsymbol{\Lambda})\mathbf{X}\beta, \boldsymbol{\Lambda}\mathbf{E})$, où $\boldsymbol{\Lambda} = (\mathbf{E}^{-1}+\mathbf{D}/\sigma^2)^{-1}\mathbf{E}^{-1}$, et $\mathbf{E} = \mathrm{diag}\{\sigma_1^2, ..., \sigma_m^2\}$, $\mathbf{D} = \lambda\mathbf{R} + (1-\lambda)\mathbf{I}$;
- $[\beta \mid \boldsymbol{\theta}, \lambda, \sigma_v^2] \sim \mathrm{MVN}[(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\boldsymbol{\theta}, \sigma_v^2(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}]$;
- $[\lambda \mid \boldsymbol{\theta}, \beta, \sigma_v^2] \propto \left|[\lambda\mathbf{R} + (1-\lambda)\mathbf{I}]^{-1}\right|^{-1/2} \times \exp\left\{-\frac{1}{2\sigma_v^2}(\boldsymbol{\theta} - X\beta)'[\lambda\mathbf{R} + (1-\lambda)\mathbf{I}](\boldsymbol{\theta} - X\beta)\right\}$

  $[\sigma_i^2 \mid y_i, \theta_i] \sim \mathrm{GI}\left(a_i + \frac{d_i+1}{2}, b_i + \frac{(y_i-\theta_i)^2 + d_i s_i^2}{2}\right)$, où $d_i = n_i - 1$, pour $i = 1, ..., m$;
- $[\sigma_v^2 \mid \boldsymbol{\theta}, \beta, \lambda] \sim \mathrm{GI}\left[a_0 + \frac{m}{2}, b_0 + \frac{1}{2}(\boldsymbol{\theta} - X\beta)'\mathbf{D}(\boldsymbol{\theta} - X\beta)\right]$.

Annexe B

Liste des 20 régions sociosanitaires de la Colombie-Britannique avec les tailles d'échantillon et les structures spatiales correspondantes

Numéro d'ID	Nom de la région sociosanitaire	Taille de l'échantillon	Nombre de régions voisines	Régions voisines
1	East Kootenay	645	3	2, 3, 15
2	West Kootenay-Boundary	705	3	1, 3, 4
3	North Okanagan	890	5	1, 2, 4, 5, 15
4	South Okanagan Similameen	1 063	4	2, 3, 5, 6
5	Thompson	982	7	3, 4, 6, 9, 11, 12, 15
6	Fraser Valley	1 125	5	4, 5, 7, 8, 9
7	South Fraser Valley	1 437	4	6, 8, 17, 19
8	Simon Fraser	1,165	5	6, 7, 9, 17, 18
9	Coast Garibaldi	623	5	5, 6, 8, 11, 18
10	Central Vancouver Island	1 077	2	11, 20
11	Upper Island/Central Coast	746	4	5, 9, 10, 12
12	Cariboo	673	4	5, 11, 13, 15
13	North West	650	3	12, 14, 15
14	Peace Liard	611	2	13, 15
15	Northern Interior	859	6	1, 3, 5, 12, 13, 14
16	Vancouver	1 285	4	17, 18, 19, 20
17	Burnaby	871	5	7, 8, 16, 18, 19
18	North Shore	842	4	8, 9, 16, 17
19	Richmond	828	3	7, 16, 17
20	Capital	1 225	2	10, 16

Nota : Vancouver (n°16) et Capital (n°20) ne sont pas des régions adjacentes sur la carte, puisqu'elles sont séparées par l'océan. Cependant, étant donné le lien intensif et étroit entre ces deux régions, nous les définissons comme voisines dans notre étude à titre d'illustration seulement.

Annexe C

Estimations ponctuelles directes et fondées sur un modèle et CV

Comparaison des estimations ponctuelles					
ID du domaine	**Est. directe**	**MFH**	**CAR-MFH**	**MYC**	**CAR-MYC**
1	0,0765	0,0793	0,0812	0,0795	0,0812
2	0,0804	0,0795	0,0793	0,0797	0,0794
3	0,0745	0,0726	0,0731	0,0725	0,0729
4	0,0893	0,0868	0,0874	0,0867	0,0873
5	0,0782	0,0739	0,0736	0,0729	0,0731
6	0,0943	0,0914	0,0927	0,0918	0,0928
7	0,0702	0,0707	0,0712	0,0711	0,0717
8	0,0858	0,0845	0,0848	0,0844	0,0849
9	0,0877	0,0763	0,0745	0,0765	0,0747
10	0,0763	0,0805	0,0799	0,0805	0,0796
11	0,0661	0,0685	0,0678	0,0679	0,0676
12	0,0717	0,0681	0,0681	0,0678	0,0677
13	0,0631	0,0687	0,0692	0,0690	0,0693
14	0,0673	0,0685	0,0680	0,0685	0,0686
15	0,0793	0,0721	0,0707	0,0728	0,0713
16	0,0657	0,0696	0,0702	0,0697	0,0704
17	0,0859	0,0778	0,0759	0,0773	0,0759
18	0,0583	0,0626	0,0633	0,0618	0,0626
19	0,0619	0,0649	0,0647	0,0653	0,0647
20	0,0877	0,0923	0,0914	0,0917	0,0908
Comparaison des CV					
ID du domaine	**Est. directe**	**MFH**	**CAR-MFH**	**MYC**	**CAR-MYC**
1	0,168	0,107	0,099	0,107	0,100
2	0,127	0,105	0,104	0,097	0,093
3	0,135	0,116	0,106	0,110	0,097
4	0,102	0,084	0,076	0,079	0,072
5	0,158	0,094	0,076	0,105	0,083
6	0,113	0,086	0,080	0,086	0,081
7	0,124	0,099	0,096	0,106	0,101
8	0,102	0,085	0,076	0,081	0,073
9	0,158	0,119	0,105	0,117	0,105
10	0,121	0,087	0,086	0,086	0,084
11	0,141	0,118	0,108	0,109	0,105
12	0,196	0,119	0,109	0,130	0,116
13	0,168	0,115	0,108	0,111	0,108
14	0,206	0,126	0,125	0,136	0,133
15	0,121	0,101	0,087	0,094	0,083
16	0,127	0,101	0,097	0,103	0,097
17	0,124	0,107	0,100	0,105	0,096
18	0,155	0,143	0,136	0,134	0,130
19	0,154	0,135	0,134	0,128	0,128
20	0,103	0,086	0,085	0,083	0,082

Bibliographie

Arora, V., et Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

Bayarri, M.J., et Berger, J.O. (2000). P values for composite null models. *Journal of the American Statistical Association*, 95, 1127-1142.

Béland, Y. (2002). Enquête sur la santé dans les collectivités canadiennes – Aperçu de la méthodologie. Rapports sur la santé, Statistique Canada, n° 82-003 au catalogue, 13, 3, 9-15.

Besag, J., York, J. et Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (avec discussion). *Annals of the Institute of Statistical Mathematics*, 43, 1-59.

Best, N., Richardson, S. et Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14, 35-39.

Brown, G., Chambers, R., Heady, P. et Heasman, D. (2001). Évaluation des méthodes d'estimation régionale dans leur application aux estimations du chômage tirées de l'Enquête sur la population active au Royaume-Uni. Recueil : Symposium 2001, *La qualité des données d'un organisme statistique* : *une perspective méthodologique*, Statistique Canada, CD-ROM, 1-10.

Chip, S., et Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49, 327-335.

Cressie, N. (1990). Prédiction du sous-dénombrement pour les petites régions à l'aide du modèle linéaire général. Recueil : Symposium 1990, *Mesure et amélioration de la qualité des données*, Statistique Canada, 103-116.

Daniels, M.J., et Gatsonis, C. (1999). Hierarchical generalized linear models in the analysis of variations in health care utilization. *Journal of the American Statistical Association*, 94, 29-42.

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.

Dick, P. (1995). Modélisation du sous-dénombrement net dans le recensement du Canada de 1991. *Techniques d'enquête*, 21, 51-61.

Fay, R.E., et Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 268-277.

Gelfand, A.E. (1996). Model determination using sampling-based methods. Dans *Markov Monte Carlo in Practice* (Éds., W.R. Gilks, S. Richardson et D.J. Spiegelhalter), Londres : Chapman & Hall, 145-161.

Gelman, A., Carlin, J.B., Stern, H.S. et Rubin, D.B. (2004). *Bayesian Data Analysis*, 2e Édition. Chapman & Hall/CRC.

Gelman, A., Meng, X.L. et Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (avec discussion). *Statistica Sinica*, 6, 733-807.

Gelman, A., et Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.

Ghosh, M., Natarajan, K., Stroud, T.W.F. et Carlin, B.P. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association*, 93, 273-282.

Ghosh, M., Natarajan, K., Walter, L.A. et Kim, D.H. (1999). Hierarchical Bayes GLMs for the analysis of spatial data: An application to disease mapping. *Journal of Statistical Planning and Inference*, 75, 305-318.

He, Z., et Sun, D. (2000). Hierarchical Bayes estimation of hunting success rates with spatial correlations. *Biometrics*, 56, 360-367.

Jiang, J., et Lahiri, P. (2006). Mixed model prediction and small area estimation (avec discussion). *Test*, 15, 1-96.

Lele, S.R., Dennis, B. et Lutscher, F. (2007). Data cloning: Easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters*, 10, 551-563.

Lele, S.R., Nadeem, K. et Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105, 1617-1625.

Leroux, B.G., Lei, X. et Breslow, N. (1999). Estimation of disease rates in small areas: A new mixed model for spatial dependence. Dans *Statistical Models in Epidemiology, the Environment and Clinical Trials*, (Éds., M.E. Halloran et D. Berry). New York : Springer Verlag, 135-178.

Liu, B., Lahiri, P. et Kalton, G. (2008). Hiearchical Bayes modeling of survey weighted small area proportions. Manuscript non-publié.

Maiti, T. (1998). Hierarchical Bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, 69, 339-348.

Mohadjer, L., Rao, J.N.K., Liu, B., Krenzke, T. et van de Kerckhove, W. (2007). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *Proceedings of the American Statistical Association, Section of Survey Method Research.*

Moura, F.A.S., et Migon, H.S. (2002). Bayesian spatial models for small area proportions. *Statistical Modelling*, 2, 3, 183-201.

MacNab, Y.C. (2003). Hierarchial Bayesian spatial modeling of small-area rates of non-rare disease. *Statistics in Medicine*, 22, 1761-1773.

Maiti, T. (1998). Hierarchical Bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, 69, 339-348.

Meng, X.L. (1994). Posterior predictive p value. *The Annals of Statistics*, 22, 1142-1160.

Mollié, A. (1996). Bayesian mapping of disease. Dans *Markov Chain Monte Carlo in Practice*. Londres : Chapman and Hall, 359-379.

Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.

Singh, A.C., Folsom, R.E., Jr. et Vaish, A.K. (2005). Small area modeling for survey data with smoothed error covariance structure via generalized design effects. Federal Committee on Statistical methods Conference proceedings, Washington, D.C., www.fcsm.gov.

Singh, B.B., Shukla, G.K. et Kundu, D. (2005). Modèles spatio-temporels pour l'estimation pour petits domaines. *Techniques d'enquête*, 31, 201-214.

Sinharay, S., et Stern, H.S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, 111, 209-221.

Spiegelhalter, D.J., Best, N., Carlin, B.P. and van de Linde, A. (2002). Bayesain measures of model complexity and fit. *Journal of Royal Statistical Society*, B, 64, 583-639.

Souza, D.F., Moura, F.A.S. et Migon, H.S. (2009). Prédiction de la population de petits domaines au moyen de modèles hiérarchiques. *Techniques d'enquête*, 35, 221-234.

Wang, J., et Fuller, W. A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y. (2008a). Une approche intégrée de modélisation de l'estimation du taux de chômage pour les régions infraprovinciales au Canada. *Techniques d'enquête*, 34, 21-31.

You, Y. (2008b). Small area estimation using area level models with model checking and applications. *Proceedings of Survey Methods Section, Statistical Society of Canada*.

You, Y., et Chapman, B. (2006). Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances d'échantillonnage. *Techniques d'enquête*, 32, 107-114.

You, Y., et Rao, J.N.K. (2000). Estimation bayésienne hiérarchique des moyennes pour petites régions à l'aide de modèles à plusieurs niveaux. *Techniques d'enquête*, 26, 197-206.

You, Y., et Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 20, 3-15.

Techniques d'enquête, juin 2011
Vol. 37, N° 1, pp. 45-58
Statistique Canada, N° 12-001-X au catalogue

Estimation sur petits domaines sous linéarisation

Hukum Chandra et Ray Chambers [1]

Résumé

L'estimation sur petits domaines fondée sur des modèles linéaires mixtes est parfois inefficace quand les relations sous-jacentes ne sont pas linéaires. Nous présentons des techniques d'estimation sur petits domaines pour des variables qui peuvent être modélisées linéairement après une transformation non linéaire. En particulier, nous étendons l'estimateur direct fondé sur un modèle de Chandra et Chambers (2005, 2009) à des données qui concordent avec un modèle linéaire mixte sur l'échelle logarithmique, en utilisant le calage sur un modèle pour définir des poids pouvant être utilisés dans cet estimateur. Nos résultats montrent que l'estimateur fondé sur la transformation que nous obtenons est à la fois efficace et robuste à la distribution des effets aléatoires dans le modèle. Une application à des données d'enquêtes auprès des entreprises démontre la performance satisfaisante de la méthode.

Mots clés : Enquête par sondage ; estimation par sondage ; enquêtes auprès des entreprises ; calage sur un modèle ; données asymétriques ; estimation directe fondée sur un modèle ; meilleur prédicteur sans biais empirique.

1. Introduction

Les méthodes utilisées habituellement pour l'estimation sur petits domaines reposent sur l'hypothèse qu'un modèle linéaire mixte peut être utilisé pour caractériser la relation de régression entre la variable étudiée Y et la variable auxiliaire X dans les petits domaines d'intérêt. En particulier, le meilleur prédicteur linéaire sans biais empirique (EBLUP) (voir Rao, 2003, chapitres 6 à 8) est habituellement fondé sur une hypothèse de modèle linéaire mixte. Toutefois, quand les données sont asymétriques comme cela est souvent le cas dans les enquêtes auprès des entreprises, la relation entre Y et X est parfois non linéaire sur l'échelle originale (brute), mais peut être linéaire sur une échelle transformée, par exemple l'échelle logarithmique (log). Le cas échéant, nous pouvons nous attendre à ce que l'estimation fondée sur un modèle linéaire mixte pour Y soit inefficace comparativement à une estimation fondée sur un modèle similaire pour la version transformée de Y. Voir Hidiroglou et Smith (2005). Le recours à des transformations en inférence n'est pas un fait nouveau (voir, par exemple, Carroll et Ruppert (1988, chapitre 4)). Récemment, Chen et Chen (1996), ainsi que Karlberg (2000a) ont étudié l'utilisation d'une approche de « linéarisation » pour l'estimation par la régression de variables étudiées qui se comportent non linéairement. Cependant, autant que nous sachions, cette idée n'a jamais été appliquée à l'estimation sur petits domaines, même si la théorie économique (et l'observation non formelle) laissent entendre que, dans le cas des données d'enquêtes auprès des entreprises, les relations de régression sont habituellement multiplicatives, et donc linéaires sur l'échelle logarithmique.

Dans le présent article, nous étendons les notions d'estimation directe fondée sur un modèle (DFM) décrites dans Chandra et Chambers (2005, 2009) à la situation où le modèle linéaire mixte qui sous-tend l'estimation sur petits domaines est vérifié sur l'échelle logarithmique, en utilisant des poids calculés par calage sur un modèle (Wu et Sitter 2001). Ce faisant, nous notons que notre approche est facilement généralisable à d'autres transformations monotones (c'est-à-dire inversibles). En revanche, l'extension de l'approche EBLUP à la situation où les données suivent un modèle linéaire mixte sous transformation est compliquée. Nous relâchons également l'hypothèse habituelle de normalité pour les effets de domaine afin d'examiner la robustesse à cette hypothèse.

À la section qui suit, nous résumons l'approche DFM de l'estimation sur petits domaines sous un modèle linéaire mixte. À la section 3, nous décrivons une alternative au modèle linéaire mixte pour des données asymétriques qui se réduit au modèle linéaire mixte sous transformation logarithmique, et à la section 4, nous adoptons une perspective fondée sur un modèle pour motiver l'estimation calée sur un modèle des quantités de populations quand la variable sous-jacente est linéaire après une transformation appropriée. À la section 5, nous regroupons ces deux idées, et introduisons le concept d'un modèle à valeurs prédites dérivé d'un modèle linéaire mixte sous l'échelle transformée. Puis, nous utilisons ce modèle à valeurs prédites pour spécifier les poids de sondage à utiliser dans un estimateur DFM pour l'estimation sur petits domaines. À la section 6, nous présentons les résultats empiriques d'un certain nombre d'études par simulation destinées à comparer l'estimateur DFM sous transformation proposé à l'estimateur

1. Hukum Chandra, Indian Agricultural Statistics Research Institute, Library Avenue, PUSA Campus, New Delhi-110012, Inde. Courriel : hchandra@iasri.res.in; Ray Chambers, Centre for Statistical and Survey Methodology, University of Wollongong, Wollongong, NSW, 2522, Australie. Courriel : ray@uow.edu.au.

EBLUP ainsi qu'à l'estimateur DFM « habituel » défini en ajustant un modèle linéaire mixte aux données, ainsi qu'un prédicteur empirique indirect basé sur le même modèle linéaire mixte sous échelle transformée. À la section 7, nous concluons l'article par une discussion des questions en suspens.

Il convient de souligner que l'approche adoptée dans le présent article est fondée sur un modèle. Par conséquent, tous les moments sont évalués par rapport à un modèle pour les données de population. En outre, nous supposons que toutes les données d'échantillon ont été obtenues par une méthode d'échantillonnage non informatif, par exemple l'échantillonnage probabiliste avec probabilités d'inclusion définies par des covariables connues du modèle.

2. Estimation directe fondée sur un modèle pour petits domaines

Pour commencer, nous établissons la notation. Soit U une population de taille N et soit $\mathbf{y}_U$, le vecteur de dimension N des valeurs de population d'une caractéristique Y d'intérêt. Supposons que notre principal objectif soit l'estimation du total $t_{Uy} = \sum_U y_j$ de ces valeurs de population (ou de leur moyenne $m_{Uy} = N^{-1}\sum_U y_j$). Soit $\mathbf{X}$ un vecteur de dimension p de variables auxiliaires qui sont reliées, d'une certaine façon, à Y et soit $\mathbf{x}_U$, la matrice de dimensions $N \times p$ correspondante des valeurs de population de ces variables. Nous supposons que les valeurs d'échantillon individuelles de $\mathbf{X}$ sont connues. Les valeurs hors échantillon de $\mathbf{X}$ ne sont pas nécessairement connues individuellement, mais sont supposées connues à un certain niveau d'agrégation. Au minimum, nous connaissons le vecteur des totaux de population $\mathbf{t}_{Ux}$ des colonnes de $\mathbf{X}$.

Supposons qu'il est raisonnable d'émettre l'hypothèse que la régression de Y sur $\mathbf{X}$ au sein de la population est linéaire, c'est-à-dire que

$$E(\mathbf{y}_U \mid \mathbf{x}_U) = \mathbf{x}_U\beta \text{ et } \mathrm{Var}(\mathbf{y}_U \mid \mathbf{x}_U) = \mathbf{v}_U \qquad (1)$$

où $\mathbf{v}_U$ est connu jusqu'à une constante multiplicative. Étant donné un échantillon s de taille n pour cette population, nous pouvons partitionner

$$\mathbf{x}_U = \begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_r \end{bmatrix}$$

et

$$\mathbf{v}_U = \begin{bmatrix} \mathbf{v}_{ss} & \mathbf{v}_{sr} \\ \mathbf{v}_{rs} & \mathbf{v}_{rr} \end{bmatrix}$$

en leurs composantes dans l'échantillon et hors échantillon. Ici, $r = U - s$ désigne les unités de population qui ne sont pas incluses dans l'échantillon. Le vecteur des poids qui définissent le meilleur prédicteur linéaire sans biais (BLUP pour *Best Linear Unbiased Predictor*) de t_{Uy} est alors (voir Royall 1976 ; Valliant, Dorfman et Royall 2000, section 2.4)

$$\begin{aligned} \mathbf{w}_s^{\mathrm{BLUP}} &= (w_j^{\mathrm{BLUP}} ;\ j \in s) \\ &= \mathbf{1}_s + \mathbf{H}_s'(\mathbf{t}_{Ux} - \mathbf{t}_{sx}) + (\mathbf{I}_s - \mathbf{H}_s'\mathbf{x}_s')\mathbf{v}_{ss}^{-1}\mathbf{v}_{sr}\mathbf{1}_r \end{aligned} \qquad (2)$$

où $\mathbf{H}_s = (\mathbf{x}_s'\mathbf{v}_{ss}^{-1}\mathbf{x}_s)^{-1}\mathbf{x}_s'\mathbf{v}_{ss}^{-1}$, $\mathbf{I}_s$ est la matrice identité d'ordre n, $\mathbf{t}_{sx}$ est le vecteur des totaux d'échantillon de $\mathbf{X}$ et $\mathbf{1}_s(\mathbf{1}_r)$ désigne un vecteur de valeurs 1 de taille n $(N - n)$.

Nous supposons maintenant que la population cible U de taille N peut être partitionnée en D petits domaines non chevauchants, chacun de taille N_i, $i = 1, ..., D$, tels que $N = \sum_{i=1}^{D} N_i$. Sachant qu'un échantillon s de n unités est tiré de cette population, nous supposerons qu'un sous-échantillon s_i de n_i unités est tiré du domaine i, avec $n = \sum_{i=1}^{D} n_i$. Il convient de noter que nous supposons que tous les petits domaines sont échantillonnés et qu'il existe au moins une unité échantillonnée dans chaque petit domaine d'intérêt.

Comme nous l'avons mentionné à la section 1, des modèles linéaires mixtes sont souvent utilisés pour l'estimation sur petits domaines. Ces modèles peuvent s'écrire sous la forme

$$\mathbf{y}_U = \mathbf{x}_U\beta + \mathbf{g}_U\mathbf{u} + \mathbf{e}_U \qquad (3)$$

où $\mathbf{u}$ est un vecteur aléatoire d'effets dits de domaine, $\mathbf{e}_U$ est un vecteur de population de dimension N d'effets individuels aléatoires et $\mathbf{g}_U$ est un matrice connue. En général, les effets de domaine sont évalués vectoriellement, de sorte que $\mathbf{u}' = (\mathbf{u}_1'\mathbf{u}_2'\cdots\mathbf{u}_D')$ et $\mathbf{g}_U = \mathrm{diag}\{\mathbf{g}_i ;\ i = 1, \ldots, D\}$, où $\mathbf{g}_i$ est de dimensions $N_i \times q$. Nous supposons que les effets particuliers au domaine $\{\mathbf{u}_i ;\ i = 1, \ldots, D\}$ sont des réalisations indépendantes et identiquement distribuées d'un vecteur aléatoire de dimension q dont la moyenne est nulle et la matrice de covariance est Σ_u. De même, nous supposons que les effets individuels scalaires constituant $\mathbf{e}_U$ sont des réalisations indépendantes et identiquement distribuées d'une variable aléatoire de moyenne nulle et de variance σ_e^2, les effets de domaine et les effets individuels étant mutuellement indépendants. Les paramètres $\theta = (\Sigma_u, \sigma_e^2)$ sont habituellement appelés les composantes de la variance de (3).

Étant donné les valeurs des composantes de la variance, il est facile de voir que (3) est simplement un cas particulier du modèle linéaire général (1) qui sous-tend les poids BLUP (2). En particulier, sous (3)

$$\begin{aligned} \mathbf{v}_{ss} &= \mathrm{diag}\{\mathbf{v}_{iss} ;\ i = 1, \ldots, D\} \\ &= \mathrm{diag}\{\mathbf{g}_{is}\Sigma_u\mathbf{g}_{is}' + \sigma_e^2\mathbf{I}_{is} ;\ i = 1, \ldots, D\} \end{aligned} \qquad (4)$$

et

$$\begin{aligned}\mathbf{v}_{sr} &= \text{diag}\{\mathbf{v}_{isr}\ ;\ i = 1, \ldots, D\} \\ &= \text{diag}\{\mathbf{g}_{is}\Sigma_u\,\mathbf{g}'_{ir}\ ;\ i = 1, \ldots, D\}. \qquad (5)\end{aligned}$$

Ici, $\mathbf{g}_{is}$ et $\mathbf{g}_{ir}$ désignent la restriction de $\mathbf{g}_i$ aux unités échantillonnées et non échantillonnées dans le domaine i, respectivement. Sachant les valeurs estimées $\hat{\theta} = (\hat{\Sigma}_u, \hat{\sigma}_e^2)$ des composantes de la variance, nous pouvons les introduire par substitution dans (4) et (5) pour obtenir les estimations $\hat{\mathbf{v}}_{ss}$ et $\hat{\mathbf{v}}_{sr}$ de $\mathbf{v}_{ss}$ et $\mathbf{v}_{sr}$, respectivement, et par conséquent calculer les poids BLUP « empiriques », ou poids EBLUP, pour le total de population de Y sous la forme

$$\begin{aligned}\mathbf{w}_s^{\text{EBLUP}} &= (w_{ij}^{\text{EBLUP}}\ ;\ j \in s_i\ ;\ i = 1, \ldots, D) \\ &= \mathbf{1}_s + \hat{\mathbf{H}}'_s(\mathbf{t}_{Ux} - \mathbf{t}_{sx}) \\ &\quad + (\mathbf{I}_s - \hat{\mathbf{H}}'_s\mathbf{x}'_s)\hat{\mathbf{v}}_{ss}^{-1}\hat{\mathbf{v}}_{sr}\mathbf{1}_r \qquad (6)\end{aligned}$$

où $\hat{\mathbf{H}}_s = (\mathbf{x}'_s\hat{\mathbf{v}}_{ss}^{-1}\mathbf{x}_s)^{-1}\mathbf{x}'_s\,\hat{\mathbf{v}}_{ss}^{-1}$. Notons que nous utilisons maintenant le double indice inférieur ij afin de faire la distinction entre les unités de population dans les différents domaines.

L'estimateur DFM de la moyenne m_{iy} de Y dans le domaine i (Chandra et Chambers 2005, 2009) fondé sur les poids EBLUP pour le total (6) est simplement la moyenne pondérée correspondante des valeurs d'échantillon de Y dans le domaine i,

$$\hat{m}_{iy}^{\text{HJ-DFMLin}} = \left\{\sum\nolimits_{j\in s_i} w_{ij}^{\text{EBLUP}}\right\}^{-1}\sum\nolimits_{j\in s_i} w_{ij}^{\text{EBLUP}}\,y_{ij}. \qquad (7)$$

Notons que (7) *n'est pas* l'EBLUP pour m_{iy} sous (3). Celui-ci est (voir Rao 2003, section 6.2.3)

$$\begin{aligned}&\hat{m}_{iy}^{\text{HT-EBLUPLin}} \\ &= \hat{E}\{m_{iy}\,|\,\mathbf{y}_{is}, \mathbf{x}_{is}, \mathbf{x}_{ir}\} \\ &= N_i^{-1}\left[\sum\nolimits_{j\in s_i} y_j + \mathbf{1}'_{ir}\left\{\mathbf{x}_{ir}\hat{\beta} + \hat{\mathbf{v}}_{irs}\hat{\mathbf{v}}_{iss}^{-1}(\mathbf{y}_{is} - \mathbf{x}_{is}\hat{\beta})\right\}\right] \\ &= N_i^{-1}\Big[n_i\bar{y}_{is} + (N_i - n_i) \\ &\quad \left\{\bar{\mathbf{x}}'_{ir}\hat{\beta} + \bar{\mathbf{g}}'_{ir}\hat{\Sigma}_u\mathbf{g}'_{is}(\mathbf{g}_{is}\hat{\Sigma}_u\mathbf{g}'_{is} + \hat{\sigma}_e^2\mathbf{I}_{is})^{-1}(\mathbf{y}_{is} - \mathbf{x}_{is}\hat{\beta})\right\}\Big]. \qquad (8)\end{aligned}$$

Ici, $\hat{E}$ désigne l'opérateur d'espérance sous (3) avec les paramètres inconnus remplacés par les estimations, $\mathbf{x}_{is}$ et $\mathbf{x}_{ir}$ sont les matrices des valeurs d'échantillon et hors échantillon de $\mathbf{X}$ dans le domaine i, $\mathbf{y}_{is}$ est le vecteur des valeurs d'échantillon de Y dans le même domaine, $\hat{\beta}$ est le prédicteur BLUE « empirique » de β, $\hat{\mathbf{v}}_{irs}$ est la transposée de la valeur estimée de $\mathbf{v}_{isr}$ avec $\hat{\mathbf{v}}_{iss}$ l'estimation correspondante de $\mathbf{v}_{iss}$, voir (4) et (5), et $\mathbf{1}_{ir}$ est un vecteur de valeurs 1 de longueur $N_i - n_i$. Notons que la dernière expression dans le deuxième membre de (8) découle directement de la substitution de (4) et (5), avec $\bar{\mathbf{x}}_{ir}$ et $\bar{\mathbf{g}}_{ir}$ désignant les vecteurs colonne d'ordre p et q définis en calculant la moyenne des colonnes de $\mathbf{x}_{ir}$ et $\mathbf{g}_{ir}$, respectivement. Comme le prédicteur EBLUP (8), l'estimateur (7) est une fonction pondérée de toutes les valeurs d'échantillon. Notons que sous la spécification de l'ordonnée à l'origine aléatoire de (3), (8) se réduit à l'expression (7.2.39) dans Rao (2003, section 7.2).

L'estimation de l'erreur quadratique moyenne (EQM) pour (8) est habituellement effectuée en appliquant la théorie décrite dans Prasad et Rao (1990). Bien que cet estimateur de l'EQM soit assez compliqué, il donne de bons résultats sous (3). Toutefois, si le modèle (3) n'est pas vérifié, les résultats peuvent être erronés. Il ne convient pas non plus comme estimateur de l'EQM de (8), sous échantillonnage répété, comme l'a fait remarquer Longford (2007). En revanche, l'estimation de l'EQM de (7) est assez simple. En effet, si l'on traite les poids qui définissent cet estimateur comme étant fixes, il s'agit d'un estimateur linéaire d'une moyenne de domaine et, par conséquent, sa variance de prédiction V_i sous (1) peut être estimée au moyen de méthodes bien connues (voir Royall et Cumberland 1978). Puisqu'en général les poids EBLUP pour le total (6) ne sont pas « calés localement » (c'est-à-dire qu'ils ne reproduisent pas la moyenne $\bar{\mathbf{x}}_i$ du domaine i de $\mathbf{X}$), (7) possède un biais B_i sous (1). Une simple estimation de ce biais par introduction des valeurs est la différence entre (7) et $\bar{\mathbf{x}}'_i\hat{\beta}$. L'estimateur final de l'EQM utilisé avec (7) est par conséquent défini en additionnant l'estimation de V_i et le carré de cette estimation de B_i. On a démontré empiriquement que cette méthode d'estimation de l'EQM possède de bonnes propriétés fondées sur un modèle ainsi que sous échantillonnage répété. Voir Chandra et Chambers (2005, 2009), Chambers et Tzavidis (2006), Chandra, Salvati et Chambers (2007), ainsi que Tzavidis, Salvati, Pratesi et Chambers (2008).

3. Estimation sur petits domaines sous transformation

À la présente section, nous étendons l'approche DFM à l'estimation sur petits domaines quand les relations de régression sous-jacentes ne sont pas linéaires. Ce faisant, nous nous concentrerons sur le cas important où les valeurs de population de Y suivent un modèle non linéaire sur leur échelle originale (brute), mais où leurs logarithmes peuvent être modélisés linéairement. L'extension à d'autres modèles de linéarisation est simple.

Sans perte de généralité, nous supposons que Y et X sont toutes deux des grandeurs scalaires strictement positives, dont la distribution marginale de population est asymétrique et pour lesquelles existent des preuves manifestes que leur relation est non linéaire, comme cela est le cas pour de nombreuses applications aux enquêtes auprès des entreprises. En outre, un modèle linéaire mixte convient pour

caractériser la façon dont la régression de $\log(Y)$ sur $\log(X)$ varie d'un petit domaine à l'autre. Autrement dit, pour $i = 1, ..., D$; $j = 1, ..., N_i$ nous avons

$$l_{ij} = \log(y_{ij}) = \beta_0 + \beta_1 \log(x_{ij}) + \mathbf{g}'_{ij}\mathbf{u}_i + e_{ij} \quad (9)$$

où y_{ij} et x_{ij} sont les valeurs de Y et X, respectivement, pour l'unité de population j dans le petit domaine i, $\mathbf{g}_{ij}$ désigne une covariable « contextuelle » de dimension q, $\mathbf{u}_i$ désigne un effet aléatoire pour le domaine i, également de dimension q, et e_{ij} est un effet aléatoire individuel scalaire. Comme d'habitude pour ce genre de modèle, nous supposons que tous les effets aléatoires suivent une loi normale et sont mutuellement non corrélés, que leurs valeurs espérées sont nulles et que $\text{Var}(\mathbf{u}_i) = \Sigma_u$ et $\text{Var}(e_{ij}) = \sigma_e^2$. Ici, Σ_u est la matrice de dimensions $q \times q$ de covariance des effets aléatoires. Notons que $\text{Var}(l_{ij} | x_{ij}) = v_{ijj} = \mathbf{g}'_{ij}\Sigma_u \mathbf{g}_{ij} + \sigma_e^2$ et $\text{Cov}(l_{ij}, l_{ik} | x_{ij}, x_{ik}, \mathbf{g}_{ij}, \mathbf{g}_{ik}) = v_{ijk} = \mathbf{g}'_{ij}\Sigma_u \mathbf{g}_{ik}$ sous (9).

Sachant les valeurs d'échantillon de y_{ij}, x_{ij} et $\mathbf{g}_{ij}$, nous pouvons appliquer des méthodes d'estimation classiques (par exemple, maximum de vraisemblance (MV) ou maximum de vraisemblance restreint (MVR), voir Harville 1977) pour estimer les paramètres de (9). Soit $\hat{\Sigma}_u$ et $\hat{\sigma}_e^2$ les estimations résultantes des composantes de la variance de ce modèle linéaire mixte. L'estimation de $\beta = (\beta_0\ \beta_1)'$ est alors

$$\hat{\beta} = \left(\sum_i \mathbf{d}'_{is}\hat{\mathbf{v}}_{iss}^{-1}\mathbf{d}_{is}\right)^{-1}\left(\sum_i \mathbf{d}'_{is}\hat{\mathbf{v}}_{iss}^{-1}\mathbf{l}_{is}\right) \quad (10)$$

où $\hat{\mathbf{v}}_{iss}$, $\mathbf{d}_{is}$ et $\mathbf{l}_{is}$ sont les composantes d'échantillon de $\hat{\mathbf{v}}_i = [\hat{v}_{ijk}] = \mathbf{g}_i\hat{\Sigma}_u\mathbf{g}'_i + \hat{\sigma}_e^2\mathbf{I}_i$, $\mathbf{d}_i = [d_{ijk}] = [\mathbf{1}_i\ \log(\mathbf{x}_i)]$ et $\mathbf{l}_i = (l_{ij};\ j = 1, \ldots, N_i)$, respectivement. Ici, $\mathbf{g}_i$ est la matrice de dimensions $N_i \times q$ définie par les covariables $\mathbf{g}_{ij}$ dans le domaine i, $\mathbf{I}_i$ est la matrice identité d'ordre N_i, $\mathbf{1}_i$ désigne un vecteur de valeurs 1 de dimension N_i et $\log(\mathbf{x}_i)$ désigne le vecteur de N_i valeurs de $\log(X)$ dans le domaine i.

Notons que, quand les composantes de la variance Σ_u et σ_e^2 sont connues, (10) est l'estimateur BLUE pour β. Par conséquent, $E(\hat{\beta}) \approx \beta$ et $\text{Var}(\hat{\beta}) \approx (\sum_i \mathbf{d}'_{is}\hat{\mathbf{v}}_{iss}^{-1}\mathbf{d}_{is})^{-1}$. Posons que $\hat{\phi}_i = (\hat{\phi}_{ij}) = \mathbf{d}_i\hat{\beta}$. Alors $E(\hat{\phi}_i) \approx \mathbf{d}_i\beta$ et $\text{Var}(\hat{\phi}_i) = \mathbf{A}_i = [a_{ijk}] \approx \mathbf{d}_i(\sum_g \mathbf{d}'_{gs}\hat{\mathbf{v}}_{gss}^{-1}\mathbf{d}_{gs})^{-1}\mathbf{d}'_i$, où $a_{ijk} = \mathbf{d}'_{ij}\text{Var}(\hat{\beta})\mathbf{d}_{ik} \to 0$ quand $n \to \infty$.

Notre but est d'utiliser le modèle linéaire mixte à échelle logarithmique (9) pour estimer les moyennes de petit domaine m_{iy}. En particulier, nous utilisons le calage sur un modèle (Wu et Sitter 2001) basé sur ce modèle pour calculer les poids de sondage à utiliser dans l'estimateur DFM (7) de cette quantité.

4. Pondération calée sur un modèle

Le calage sur un modèle a été introduit par Wu et Sitter (2001) en tant que méthode de pondération par calage assistée par un modèle quand les relations de régression sous-jacentes sont non linéaires. Ici, nous donnons une perspective fondée sur un modèle de la méthode, en tant que précurseur à son utilisation pour construire les poids qui doivent être utilisés dans l'estimateur DFM dans une situation similaire.

Supposons que le modèle de population sous-jacent est non linéaire, la relation entre Y et $\mathbf{X}$ dans la population étant de la forme

$$E(y_j | \mathbf{x}_j) = h(\mathbf{x}_j;\ \eta) \text{ et } \text{Var}(y_j | \mathbf{x}_j) = \sigma_j^2. \quad (11)$$

Ici, $j = 1, \ldots, N$, η (habituellement évalué vectoriellement) et σ_j^2 sont les paramètres inconnus du modèle, et la fonction de la moyenne $h(\mathbf{x}_j;\ \eta)$ est une fonction connue de $\mathbf{x}_j$ et η. Nous supposons aussi que les unités de population sont mutuellement non corrélées, sachant leurs valeurs respectives de $\mathbf{X}$. Notons que (11) est une expression assez générale qui englobe les modèles linéaire, non linéaire et linéaire généralisé comme cas particuliers. Dans cette situation, Wu et Sitter (2001) définissent l'estimateur calé sur un modèle (cm) du total de population t_{Uy} comme étant $\hat{t}_y^{cm} = \sum_{j\in s} w_j^{cm} y_j$, où le vecteur de poids $\mathbf{w}_s^{cm} = (w_j^{cm})$ et est choisi de manière à minimiser une mesure appropriée de la distance entre $\mathbf{w}_s^{cm}$ et le vecteur de poids d'Horvitz-Thompson $\mathbf{w}_s^{\pi} = (\pi_j^{-1})$, sous les contraintes de calage sur le modèle

$$\sum_{j\in s} w_j^{cm} = N$$
$$\text{et} \quad (12)$$
$$\sum_{j\in s} w_j^{cm} h(\mathbf{x}_j;\ \hat{\eta}_\pi) = \sum_{j\in U} h(\mathbf{x}_j;\ \hat{\eta}_\pi)$$

où $\hat{\eta}_\pi$ est un estimateur convergent sous le plan de η. Notons que, contrairement au calage classique, les contraintes (12) exigent que nous connaissions les valeurs de population individuelles de $\mathbf{X}$. L'idée essentielle sur laquelle s'appuie cette approche est que, à condition que (11) soit ajusté raisonnablement, y_j est (du moins approximativement) une fonction linéaire de sa valeur prédite $h(\mathbf{x}_j;\ \hat{\eta}_\pi)$ sous ce modèle et que nous pouvons donc exécuter l'estimation linéaire en utilisant ces valeurs prédites comme information auxiliaire.

Une perspective fondée sur un modèle du calage sur un modèle peut être élaborée comme il suit. Soit $\hat{\eta}$ un estimateur « efficace sous le modèle » de η dans (11), par exemple son estimateur par le maximum de vraisemblance (MV), avec les valeurs prédites $h(\mathbf{x}_j;\ \hat{\eta})$. En général, ces valeurs prédites ne sont pas sans biais. Elles sont également corrélées. Toutefois, il persiste une relation systématique entre les valeurs réelles de Y et les valeurs prédites correspondante que nous pouvons approximer. Bien que rien ne nous empêche d'examiner des approximations plus complexes, un modèle linéaire pour la relation entre les valeurs

de population y_j et les valeurs prédites $\hat{y}_j = h(\mathbf{x}_j\,;\hat{\eta})$ semble être un point de départ raisonnable. Nous remplaçons par conséquent le modèle non linéaire (11) par le modèle linéaire

$$E(y_j|\,\hat{y}_j) = \alpha_0 + \alpha_1\hat{y}_j$$

et (13)

$$\mathrm{Cov}(y_j,\, y_k|\,\hat{y}_j,\hat{y}_k) = \omega_{jk}.$$

Nous désignons (13) comme étant le modèle « sous valeurs prédites » correspondant à (11). Soit $\mathbf{J}_U$ la « matrice de plan » de population sous le modèle (13), c'est-à-dire $\mathbf{J}_U = [\mathbf{1}_U\ \hat{\mathbf{y}}_U]$, où $\mathbf{1}_U$ désigne le vecteur unitaire de taille N et $\hat{\mathbf{y}}_U = (\hat{y}_j\,;\ j = 1, \ldots, N)$, et posons que $\Omega_U = [\omega_{jk}\,;\ j = 1, \ldots, N\,;\ k = 1, \ldots, N]$. Nous pouvons maintenant partitionner $\mathbf{J}_U$ et Ω_U en fonction des unités échantillonnées (s) et non échantillonnées (r) comme il suit

$$\mathbf{J}_U = \begin{bmatrix} \mathbf{J}_s \\ \mathbf{J}_r \end{bmatrix}$$

et

$$\Omega_U = \begin{bmatrix} \Omega_{ss} & \Omega_{sr} \\ \Omega_{rs} & \Omega_{rr} \end{bmatrix},$$

et donc écrire les poids qui définissent l'estimateur BLUP de t_{Uy} sous (13). Il s'agit des poids calés sur un modèle fondés sur un modèle (cmfm)

$$\mathbf{w}^{cmfm} = (w_j^{cmfm}\,;\ j \in s)$$
$$= \mathbf{1}_s + \mathbf{H}'_{cm}(\mathbf{J}'_U\mathbf{1}_U - \mathbf{J}'_s\mathbf{1}_s) + (\mathbf{I}_s - \mathbf{H}'_{cm}\mathbf{J}'_s)\Omega_{ss}^{-1}\Omega_{sr}\mathbf{1}_r \quad (14)$$

où $\mathbf{H}_{cm} = (\mathbf{J}'_s\Omega_{ss}^{-1}\mathbf{J}_s)^{-1}\mathbf{J}'_s\Omega_{ss}^{-1}$. De toute évidence, ces poids sont calés sur un modèle puisque $\sum_{j\in s} w_j^{cmfm} = N$ et $\sum_{j\in s} w_j^{cmfm}\hat{y}_j = \sum_{j\in U}\hat{y}_j$. Cependant, contrairement aux poids EBLUP du modèle linéaire (2), ils *ne sont pas* calés sur $\mathbf{X}$. En pratique, les composantes de Ω_U sont inconnues et doivent être estimées. Quand ces estimations sont introduites par substitution dans (14), nous obtenons la version empirique $\mathbf{w}^{cmfme}$ de ces poids calés sur un modèle.

5. Pondération calée sur un modèle pour l'estimation sur petits domaines

Nous utilisons maintenant le calage sur un modèle fondé sur le modèle linéaire mixte à échelle logarithmique (9) pour obtenir les poids de sondage à utiliser dans l'estimateur DFM (7). L'exposé de la section précédente montre que, pour cela, nous devons d'abord spécifier un modèle sous valeurs prédites (13) pour Y basé sur (9), c'est-à-dire que nous devons calculer les valeurs prédites appropriées $\hat{y}_{ij}$ ainsi que les estimations $\hat{\omega}_{ijk}$ de $\omega_{ijk} = \mathrm{Cov}(y_{ij},\, y_{ik}|\, x_{ij},\, x_{ik},\, \mathbf{g}_{ij},\, \mathbf{g}_{ik})$ sous le modèle (9). Les poids de sondage à utiliser dans l'estimateur DFM (7) sont alors donnés par (14).

Une méthode simple de définition de valeurs prédites $\hat{y}_{ij}$ sous (9) consiste à utiliser les estimations des paramètres calculées sous ce modèle pour obtenir les valeurs prédites sur l'échelle logarithmique, qui sont ensuite rétrotransformées. Malheureusement, il est bien connu que cette approche est biaisée. Par conséquent, nous développons les moments de premier et de deuxième ordre d'un modèle sous valeurs prédites corrigé du biais et fondé sur (9) approprié. Soit $\mathbf{x}_s$ et $\mathbf{g}_s$ les valeurs d'échantillon de x_{ij} et $\mathbf{g}_{ij}$, respectivement. Sous (9),

$$E(y_{ij}|\, x_{ij}, \mathbf{g}_{ij}) = E\{e^{l_{ij}}|\, x_{ij},\, \mathbf{g}_{ij}\} = e^{\phi_{ij}+v_{ijj}/2}$$
$$\neq E(e^{\hat{\phi}_{ij}+\hat{v}_{ijj}/2}|\mathbf{x}_s,\, \mathbf{g}_s) = E(\hat{y}_{ij}|\, x_{ij},\, \mathbf{g}_{ij})$$

de sorte que la correction habituelle du biais s'appuyant sur le fait que la loi conditionnelle de y_{ij} est lognormale est inadéquate. Soit $\hat{\eta}_{ij} = (\hat{\beta}, \hat{v}_{ijj})'$ une estimation de $\eta_{ij} = (\beta,\, v_{ijj})'$ telle que $E(\hat{\eta}_{ij} - \eta_{ij}) \approx 0$ pour n grand. Posons que $z(\eta_{ij}) = e^{\phi_{ij}+v_{ijj}/2}$. En utilisant une approximation par développement en série de Taylor de deuxième ordre, nous pouvons écrire

$$z(\hat{\eta}_{ij}) \approx z(\eta_{ij}) + (\hat{\eta}_{ij} - \eta_{ij})'z^{(1)}(\eta_{ij}) + \frac{1}{2}(\hat{\eta}_{ij} - \eta_{ij})'z^{(2)}(\eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})$$

et donc

$$E\{z(\hat{\eta}_{ij})\} \approx z(\eta_{ij}) + \frac{1}{2}tr[E\{z^{(2)}(\eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})'\}].$$

Ici

$$z^{(1)}(\eta_{ij}) = \left(\mathbf{d}'_{ij}e^{\phi_{ij}+v_{ijj}/2}\ \ \frac{1}{2}e^{\phi_{ij}+v_{ijj}/2}\right)'$$

et

$$z^{(2)}(\eta_{ij}) = \begin{pmatrix} \mathbf{d}_{ij}\mathbf{d}'_{ij}e^{\phi_{ij}+v_{ijj}/2} & \frac{1}{2}\mathbf{d}_{ij}e^{\phi_{ij}+v_{ijj}/2} \\ \frac{1}{2}\mathbf{d}'_{ij}e^{\phi_{ij}+v_{ijj}/2} & \frac{1}{4}e^{\phi_{ij}+v_{ijj}/2} \end{pmatrix}$$

sont le vecteur et la matrice, respectivement, contenant les dérivées de premier et de deuxième ordre de $z(\eta_{ij})$ par rapport à η_{ij}. Puisque la covariance asymptotique entre les estimateurs MV (ou MVR) des composantes fixes et des composantes de variance d'un modèle linéaire mixte est

nulle (McCulloch et Searle 2001, chapitre 2, pages 40-45), la covariance entre $\hat{\beta}$ et $\hat{v}_{ijj}$ sera négligeable. Il s'ensuit que

$$\begin{aligned} & tr[E\{z^{(2)}(\eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})'\}] \\ & \quad = tr[z^{(2)}(\eta_{ij})E\{(\hat{\eta}_{ij} - \eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})'\}] \\ & \quad \approx e^{\phi_{ij} + \frac{v_{ijj}}{2}}\left[\mathbf{d}_{ij}'\left(\sum_g \mathbf{d}_{gs}'\hat{\mathbf{v}}_{gss}^{-1}\mathbf{d}_{gs}\right)^{-1}\mathbf{d}_{ij} + \frac{1}{4}\mathrm{Var}(\hat{v}_{ijj})\right] \\ & \quad = E(y_{ij}|\, x_{ij}, \mathbf{g}_{ij})\left[\hat{a}_{ij} + \frac{1}{4}\mathrm{Var}(\hat{v}_{ijj})\right] \end{aligned}$$

où $\hat{a}_{ijj} = \mathbf{d}_{ij}'\hat{V}(\hat{\beta})\mathbf{d}_{ij}$ et $\hat{V}(\hat{\beta}) = (\sum_i \mathbf{d}_{is}'\hat{\mathbf{v}}_{iss}^{-1}\mathbf{d}_{is})^{-1}$ est l'estimateur habituel de $\mathrm{Var}(\hat{\beta})$. Nos valeurs prédites sont par conséquent définies par l'estimateur corrigé du biais de deuxième ordre de $E(y_{ij}|x_{ij}, \mathbf{g}_{ij})$,

$$\hat{y}_{ij} = h(\mathbf{d}_{ij}\,;\, \hat{\eta}_{ij}) = \hat{k}_{ij}^{-1}e^{\hat{\phi}_{ij} + \hat{v}_{ijj}/2} \qquad (15)$$

où

$$\hat{k}_{ij} = 1 + \frac{1}{2}\left\{\hat{a}_{ijj} + \frac{1}{4}\hat{V}(\hat{v}_{ijj})\right\}$$

et $\hat{V}(\hat{v}_{ijj})$ est la variance asymptotique estimée de $\hat{v}_{ijj}$. Sous estimation du MV et du MVR des composantes de la variance de (9), cette variance asymptotique estimée s'obtient à partir de l'inverse de la matrice d'information pertinente. Notons que la correction du biais de Karlberg (2000a) est un cas particulier de (15).

Afin d'utiliser (14) pour définir les poids de sondage calés sur un modèle fondé sur un modèle, nous avons également besoin des estimations des moments de deuxième ordre des valeurs de population de Y sachant les valeurs prédites. Les moments conditionnels ω_{ijk} représentent une approximation de premier ordre de ces moments. En particulier, si les effets aléatoires sont normaux,

$$\omega_{ijk} = e^{(\phi_{ij} + \phi_{ik}) + (v_{ijj} + v_{ikk})/2}(e^{v_{ijk}} - 1) \qquad (16)$$

Nous obtenons notre estimation $\hat{\omega}_{ijk}$ de ω_{ijk} en substituant $\hat{\phi}_{ij}$ et $\hat{v}_{ijk}$ à ϕ_{ij} et v_{ijk} dans (16).

Les poids calés sur un modèle fondés sur un modèle empiriques (14) correspondants au modèle sous valeurs prédites défini par (15) et (16) sont

$$\begin{aligned} \mathbf{w}^{cmfme} &= (w_{ij}^{cmfme}\,;\; j \in s_i\,;\; i = 1, \ldots, D) \\ &= \mathbf{1}_s + \hat{\mathbf{H}}_{cm}'(\mathbf{J}_U'\mathbf{1}_U - \mathbf{J}_s'\mathbf{1}_s) \\ &\quad + (\mathbf{I}_s - \hat{\mathbf{H}}_{cm}'\mathbf{J}_s')\hat{\Omega}_{ss}^{-1}\hat{\Omega}_{sr}\mathbf{1}_r. \end{aligned} \qquad (17)$$

Ici, $\mathbf{J}_U = [\mathbf{1}_U\ \hat{\mathbf{y}}_U]$, donc

$$\mathbf{J}_U'\mathbf{1}_U - \mathbf{J}_s'\mathbf{1}_s = \begin{pmatrix} N - n \\ \sum_i \sum_{j \in r_i} \hat{y}_{ij} \end{pmatrix},$$

et $\hat{\mathbf{H}}_{mc} = (\mathbf{J}_s'\hat{\Omega}_{ss}^{-1}\mathbf{J}_s)^{-1}\mathbf{J}_s'\hat{\Omega}_{ss}^{-1}$. En outre $\hat{\Omega}_{ss} = \mathrm{diag}\{\hat{\Omega}_{iss}\,;\, i = 1, \ldots, D\}$ et $\hat{\Omega}_{sr} = \mathrm{diag}\{\hat{\Omega}_{isr}\,;\, i = 1, \ldots, D\}$, où $\hat{\Omega}_{iss}$ et $\hat{\Omega}_{isr}$ sont définis par la décomposition échantillon/hors échantillon de $\hat{\Omega}_i$. Par exemple, quand (9) correspond à une spécification d'ordonnées à l'origine aléatoires, $\hat{v}_{ijk} = \hat{\sigma}_u^2 + \hat{\sigma}_e^2 I(j = k)$ et les composantes de $\hat{\Omega}_i$ sont donc

$$\hat{\omega}_{ijk} = e^{\hat{\phi}_{ij} + \hat{\phi}_{ik} + \hat{\sigma}_u^2 + \hat{\sigma}_e^2}[e^{\hat{\sigma}_u^2}\{1 + I(j = k)(e^{\hat{\sigma}_e^2} - 1)\} - 1].$$

Jusqu'à présent, notre exposé s'est appuyé sur l'hypothèse de normalité des effets aléatoires sur l'échelle logarithmique. Toutefois, il n'existe aucune bonne raison (outre la commodité) de supposer que, dans le cas de données asymétriques, ces effets aléatoires de domaine doivent être normaux. Une autre option, étant donné un effet de domaine scalaire dans (9), consiste à supposer que les effets aléatoires dans ce modèle sont tirés de la famille de lois *gamma*. Partant des propriétés de cette loi et en utilisant les développements des fonctions binominales et exponentielles (en ignorant les termes d'ordre élevé), nous pouvons montrer que $E(y_{ij}|\, x_{ij}, \mathbf{g}_{ij}) \approx e^{\phi_{ij} + v_{ijj}/2} = z(\eta_{ij})$, comme dans le cas normal. Cela indique qu'un estimateur DFM basé sur les poids calés sur un modèle fondés sur un modèle (17) devrait être robuste en ce qui concerne la distribution des effets aléatoires dans (9).

Enfin, nous considérons la définition de l'estimateur DFM proprement dit. Comme nous l'avons mentionné à la section 2, cet estimateur est simplement la moyenne pondérée des valeurs d'échantillon de Y dans un domaine. Cependant, l'utilisation de cette moyenne pondérée présuppose que les poids sont raisonnablement proches d'être calés localement sur N', autrement dit que, si nous effectuons la sommation sur les unités échantillonnées dans le domaine i, nous obtenons une valeur qui ne diffère pas trop de la taille de population réelle du petit domaine N_i. Cette propriété est habituellement vérifiée si les poids sont les poids EBLUP pour le total (6) défini par un modèle linéaire mixte pour Y. Elle ne l'est pas nécessairement pour les poids calés sur un modèle fondés sur un modèle (17). Par conséquent, nous considérons deux spécifications pour l'estimateur DFM sachant ces poids. La première, que nous appelons « spécification de Hájek », est simplement la moyenne pondérée (7), les poids étant définis par (17). La deuxième, que nous appelons « spécification d'Horvitz-Thompson », remplace le dénominateur de (7) par la valeur réelle de N_i. Autrement dit, les deux types d'estimateurs DFM sous la pondération calée sur un modèle fondée sur un modèle que nous considérons sont

$$\hat{m}_{iy}^{\text{HJ-DFMTr}} = \left\{\sum_{j\in s_i} w_{ij}^{cmfme}\right\}^{-1} \sum_{j\in s_i} w_{ij}^{cmfme} y_{ij} \qquad (18)$$

et

$$\hat{m}_{iy}^{\text{HT-DFMTr}} = N_i^{-1} \sum_{j\in s_i} w_{ij}^{cmfme} y_{ij}. \qquad (19)$$

Nous pouvons aussi adopter une approche fondée sur la prédiction pour obtenir un prédicteur indirect de rechange de la moyenne de petit domaine sous le modèle transformé logarithmiquement (9). Notre approche étend celle de Karlberg (2000a). Dans ces conditions, en supposant que le modèle (9) est vérifié, nous prédisons chaque valeur de Y hors échantillon dans le petit domaine i, puis nous additionnons ces prédictions. Notons que nous devons corriger le biais après la rétrotransformation à l'échelle originale quand nous calculons ces valeurs prédites pour les valeurs de Y hors échantillon. Sous le modèle (9), le prédicteur empirique résultant de la moyenne m_{iy} de Y dans le domaine i (désigné PETr) peut être défini comme étant

$$\hat{m}_{iy}^{\text{PETr}} = N_i^{-1}\left\{\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \hat{y}_{ij}\right\}, \qquad (20)$$

où $\hat{y}_{ij}$ est donné par (15).

L'estimation de l'EQM de (18) et de (19) est effectuée de la façon habituelle pour les estimateurs DFM, c'est-à-dire suivant l'approche d'estimation de l'EQM décrite à la section 2. L'estimation de l'EQM de (20) n'est pas facile, parce que ce prédicteur est une fonction non linéaire des valeurs de Y. Nous ne poursuivons pas l'examen de cette question dans le présent article.

6. Une évaluation empirique

À la présente section, nous présentons les résultats empiriques d'une étude comparative des performances de cinq méthodes différentes d'estimations sur petits domaines. Ces méthodes comprennent les deux estimateurs directs fondés sur un modèle (DFM) « basé sur une transformation » (18) et (19), et s'appuient tous les deux sur les poids calés sur un modèle fondés sur un modèle (17) et sont désignés par HJ-DFMTr et HT-DFMTr, respectivement ; le prédicteur (20) fondé sur une transformation logarithmique sous le modèle (9), désigné par PEtr, l'estimateur DFM « classique » (7) basé sur le modèle linéaire mixte (3) et les poids empiriques EBLUP pour le total (6), que nous désignons par HJ-DFMLin afin de mettre l'accent sur le fait qu'il s'agit d'une moyenne pondérée de type Hájek fondée sur les poids calculés sous un modèle linéaire mixte, ainsi que l'estimateur EBLUP (8) calculé sous le même modèle linéaire mixte, que nous désignons par HT-EBLUPLin. Notons que les EQM pour les trois estimateurs DFM ont été estimées en utilisant la méthode décrite à la section 2, tandis que l'EQM de HT-EBLUPLin a été estimée en utilisant la méthode décrite dans Prasad et Rao (1990). Il convient de souligner que nous n'avons pas considéré l'estimation de l'EQM de PETr.

Nos résultats empiriques sont fondés sur deux types d'études par simulation. Le premier comprend l'utilisation d'une simulation fondée sur un modèle pour générer une population artificielle et des données d'échantillon. Autrement dit, à chaque simulation, nous avons d'abord généré des données de population sous le modèle, puis tiré un échantillon unique de cette population simulée par échantillonnage aléatoire simple stratifié sans remise en utilisant les petits domaines comme strates. Nous avons ensuite utilisé ces données pour comparer les propriétés des divers estimateurs. À la section 6.1, nous présentons les résultats de ces simulations fondées sur un modèle. Nous avons exécuté deux ensembles de ces simulations. Dans le premier (ensemble A), nous avons étudié la performance de ces estimateurs étant donné les données de population générées en utilisant le modèle linéaire mixte à échelle logarithmique (9). Dans le deuxième ensemble de simulations (ensemble B), nous avons examiné la robustesse de ces estimateurs à l'erreur de spécification de ce modèle. Le deuxième type d'études par simulation était fondé sur le plan de sondage. À la section 6.2, nous décrivons les simulations fondées sur le plan de sondage. Ici, nous avons évalué les estimateurs dans le contexte de l'échantillonnage répété à partir d'une population réelle en utilisant des méthodes d'échantillonnage réalistes. Autrement dit, nous avons d'abord utilisé des données d'enquête réelles pour simuler une population, puis nous avons échantillonné cette population fixe à plusieurs reprises conformément à un plan préspécifié. En particulier, nous avons utilisé un plan d'échantillonnage aléatoire stratifié dans lequel les strates correspondaient aux petits domaines d'intérêt et la répartition de l'échantillon entre les strates correspondait aux tailles d'échantillon de petit domaine dans les ensembles de données originaux.

Nous avons calculé quatre mesures de la performance des estimateurs en nous servant des diverses estimations produites dans les études par simulation. Ces mesures sont le biais relatif (BR) et la racine carrée de l'erreur quadratique moyenne (RCEQM) relative de ces estimations, ainsi que le taux de couverture et la largeur moyenne des intervalles de confiance à 95 % nominaux qui sont fondés sur ces taux. Aux tableaux 2 à 4, ces mesures sont présentées sous forme de moyennes sur les petits domaines d'intérêt.

6.1 L'étude par simulation fondée sur un modèle

Les simulations fondées sur un modèle sont utilisées fréquemment pour illustrer la sensibilité d'une méthode d'estimation à la variation des hypothèses au sujet de la structure de la population d'intérêt. Ici, nous avons fixé la taille de la population à $N = 15\,000$ et avons produit aléatoirement les tailles de population de petit domaine N_i, $i = 1, ..., D = 30$ de sorte que $\sum_i N_i = N$. Nous avons utilisé une taille globale d'échantillon de $n = 600$ avec un ensemble de tailles d'échantillon de petit domaine tel que ces tailles étaient proportionnelles aux tailles de population de petit domaine correspondantes. Ces tailles de population et d'échantillon propres aux domaines ont été maintenues fixes dans toutes les simulations. Les tailles de population et d'échantillon sont présentées au tableau 1a.

Tableau 1a
Tailles de population (N_i) et d'échantillon (n_i) propres au domaine pour la simulation fondée sur un modèle

Domaine	1	2	3	4	5	6	7	8	9	10
N_i	525	538	510	468	526	484	516	458	529	518
n_i	21	22	20	19	21	19	21	19	21	21
Domaine	11	12	13	14	15	16	17	18	19	20
N_i	502	524	509	484	487	459	542	498	512	500
n_i	20	21	20	19	19	18	22	20	20	20
Domaine	21	22	23	24	25	26	27	28	29	30
N_i	497	492	443	506	513	536	506	495	463	460
n_i	20	20	18	20	21	21	20	20	19	18

Dans l'ensemble A de nos simulations fondées sur un modèle, les valeurs de population y_{ij} ont été produites en utilisant le modèle multiplicatif $y_{ij} = 5,0 x_{ij}^{\beta} u_i e_{ij} (j = 1,..., N_i ; i = 1, ..., 30)$, puis des échantillons aléatoires ont été tirés de chaque petit domaine. Ici, les valeurs de x_{ij} ont été tirées indépendamment de la loi log normale $\log(x_{ij}) \sim N(6, \sigma_x^2)$, avec les effets individuels et les effets de domaine tirés indépendamment en tant que $\log(e_{ij}) \sim N(0, \sigma_e^2)$ et $\log(u_i) \sim N(0, \sigma_u^2)$, respectivement. Le valeurs de population de x ont été générées à nouveau dans chaque simulation. En particulier, dans chaque simulation, nous avons d'abord généré les valeurs de x pour une population de taille N, puis attribué aléatoirement ces valeurs à divers domaines de tailles N_i. Les valeurs de σ_e et σ_u ont été choisies de manière que la corrélation intra-domaine dans la population varie entre 0,20 et 0,25. Le tableau 1b montre les six ensembles distincts de valeurs des paramètres qui ont été utilisés dans l'ensemble A. Ces divers ensembles ont assuré que les populations simulées contiennent une grande gamme de variations. Pour chaque population générée et pour chaque domaine i, nous avons sélectionné un échantillon aléatoire simple (sans remise) de taille n_i, ce qui a donné une taille globale d'échantillon de $n = 600$. Les valeurs d'échantillon de y et les valeurs de population de x obtenues dans chaque simulation ont ensuite été utilisées pour estimer les moyennes de petit domaine. Autrement dit, en utilisant les données d'échantillon dans chaque cas, nous avons estimé les valeurs des paramètres en utilisant la fonction *lme* dans R (Bates et Pinheiro 1998), puis nous avons calculé les estimations des moyennes de petit domaine, ainsi que les intervalles de confiance à 95 % nominaux appropriés. Le processus de génération des données de population et d'échantillon, d'estimation des paramètres et de calcul des estimations sur petits domaines a été répété indépendamment 1 000 fois. Les résultats de cette partie de l'étude par simulation sont présentés au tableau 2.

Tableau 1b
Spécifications des populations pour l'ensemble A de simulations fondées sur un modèle

Ensemble de paramètres	β	σ_u	σ_e	σ_x
1	0,5	0,30	0,50	3,00
2	0,8	0,35	0,60	2,50
3	1,0	0,40	0,70	2,25
4	1,3	0,45	0,80	1,75
5	1,5	0,50	0,90	1,50
6	2,0	0,60	1,00	1,20

Dans l'ensemble B de simulations fondées sur un modèle, les données de population ont été générées en utilisant le modèle $y_{ij} = 5,0 x_{ij} [\exp(\log^2 (x_{ij}))]^{\gamma} u_i e_{ij}$. Ici, les effets individuels e_{ij} et les effets de domaine u_i ont été tirés indépendamment en tant que $\log(e_{ij}) \sim N(0, 1)$ et $\log(u_i) \sim N(0, 0,25)$, respectivement, tandis que les valeurs des covariables x_{ij} ont été tirées en tant que $\log(x_{ij}) \sim N(3, 0,04)$. Cinq valeurs différentes du paramètre γ (-1,0, -0,5, 0,0, 0,5, 1,0) ont été examinées, ce qui a produit des données de population ayant divers degrés de courbure. Tous les autres aspects de ces simulations, y compris les estimateurs pris en considération, étaient les même que pour l'ensemble A. Le tableau 3 donne les résultats pour cette composante de l'étude par simulation.

Tableau 2
Biais relatif moyen (BRM), moyenne des RCEQM relatives (MRCEQMR), taux de couverture moyen (TCM) et largeur d'intervalle moyenne (LIM) pour l'ensemble A de simulations fondées sur un modèle

Critère	Estimateur	Ensemble de paramètres					
		1	2	3	4	5	6
BRM, %	HJ-DFMTr	-82,68	-95,02	-98,08	-98,50	-98,29	-99,00
	HT-DFMTr	0,09	0,10	-0,14	-0,25	-0,03	0,04
	PETr	0,08	0,09	-0,18	-0,48	-0,05	0,01
	HJ-DFMLin	12,01	4,09	-1,35	-5,54	-6,60	-9,88
	HT-EBLUPLin	13,39	5,18	-0,67	-5,24	-6,41	-9,67
MRCEQMR	HJ-DFMTr	4,80	1,39	1,25	1,44	1,42	1,62
	HT-DFMTr	0,15	0,26	0,45	0,64	0,66	0,91
	PETr	0,30	0,41	0,58	0,80	0,81	1,09
	HJ-DFMLin	1,11	1,41	1,85	1,99	2,06	2,69
	HT-EBLUPLin	0,79	0,54	0,64	0,92	0,93	1,31
TCM	HJ-DFMTr	0,99	0,98	0,97	0,95	0,94	0,92
	HT-DFMTr	0,94	0,91	0,89	0,89	0,89	0,88
	HJ-DFMLin	0,87	0,85	0,85	0,88	0,88	0,87
	HT-EBLUPLin	0,85	0,85	0,86	0,87	0,88	0,87
LIM	HJ-DFMTr	1 592	22 688	140 452	52×10^4	35×10^5	44×10^6
	HT-DFMTr	219	4 414	34 105	14×10^4	11×10^5	15×10^6
	HJ-DFMLin	1 005	19 232	139 420	57×10^4	41×10^5	56×10^6
	HT-EBLUPLin	382	7 099	57 039	26×10^4	21×10^5	32×10^6

Tableau 3
Biais relatif moyen (BRM), moyenne des RCEQM relatives (MRCEQMR), taux de couverture moyen (TCM) et largeur d'intervalle moyenne (LIM) pour l'ensemble B de simulations fondées sur un modèle

Critère	Estimateur	$\gamma = -1,0$	$\gamma = -0,5$	$\gamma = 0,0$	$\gamma = 0,5$	$\gamma = 1,0$
BRM, %	HT-DFMTr	4,92	0,66	0,14	-1,50	-8,75
	HJ-DFMLin	-0,21	0,04	0,12	0,16	-0,85
	HT-EBLUPLin	-0,19	0,04	0,13	0,17	-0,77
MRCEQMR	HT-DFMTr	0,38	0,35	0,33	0,37	0,41
	HJ-DFMLin	0,56	0,36	0,34	0,53	1,20
	HT-EBLUPLin	0,38	0,30	0,29	0,36	0,56
TCM	HT-DFMTr	0,94	0,92	0,92	0,91	0,87
	HJ-DFMLin	0,91	0,92	0,92	0,92	0,90
	HT-EBLUPLin	0,93	0,94	0,94	0,93	0,92
LIM	HT-DFMTr	0,04	2,50	211	29 070	5×10^6
	HJ-DFMLin	0,06	2,70	214	38 660	13×10^6
	HT-EBLUPLin	0,05	2,60	214	33 442	10×10^6

6.2 L'étude par simulation fondée sur le plan de sondage

Pour cette étude, nous nous sommes servis de la même population et des mêmes échantillons que dans les études par simulation décrites dans Chandra et Chambers (2005) et dans Chambers et Tzavidis (2006), qui étaient fondées sur des données obtenues auprès d'un échantillon de 1 652 entreprises agricoles qui avaient participé à l'Australian Agricultural and Grazing Industries Survey (AAGIS). Une population réaliste de 81 982 exploitations agricoles a été définie par échantillonnage avec remise à partir de l'échantillon original de 1 652 exploitations agricoles avec probabilités proportionnelles aux poids de sondage, qui étaient tous strictement supérieurs à 1. En tout, 1 000 échantillons indépendants, chacun de taille $n = 1\,652$, ont été tirés de cette population fixe par échantillonnage aléatoire simple sans remise dans des strates définies par les 29 régions agricoles australiennes représentées dans l'échantillon de l'AAGIS. Ces régions sont les petits domaines d'intérêt. Les tailles des échantillons régionaux ont été fixées à la même valeur que dans l'échantillon original, variant de 6 à 117, ce qui permet d'évaluer la performance des diverses méthodes d'estimation sur une gamme de tailles d'échantillon de petit domaine réalistes. Notons que les fractions d'échantillonnage dans ces strates variaient également de manière disproportionnée, allant de 0,70 % à 15,87 %. L'objectif était

d'estimer les coûts agricoles annuels moyens (coûts décaissés totaux (CDT), mesurés en $A) dans chaque région en utilisant la taille de l'exploitation agricole (hectares) comme variable auxiliaire. Nous nous sommes servis de la même spécification du modèle mixte que dans Chandra et Chambers (2005). Cette spécification comprend un terme d'interaction (zone par taille) dans les effets fixes et la spécification d'une pente aléatoire pour l'effet de domaine. Sous sa forme linéaire, le modèle n'est pas très bien ajusté aux données d'échantillon de l'AAGIS. L'adéquation du modèle est améliorée (quoique marginalement) quand on utilise une spécification linéaire à échelle logarithmique. Nos résultats sont résumés au tableau 4.

6.3 Discussion des résultats des simulations

Dans le tableau 2, les très grandes valeurs du biais relatif moyen de HJ-DFMTr sous la pondération calée sur un modèle fondée sur un modèle est l'aspect le plus frappant. Les deux estimateurs donnant les meilleurs résultats en ce qui concerne le biais relatif sont HT-DFMTr, qui est fondé sur les mêmes pondérations que HJ-DFMTr, et PETr. Un examen en vue de déterminer la raison de la performance médiocre de HJ-DFMTr a révélé que la sommation des poids calés sur un modèle fondés sur un modèle (17) à l'intérieur des petits domaines produisait des estimations extrêmement variables des tailles de population de petit domaine, ce qui sous-entend que ces poids ne peuvent pas être considérés comme « polyvalents » – ils fonctionnent bien quand ils sont utilisés avec des variables qui sont raisonnablement corrélées avec la variable qui définit le modèle sous valeurs prédites, mais peut donner de mauvais résultats avec d'autres variables moins bien corrélées (par exemple la variable indicatrice d'inclusion du petit domaine). Nous constatons en outre que ce problème ne se pose pas avec les poids EBLUP empiriques « classiques » pour le total (6), car HJ-DFMLin produit des résultats cohérents pour les six scénarios examinés dans l'ensemble A de l'étude par simulation. À partir de maintenant, nous concentrons par conséquent notre discussion sur les quatre estimateurs HT-DFMTr, PEtr, HJ-DFMLin et HT-EBLUPLin.

Le tableau 2 montre que les biais relatifs moyens et les moyennes des RCEQM relatives pour HT-DFMTr sont systématiquement inférieurs à ceux produits par HJ-DFMLin et HT-EBLUPLin. Le biais relatif moyen de HT-DFMTr et PETr sont comparables. Cependant, les moyennes des RCEQM relatives de HT-DFMTr sont systématiquement plus faibles que pour PETr. En outre, les taux de couverture moyens et les largeurs d'intervalle moyennes pour HT0-DFMTr sont meilleurs que ceux produits par HJ-DFMLin et HT-EBLUPLin. Par comparaison, pour le même ordre de biais relatif, la RCEQM relative de HT-EBLUPLin est plus faible que celle de HJ-DFMLin, et, bien que les deux estimateurs produisent des taux de couverture fort semblables, les intervalles de confiance générés par le HT-EBLUPLin ont tendance à avoir une largeur moyenne plus faible que ceux générés par HJ-DFMLin.

Les graphiques de la figure 1 représentent les mesures de performance selon la région obtenues pour ces quatre estimateurs pour l'ensemble A de simulations. Ces graphiques montrent que les valeurs du biais relatif et de la RCEQM relative générées par HT-DFMTr sont plus faibles que les valeurs correspondantes pour HJ-DFMLin et HT-EBLUPLin dans toutes les régions. Alors que les valeurs du biais relatif sont presque identiques, l'estimateur HT-DFMTr produit des valeurs plus faibles de la RCEQM relative que le PETr dans toutes les régions. En outre, le biais relatif et la RCEQM relative de HJ-DFMLin et HT-EBLUPLin augmentent à mesure que s'accroît la non-linéarité des données (c'est-à-dire quand nous passons de l'ensemble de paramètres 1 à l'ensemble de paramètres 6). Nous constatons également que HT-DFMTr produit dans toutes les régions de meilleurs taux de couverture que HT-EBLUPLin et HJ-DFMLin.

Tableau 4
Biais relatif moyen (BRM), moyenne des RCEQM relatives (MRCEQMR) et taux de couverture moyen (TCM) pour les simulations fondées sur le plan de sondage en utilisant les données de l'AAGIS. Les erreurs-types de simulation du BRM et de la MRCEQMR sont données entre parenthèses

Critère	Estimateur	Moyenne de 29 régions	Moyenne de 28 régions
BRM, %	HT-DFMTr	1,96 (0,20)	1,92 (0,11)
	HJ-DFMLin	-2,13 (0,15)	-2,21 (0,12)
	HT-EBLUPLin	2,98 (0,18)	3,36 (0,16)
	Pseudo-EBLUP	4,01 (0,22)	4,41 (0,20)
	JL	1,89 (0,19)	2,23 (0,17)
MRCEQMR, %	HT-DFMTr	21,93 (4,47)	17,41 (1,18)
	HJ-DFMLin	20,15 (3,80)	16,91 (2,20)
	HT-EBLUPLin	19,87 (1,78)	19,30 (1,63)
	Pseudo-EBLUP	22,42 (2,52)	21,95 (2,46)
	JL	20,97 (1,48)	20,48 (1,31)
TCM	HT-DFMTr	0,89	0,92
	HJ-DFMLin	0,93	0,95
	HT-EBLUPLin	0,85	0,85

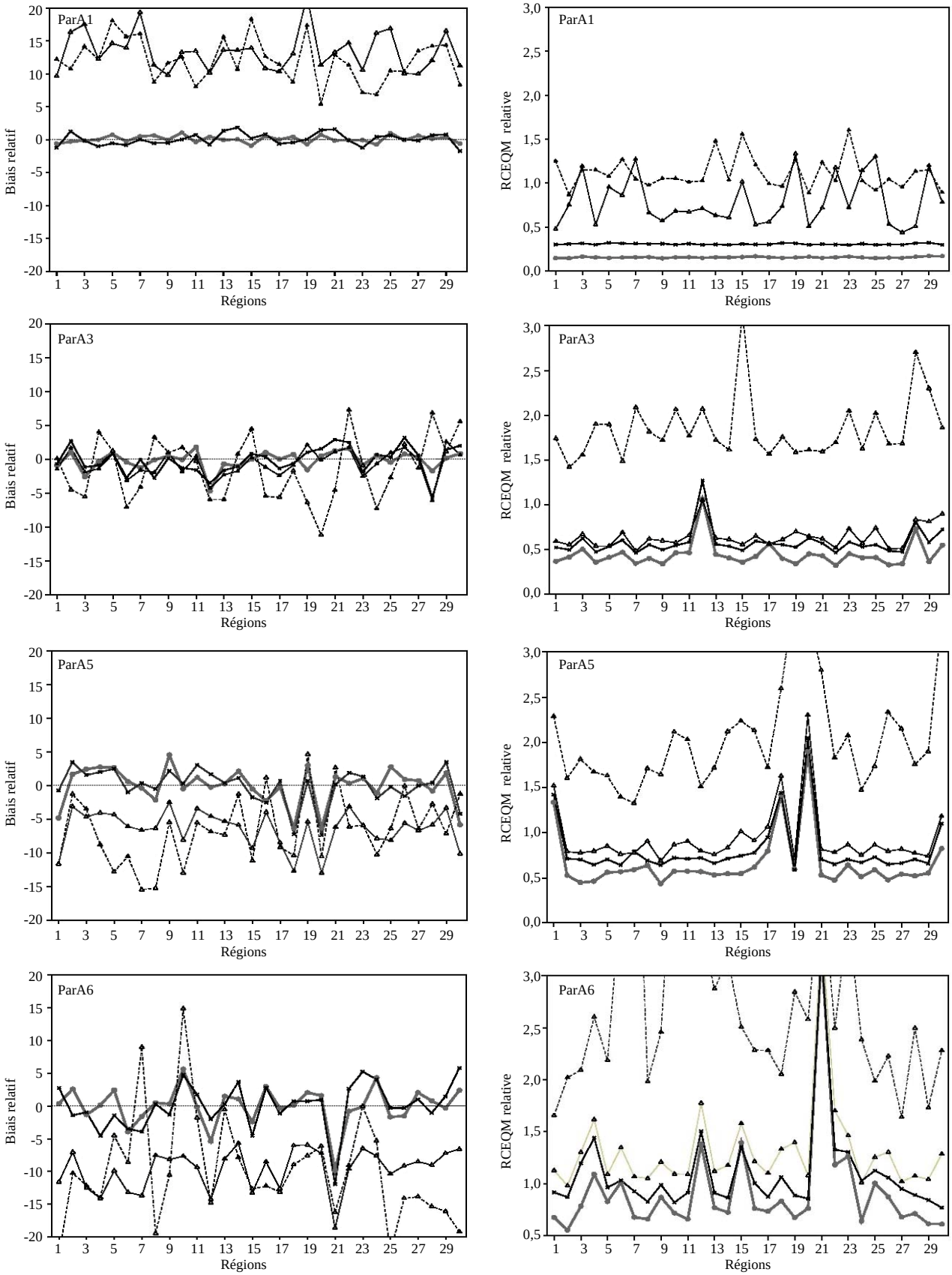

Figure 1 Résultats par région pour HT-DFMTr (trait plein, ●), PETr (trait épais, ×), HT-EBLUPLin (trait fin, Δ) et HJ-DFMLin (trait pointillé, Δ) sous les ensembles de paramètres 1 (ParA1), 3 (ParA3), 5 (ParA5) et 6 (ParA6). La colonne de gauche donne le biais relatif (%) et la colonne de droite, la RCEQM relative

Dans l'ensemble, les résultats montrent que, quand le modèle de population sous-jacent n'est pas linéaire, l'utilisation des estimateurs DFM de type HT pour calculer les moyennes de petit domaine fondés sur les poids calés sur un modèle (17) produit des gains importants comparativement aux estimateurs fondés sur un modèle linéaire mixte classiques, tels que HJ-DFMLin et HT-EBLUPLin. Les résultats montrent aussi que l'estimateur indirect HT-EBLUPLin donne de relativement meilleurs résultats que l'estimateur direct HJ-DFMLin dans ces situations. Le prédicteur indirect PETr fondé sur le modèle à transformation logarithmique (9) a de bonnes propriétés en ce qui concerne le biais relatif, mais est moins efficace que l'estimateur DFM sous le même modèle.

Pour l'ensemble B de simulations fondées sur un modèle, nous avons étudié la robustesse de l'estimation directe calée sur un modèle fondée sur un modèle à l'erreur de spécification du modèle non linéaire. Les résultats du tableau 3 montrent que, dans ce cas, le biais généré par HT-DFMTr augmente à mesure que le modèle non linéaire réel s'écarte du modèle non linéaire supposé ($\gamma = 0,0$ dans le tableau). Cependant, ces biais sont compensés par une faible variabilité, de sorte que si l'on considère la moyenne des RCEQM relative, HT-DFMTr donne encore d'aussi bons, voire de meilleurs, résultats que HT-EBLUPLin et continue d'être supérieur à HJ-DFMLin. Les biais produits par HJ-DFMLin et HT-EBLUPLin sont du même ordre, tandis que la moyenne des RCEQM relative de HT-EBLUPLin est plus grande que celle de HJ-DFMLin. Les taux de couverture moyen de HT-EBLUPLin sont marginalement meilleurs que ceux de HJ-DFMLin et de HT-DFMTr, mais la largeur moyenne des intervalles de confiance qui sous-tendent ces taux la plus petite a tendance à être observée pour HT-DFMTr, suivi par HT-EBLUPLin, puis par HJ-DFMLin. Globalement, les résultats de nos simulations fondées sur un modèle pour l'ensemble B indiquent que, même si l'estimation sur petits domaines au moyen d'un estimateur DFM avec des poids calés sur un modèle fondés sur un modèle peut donner lieu à un biais dû à l'erreur de spécification du modèle, la performance globale de cette approche semble être assez peu affectée par de légers écarts par rapport au modèle non linéaire supposé.

Dans le tableau 4 et à la figure 2, nous présentons les mesures de performance moyennes et par région produites par diverses méthodes d'estimation sur petits domaines pour les données de l'AAGIS, respectivement. Ces résultats montrent que le biais relatif moyen de HT-DFMTr est plus faible que celui de HT-EBLUPLin ainsi que celui de HJ-DFMLin, tandis que la moyenne des RCEQM relative de HT-DFMTr est légèrement supérieure aux valeurs correspondantes pour HJ-DFMLin et HT-EBLUPLin. L'examen de la figure 2 révèle que ce résultat est essentiellement dû à une région (21) de l'échantillon original de l'AAGIS qui contenait une valeur fortement aberrante (CDT >30 000 000 $A). Cette valeur aberrante a été incluse dans la population utilisée pour la simulation (deux fois) puis sélectionnée (dans un cas, deux fois) dans 37 des 1 000 échantillons de simulation, ce qui a donné lieu à des estimations entièrement irréalistes produites par HT-DFMTr et par HJ-DFMLin pour la région 21. La colonne de droite du tableau 4 donne par conséquent les mesures de performance moyennes de diverses méthodes quand cette région est exclue. Nous voyons ainsi que HT-DFMTr et HJ-DFMLin sont essentiellement à égalité, dominant tous deux HT-EBLUPLin. La raison pour laquelle HT-DFMTr ne fournit pas de gains significatifs par rapport à HJ-DFMLin dans ce cas tient au fait que les modèles linéaires mixtes sur l'échelle originale et sur l'échelle logarithmique sont tous deux relativement mal ajustés aux données de l'AAGIS.

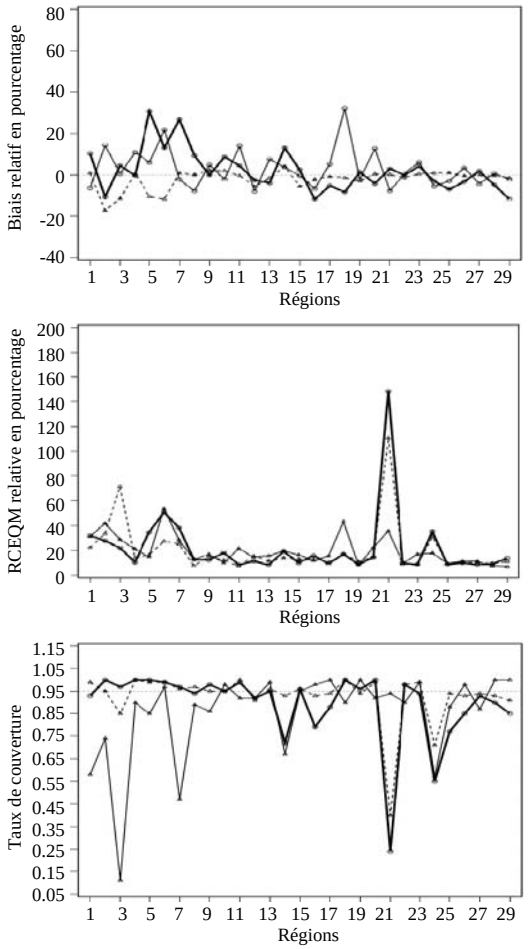

Figure 2 Résultats de simulation par région pour HT-DFMTr (trait épais, 0), HT-EBLUPLin (trait fin, Δ) et HJ-DFMLin (trait pointillé, Δ) dans les simulations fondées sur le plan de sondage basé sur les données de l'AAGIS. Les graphiques représentent (de haut en bas) BR (%), RCEQM relative (%) et TC. Les régions sont classées par ordre croissant de taille de population

7. Conclusion et travaux à venir

Les résultats des simulations exposées à la section précédente montrent que le fait de combiner des poids calés sur un modèle fondés sur un modèle avec l'estimation directe peut accroître considérablement l'efficacité de l'estimation sur petits domaines si la relation entre les donnés de population est clairement non linéaire. Comme il faut s'y attendre, ces gains sont moins importants quand le modèle non linéaire supposé est mal spécifié. Bien que nous ne donnions pas les détails, nos conclusions n'ont essentiellement pas changé quand nous avons exécuté des simulations similaires en utilisant des effets aléatoires suivant une loi gamma.

Notre mise en garde la plus importante en ce qui concerne l'utilisation des poids calés sur un modèle fondés sur un modèle (17) pour l'estimation sur petits domaines tient à leur spécificité. Ces poids ne semblent pas posséder les mêmes caractéristiques « polyvalentes » que les poids EBLUP classiques pour le calcul du total fondés sur des modèles linéaires mixtes. D'autres travaux seront donc nécessaires afin de déterminer comment construire, pour l'estimation sur petits domaines, des poids calés sur un modèle qui sont plus « polyvalents ». Nous nous attendons à ce que de tels poids ne soient pas aussi efficaces que les poids propres aux variables (17), mais nous espérons que cela sera plus que compensé par leur plus grande utilité. Un autre problème très important en pratique est que les variables d'enquête dont la distribution est positivement asymétrique peuvent également prendre des valeurs nulles (voire même négatives). Par exemple, les variables économiques, telles que le passif et les dépenses en immobilisations, prennent souvent des valeurs nulles, tandis que les variables définies comme étant la différence entre deux quantités non négatives (par exemple les bénéfices, qui correspondent à la différence entre les revenus et les dépenses) peuvent être négatives. Karlberg (2000b) utilise un mélange de modèles pour caractériser les données qui comprennent un mélange de valeurs nulles et de valeurs strictement positives. Ce genre de modèle peut être utilisé dans la pondération calée sur un modèle fondée sur un modèle.

Enfin, nous notons que l'utilisation d'une approche DFM fondée sur une transformation dans laquelle les hypothèses habituelles concernant le modèle linéaire ne sont qu'approximativement valides (la situation considérée dans le présent article) n'est pas la seule qui a été proposée pour résoudre ce problème. Deux autres approches décrites dans la littérature sont celles du pseudo-EBLUP (Rao 2003, section 7.2.7) et de l'estimateur de type bayésien empirique (BE) assisté par un modèle de Jiang et Lahiri (2006). Rappelons que, selon (8), le prédicteur EBLUP est défini en remplaçant la moyenne de domaine i m_{iy} inconnue par une estimation de sa valeur prévue donnée par les valeurs d'échantillon observées de Y dans le domaine i et les valeurs de $\mathbf{X}$ dans le domaine i. Soit π_{ij} la probabilité d'inclusion dans l'échantillon de l'unité de population j dans le petit domaine i. Le pseudo-EBLUP est alors défini en remplaçant m_{iy} par une estimation de sa valeur prévue donnée par la valeur de son estimation convergente sous le plan

$$\hat{m}_{iy}^{\pi} = \left(\sum_{j\in s_i} \pi_{ij}^{-1}\right)^{-1} \sum_{j\in s_i} \pi_{ij}^{-1} y_{ij} = \sum_{j\in s_i} \tilde{w}_{ij} y_{ij} \quad (21)$$

et les valeurs de $\mathbf{X}$ dans le domaine i. Autrement dit, sous (3), le pseudo-EBLUP de m_{iy} est

$$\begin{aligned}\hat{m}_{iy}^{\text{psuedoEBLUP}} &= \hat{E}\{m_{iy} \mid \hat{m}_{iy}^{\pi}, \mathbf{x}_{is}, \mathbf{x}_{ir}\} \\ &= \overline{\mathbf{x}}_i' \hat{\beta}_{\tilde{w}} + (\overline{\mathbf{g}}_i' \hat{\Sigma}_{u\tilde{w}} \overline{\mathbf{g}}_{i\tilde{w}}) \\ &\quad \left(\overline{\mathbf{g}}_{i\tilde{w}}' \hat{\Sigma}_{u\tilde{w}} \overline{\mathbf{g}}_{i\tilde{w}} + \hat{\sigma}_{e\tilde{w}}^2 \sum_{j\in s_i} \tilde{w}_{ij}^2\right)^{-1} (\hat{m}_{iy}^{\pi} - \overline{\mathbf{x}}_{i\tilde{w}}' \hat{\beta}_{\tilde{w}}) \quad (22)\end{aligned}$$

où $\hat{\beta}_{\tilde{w}}$, $\hat{\Sigma}_{u\tilde{w}}$ et $\hat{\sigma}_{e\tilde{w}}^2$ sont les estimations du pseudo-maximum de vraisemblance basées sur les poids $\tilde{w}_{ij}$, et $\overline{\mathbf{g}}_{i\tilde{w}}$ et $\overline{\mathbf{x}}_{i\tilde{w}}'$ sont les estimations convergentes sous le plan de $\overline{\mathbf{g}}_i$ et $\overline{\mathbf{x}}_i$, qui sont définies exactement de la même manière que $\hat{m}_{iy}^{\pi}$ susmentionnée. Sous le même modèle, l'approche de type BE assistée par modèle de Jiang et Lahiri (2006) mène à un estimateur qui est également défini par conditionnement sur la valeur de $\hat{m}_{iy}^{\pi}$,

$$\begin{aligned}\hat{m}_{iy}^{JL} &= \sum_{j\in s_i} \tilde{w}_{ij} \hat{E}\{\hat{E}(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i) \mid \hat{m}_{iy}^{\pi}, \mathbf{x}_i\} \\ &= \overline{\mathbf{x}}_{i\tilde{w}}' \hat{\beta} + \{\tilde{\mathbf{w}}_{is}' (\mathbf{g}_{is} \hat{\Sigma}_u \mathbf{g}_{is}' + \hat{\sigma}_e^2 \mathbf{I}_{is}) \tilde{\mathbf{w}}_{is}\}^{-1} \\ &\quad \{\tilde{\mathbf{w}}_{is}' \mathbf{g}_{is} \hat{\Sigma}_u \mathbf{g}' \tilde{\mathbf{w}}_{is}\} (\hat{m}_{iy}^{\pi} - \overline{\mathbf{x}}_{i\tilde{w}}' \hat{\beta}) \quad (23)\end{aligned}$$

où $\tilde{\mathbf{w}}_{is}$ est le vecteur des poids d'échantillon normalisés $\tilde{w}_{ij}$ dans le domaine i. Notons que, dans (23), nous utilisons des estimations optimales (c'est-à-dire MV ou MVR) pour les paramètres du modèle.

Les estimateurs (22) et (23) sont essentiellement motivés par l'idée d'estimer la moyenne du domaine i par son espérance conditionnelle sous (3), sachant la valeur de l'estimateur convergent sous le plan habituel (21) pour cette quantité. Il s'agit donc d'estimateurs indirects tels que l'estimateur HT-EBLUPLin. Sous (3), ni l'un ni l'autre n'est aussi efficace que l'estimateur HT-EBLUPLin, tandis que si (9) plutôt que (3) est vérifié, les deux estimateurs tirent parti de la convergence sous le plan de $\hat{m}_{iy}^{\pi}$ pour leur robustesse. Puisque s'appuyer sur une propriété en grand échantillon d'une statistique sur petit échantillon semble assez optimiste, nous préférons nous attaquer directement au problème de spécification du modèle, en remplaçant (3) par (9)

et en utilisant l'approche DFM fondée sur une transformation décrite à la section 5. Les valeurs du biais relatif moyen et de la moyenne des RCEQM relatives pour le pseudo-EBLUP (22) et l'estimateur de Jiang et Lahiri (23) sont présentées au tableau 4. Il est intéressant de souligner que ni l'un ni l'autre estimateur ne semble donner de meilleurs résultats que l'EBLUP classique dans ces simulations fondées sur le plan de sondage, et que tous trois sont surpassés considérablement en ce qui concerne la moyenne des RCEQM relatives par les deux estimateurs de type DFM étudiés dans le cadre de la présente étude. Manifestement, les résultats d'une étude par simulation unique (mais raisonnablement réaliste) ne peuvent être vus que comme des indications. Cependant, ils donnent certaines preuves que les propriétés asymptotiques sous le plan n'offrent aucune garantie quant à la performance de l'estimation sur petits domaines.

Le prédicteur indirect (20) de la moyenne de petit domaine est obtenu en appliquant des notions de prédiction bien connues. Sous les modèles avec transformation logarithmique, diverses approches existent pour obtenir le meilleur prédicteur indirect de la moyenne de petit domaine. Par exemple, Slud et Maiti (2006) ont décrit un prédicteur indirect de la moyenne de petit domaine sous une version au niveau du domaine du modèle logarithmiquement transformé (9). Berg (2009, communication privée) suit l'approche de Slud-Maiti pour obtenir un prédicteur de la moyenne de petit domaine sous une spécification avec ordonnée aléatoire du modèle avec transformation logarithmique (9) au niveau de l'unité. Cependant, comme celui de Slud-Maiti, le prédicteur de Berg ne tient pas compte de la correction du biais qui est nécessaire après la rétrotransformation pour revenir à l'échelle originale. Les propriétés empiriques de ce prédicteur n'ont pas encore été examinées.

Remerciements

Le premier auteur est reconnaissant à la Commonwealth Scholarship Commission du Royaume-Uni du soutien financier que lui a procuré la bourse de doctorat qui lui a été accordée. Les auteurs remercient également le rédacteur en chef, le rédacteur associé et deux examinateurs de leurs commentaires constructifs qui leur ont permis de réviser la version originale de l'article et de l'améliorer considérablement.

Bibliographie

Bates, D.M., et Pinheiro, J.-C. (1998). Computational Methods for Multilevel Models. http://franz.stat.wisc.edu/pub/NLME/.

Carroll, R., et Ruppert, D. (1988). *Transformation and Weighting in Regression*. New York : Chapman and Hall.

Chambers, R., et Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255-268.

Chandra, H., et Chambers, R.L. (2005). Comparing EBLUP and C-EBLUP for small area estimation. *Statistics in Transition*, 7, 637-648.

Chandra, H., et Chambers, R. (2009). Multipurpose weighting for small area estimation. *Journal of Official Statistic*, 25(3), 379-395.

Chandra, H., Salvati, N. et Chambers, R. (2007) Small area estimation for spatially correlated populations. A comparison of direct and indirect model-based methods. *Statistics in Transition*, 8, 887-906.

Chen, G., et Chen, J. (1996). Une méthode de transformation applicable à l'échantillonnage de population finies calés par une méthode de vraisemblance empirique. *Techniques d'enquête*, 22, 139-147.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-338.

Hidiroglou, M.A., et Smith, P.A. (2005). Developing small area estimates for business surveys at the ONS. *Statistics in Transition*, 7, 527-539.

Jiang, J., et Lahiri, P. (2006). Estimation of finite population domain means: A model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.

Karlberg, F. (2000a). Population total prediction under a lognormal superpopulation model. *Metron*, LVIII, 53-80.

Karlberg, F. (2000b). Survey estimation for highly skewed populations in the presence of zeroes. *Journal of Official Statistics*, 16, 229-241.

Longford, N.T. (2007). De l'erreur-type des estimateurs pour petits domaines fondés sur un modèle. *Techniques d'enquête*, 33, 81-92.

McCulloch, C.E., et Searle, S.R. (2001). *Generalized, Linear and Mixed Models*. New York : John Wiley & Sons, Inc.

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*. New York : John Wiley & Sons, Inc.

Royall, R.M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.

Royall, R.M., et Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.

Slud, E. V., et Maiti, T. (2006). Mean squared error estimation in transformed Fay–Herriot models. *Journal of the Royal Statistical Society*, Séries B, 68(2), 239-257.

Tzavidis, N., Salvati, N., Pratesi, M. et Chambers, R. (2008). M-quantile models with application to poverty mapping. *Statistical Methods And Applications*, 17, 393-411.

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

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.

Techniques d'enquête, juin 2011
Vol. 37, N° 1, pp. 59-72
Statistique Canada, N° 12-001-X au catalogue

Élaboration de plans stratifiés en R à l'aide du programme *stratification*

Sophie Baillargeon et Louis-Paul Rivest [1]

Résumé

Ce document présente un programme R pour la stratification d'une population d'enquête à l'aide d'une variable unidimensionnelle X et pour le calcul de tailles d'échantillon dans les strates. Nous y employons des méthodes non itératives pour délimiter les strates, comme la méthode de la fonction cumulative de la racine carrée des fréquences et la méthode géométrique. Nous pouvons élaborer des plans optimaux où les bornes de strates minimisent soit le CV de l'estimateur simple par dilatation pour une taille fixe d'échantillon n, soit la valeur n pour un CV fixe. Nous disposons de deux algorithmes itératifs pour le calcul des bornes optimales. Le plan peut comporter des strates à tirage obligatoire qui sont définies par l'utilisateur et dont toutes les unités sont échantillonnées. Il est également possible d'inclure dans le plan stratifié des strates à tirage complet et à tirage nul qui permettent souvent de réduire les tailles d'échantillon. Les calculs de taille d'échantillon sont fondés sur les moments anticipés de la variable d'enquête Y étant donné la variable de stratification X. Le programme traite les distributions conditionnelles de Y étant donné X qui sont soit un modèle linéaire hétéroscédastique soit un modèle loglinéaire. Nous pouvons tenir compte de la non-réponse par strate dans l'élaboration du plan d'échantillonnage et dans les calculs de taille d'échantillon.

Mots clés : Modèles linéaires ; modèles loglinéaires ; stratification optimale ; échantillonnage d'enquête ; strate à tirage complet ; strate à tirage nul.

1. Introduction

L'établissement de strates et l'élaboration d'un plan stratifié sont une question importante en méthodologie d'enquête depuis les travaux de pionnier de Dalenius il y a plus de soixante ans. Il sera question ici d'une stratification à l'aide d'une variable unidimensionnelle positive X connue pour toutes les unités de la population. On pose que X est en relation avec la variable d'enquête Y. La strate h contient toutes les unités ayant une valeur X dans l'intervalle $[b_{h-1}, b_h)$ pour $h = 1, \ldots, L$, de sorte que $b_0 = \min X$ et $b_L = \max X + 1$, où $\min X$ et $\max X$ sont respectivement les valeurs minimale et maximale de la variable de stratification.

La délimitation optimale de strates a une longue histoire ; on peut consulter à ce sujet le chapitre 5A dans Cochran (1977). La méthode de la fonction cumulative de la racine carrée des fréquences ($\text{cum}\sqrt{f}$) de Dalenius et Hodges (1959) apporte une solution approximative à ce problème. Les cas où X a une distribution asymétrique sont fréquents dans les enquêtes auprès des entreprises et ont reçu une attention particulière. Gunning et Horgan (2004) ont proposé une méthode géométrique de stratification et Hidiroglou (1986) a fait valoir que les grandes unités devraient être mises dans une strate à tirage complet. Plutôt que de s'en remettre à une méthode approximative pour la construction des strates, Lavallée et Hidiroglou (1988) ont proposé un algorithme itératif qui donne des bornes optimales pour une variable X déterminée. Parfois cet algorithme ne converge pas (Detlefsen et Veum 1991). Slanta et Krenzke (1996) ont démontré que, dans certains cas, les bornes optimales ne sont pas définies d'une manière unique. On a proposé d'autres méthodes comme l'algorithme de recherche aléatoire de Kozak (2004) pour parer à certaines de ces difficultés. L'hypothèse d'une correspondance exacte entre la variable d'enquête Y et la variable de stratification X est irréaliste lorsqu'on calcule des tailles d'échantillon et plusieurs auteurs, dont Dayal (1985) et Sigman et Monsour (1995), ont voulu répartir l'échantillon entre les strates en se fondant sur les moments anticipés de Y si on sait que X se trouve dans $[b_{h-1}, b_h)$. Sweet et Sigman (1995) et Rivest (1999, 2002) ont suggéré de faire entrer ces moments anticipés dans l'algorithme de stratification de Lavallée et Hidiroglou (1988). Récemment, Baillargeon et Rivest (2009) ont montré que, en mettant les petites unités dans une strate à tirage nul, c'est-à-dire hors échantillonnage, on pouvait réduire la taille d'échantillon nécessaire à l'obtention d'un degré de précision préétabli.

Dans cet article, nous présentons le programme R *stratification* qui met en oeuvre la plupart des méthodes que nous venons de mentionner. Il s'agit d'un environnement convivial pour l'élaboration de plans stratifiés et l'évaluation de leur rendement pour des populations réelles. Nous présentons ce programme en revisitant des exemples déjà traités dans la litérature qui illustrent ses caractéristiques importantes. Les quatre fonctions de *stratification* dont le nom débute par `strata` construisent des plans d'échantillonnage stratifiés. Ce sont les fonctions `strata.cumrootf`, `strata.geo`, `strata.LH` et `strata.bh`. Les deux premières correspondent aux méthodes simples du $\text{cum}\sqrt{f}$ et des

1. Sophie Baillargeon, Département de mathématiques et de statistique, 1045, avenue de la Médecine, Université Laval, Québec, Canada G1V 0A6. Courriel : sophie.baillargeon@mat.ulaval.ca ; Louis-Paul Rivest, Département de mathématiques et de statistique, 1045, avenue de la Médecine, Université Laval, Québec, Canada G1V 0A6. Courriel : louis-paul.rivest@mat.ulaval.ca.

bornes géométriques. Quant à `strata.LH`, elle construit des plans stratifiés optimaux à l'aide d'algorithmes itératifs. La dernière fonction traite des bornes de strates fournies par l'utilisateur. Avec ces quatre fonctions, on construit des strates, établit des tailles d'échantillon de strates et calcule le degré de précision de l'estimateur simple par dilatation $\bar{y}_s$ de $\bar{Y}$, la moyenne de la variable d'intérêt Y dans la population qui est liée à la variable de stratification X.

Dans les quatre fonctions `strata`, on emploie la règle de Hidiroglou et Srinath (1993) pour attribuer aux strates les n unités de l'échantillon. Les tailles d'échantillon des strates sont proportionnelles à $N_h^{2q_1} \bar{Y}_h^{2q_2} S_{yh}^{2q_3}$, où N_h est la taille de la strate h, et où $\bar{Y}_h$ et S_{yh}^2 sont respectivement la moyenne et la variance anticipées de Y dans cette strate. Dans les fonctions `strata`, on se trouve à spécifier une règle d'allocation par l'argument `alloc` qui contient les exposants (q_1, q_2, q_3) ; l'attribution de Neyman correspond à `alloc=c(1/2,0,1/2)`. Dans une fonction `strata`, on prend comme entrées le vecteur de population de la variable de stratification X, le nombre de strates `Ls` et une taille totale d'échantillon n ou un CV cible pour l'estimateur simple par dilatation $\bar{y}_s$. La sortie est un objet `R` de classe strata qui définit un plan stratifié. On y trouve un ensemble de strates déterminées par leurs bornes supérieures $\{b_h\}$ et les tailles de population et d'échantillon de strates N_h et n_h. Dans le programme *stratification*, il existe une cinquième fonction appelée `var.strata` ayant pour entrées un objet `R` de classe strata et un vecteur donnant les valeurs d'une variable d'intérêt Y dans la population. Sa sortie est la variance de $\bar{y}_s$ pour la variable Y et le plan stratifié donné en entrée.

Le texte comporte des instructions en `R` à taper dans une fenêtre de commande. Ces lignes commencent par `>`. Le texte présente aussi des sorties telles qu'elles s'impriment dans une fenêtre de commande. Une police spéciale permet de reconnaître facilement les commandes et les sorties R, qu'elles se retrouvent dans un paragraphe disctint ou encore dans le texte. L'annexe offre un tableau récapitulatif énumérant tous les arguments possibles des cinq fonctions de *stratification*. Lorsqu'on utilise ce programme, l'instruction `help(stratification)` ouvre un fichier d'assistance cliquable qui renseigne en détail sur le programme et donne des exemples pouvant être collés dans une fenêtre de commande.

2. Méthodes de stratification de base

Dans cette section, nous examinerons deux méthodes élémentaires de stratification, soit la méthode cum $\sqrt{f}$ de Dalenius et Hodges (1959) et la méthode géométrique de Gunning et Horgan (2004). Ces méthodes sont à calcul exact et ne font donc pas intervenir d'algorithme itératif. Dans toute cette section, $Y = X$. L'emploi d'une même variable pour la stratification et l'évaluation de la précision des estimations d'enquête pourrait avoir pour conséquence une sous-estimation des variances. Nous traiterons à la section 4 de la question du calcul des variances en cas de non-correspondance $Y \neq X$.

2.1 Méthode cum $\sqrt{f}$

Cet algorithme de stratification présenté au chapitre 5A de Cochran (1977) s'exécute par la fonction `strata.cumrootf`. Ses arguments sont `x`, vecteur de la variable de stratification dans la population, `nclass`, nombre de classes de taille égale pour la variable `x`, un `CV` cible pour $\bar{y}_s$ ou une taille d'échantillon préétablie `n` avec `Ls`, nombre de strates, et `alloc`, règle d'attribution. Cet algorithme réunit les `nclass` classes en `Ls` strates de sorte que la somme des racines carrées des fréquences dans les classes soit approximativement égale d'une strate à l'autre.

En guise d'illustration, prenons la proportion de prêts industriels de $N = 13\,435$ banques dans Cochran (1961). Nous stratifions cette population et évaluons la taille d'échantillon à prévoir pour que $\bar{y}_s$ ait un CV de 5 % avec l'allocation de Neyman. Le code `R` qui suit crée le vecteur de la variable de stratification `loans` donnée au tableau 2 de McEvoy (1956). Nous appliquons ensuite la fonction `strata.cumrootf` à cette variable. Comme dans le tableau 2 de Cochran (1961), trois strates seront créées à partir de 20 classes, donc nclass est fixé à 20 et `Ls=3`. La sortie est enregistrée dans `cum`, un objet `R` de classe strata. Si on tape `cum` ou `print(cum)` dans la fenêtre de commande `R`, on imprime le plan d'échantillonnage en détail. Les arguments d'entrée, par défaut ou spécifiés par l'utilisateur, s'affichent en premier. Les informations relatives aux strates s'affichent ensuite : bornes, tailles de strate N_h et tailles d'échantillon n_h. La troisième partie de la sortie renseigne sur les propriétés échantillonnales de $\bar{y}_s$.

```
> values <- c(seq(0.5, 9.5, 1), seq(12.5, 97.5, 5))
> nrep <- c(1985, 261, 339, 405, 474, 478, 506, 569, 464, 499,
  2157, 1581, 1142, 746, 512, 376, 265, 207, 126, 107, 82, 50,
  39, 25, 16, 19, 2, 3)
> loans <- rep(values, nrep)
> cum <- strata.cumrootf(x = loans, nclass = 20, CV = 0.05,
  Ls = 3, alloc = c(0.5, 0, 0.5))
> cum
Given arguments:
x = loans
nclass = 20, CV = 0.05, Ls = 3
allocation : q1 = 0.5, q2 = 0, q3 = 0.5
model = none
Strata information:
          rh |   bh  anticip.Mean anticip.var    Nh   nh    fh
Stratum 1   1 | 10.2          4.12       10.46  5980   14 0.00
Stratum 2   1 | 29.6         17.92       27.74  5626   20 0.00
Stratum 3   1 | 98.5         44.47      165.83  1829   16 0.01
Total                                    13435   50 0.00

Total sample size: 50
Anticipated population mean: 15.39408
Anticipated CV: 0.0494897
```

Dans `Given arguments, model=none` signifie que les propriétés d'échantillonnage de $\bar{y}_s$, présentées à la fin de la sortie sont évaluées avec $Y = X$, en l'occurrence pour la variable `loans`. La moyenne est 15,39408 et le CV prévu de 0,0494897 est celui de l'estimateur $\bar{y}_s$ de la moyenne de la variable `loans` obtenue avec ce plan d'échantillonnage. Les limites des strates dans cette sortie sont (10,2, 29,6, 98,5) ; elles sont égales à celles qui figurent au bas de la page 349 dans Cochran (1961), une fois pris en compte les arrondis. Dans `Strata Information`, r_h renvoie aux taux de réponse des strates dont il sera question à la section 5.1. L'objet cum en `R` contient plusieurs éléments énumérés par la commande `names(cum)`.

```
> names(cum)
[1]  "Nh"            "nh"   "n"          "nh.nonint" "certain.info"
[6]  "opti.criteria" "bh"   "meanh"      "varh"      "mean"
[11] "stderr"        "CV"   "stratumID" "nclassh"   "takeall"
[16] "call"          "date" "args"
```

On peut imprimer un élément de l'objet `cum` de classe strata en tapant `cum$` suivi du nom de l'objet. Ainsi, la commande `cum$stratumID` imprime la strate de chaque unité de la population. La variable `cum$nclassh` est propre à la fonction `strata.cumrootf` et indique comment les `nclass=20` classes initiales ont été réunis en trois strates.

```
> cum$nclassh
[1] 2 4 14
```

Dans cette stratification, les strates 1, 2 et 3 contiennent respectivement 2, 4 et 14 des `nclass=20` classes initiales.

2.2 Méthode géométrique

La méthode géométrique de stratification a été introduite par Gunning et Horgan (2004). Elle délimite les strates par $b_h = \min X \times (\max X / \min X)^{h/L}$, pour $h = 1, ..., L-1$. Une fois les bornes b_h établies, les calculs de taille d'échantillon des strates sont les mêmes qu'avec la méthode `strata.cumrootf`.

Pour illustrer, nous avons stratifié les quatre populations présentées dans Gunning et Horgan (2004), à savoir Debtors, USbanks, UScities et UScolleges, en `Ls=5` strates. Les trois dernières populations ont été examinées dans Cochran (1961). Les quatre sont stockées dans *stratification*. On appelle la première par la commande `data(Debtors)`. Plutôt que de spécifier un CV cible, nous fixons la taille d'échantillon à $n = 100$ tel que fait par Gunning et Horgan (2004). Les commandes qui suivent créent l'objet `R pop1` qui contient le plan stratifié de la population Debtors.

```
> data(Debtors)
> pop1 <- strata.geo(x = Debtors, n = 100, Ls = 5,
  alloc = c(0.5, 0, 0.5))
```

Le tableau 1 récapitule les plans stratifiés obtenus en appliquant la méthode géométrique aux quatre populations en question. Il reproduit en partie le tableau 4 de Gunning et Horgan (2004). Notons cependant un certain nombre de différences légères tenant à des différences de stratégie d'arrondissement. On trouvera plus de détails sur les méthodes d'arrondissement dans *stratification* en consultant le fichier d'assistance.

Tableau 1
Plans stratifiés pour quatre populations avec n = 100

Population	CV		1	2	3	4	5
Debtors	0,0359	b_h	148,28	549,67	2 037,60	7 553,33	
		N_h	1 054	1 267	732	265	51
		n_h	3	14	27	33	23
UScities	0,0145	b_h	18,17	33,01	59,98	108,98	
		N_h	364	418	130	87	39
		n_h	18	28	17	20	17
UScolleges	0,0183	b_h	434,00	941,76	2 043,61	4 434,60	
		N_h	94	255	198	74	56
		n_h	3	15	27	20	35
USbanks	0,0107	b_h	118,59	200,92	340,39	576,68	
		N_h	114	116	64	39	24
		n_h	13	20	25	18	24

2.3 Strate à tirage complet

Au tableau 1, la cinquième strate pour la population USbanks est à tirage complet puisque $n_5 = N_5 = 24$. Avec l'allocation de Neymann, cette strate a une taille d'échantillon n_5 supérieure à la taille de population N_5. `strata.geo` désigne automatiquement cette strate comme étant à tirage complet et attribue les $n - N_5$ unités aux quatre premières strates selon l'allocation de Neyman. Cette correction importe pour l'obtention d'une taille d'échantillon de $n = 100$, comme il est spécifié dans les arguments de la fonction `strata.geo`.

Pour illustrer ce point, nous employons la fonction `strata.bh` et procédons à l'attribution sans correction pour la strate à tirage complet. Nous répartissons l'échantillon et calculons le degré de précision de $\bar{y}_s$ pour un ensemble préétabli de bornes de strates. En réglant `takeall.adjust=FALSE`, nous nous servons de l'attribution de Neyman dans les cinq strates et, comme $n_5 > N_5$, le résultat affiché est $n_5 = N_5$. Le code qui suit en R prend les bornes de strates géométriques $\{b_h\}$ dans l'objet de classe strata `adjust`. La fonction `strata.bh` est ensuite appliquée pour obtenir le plan d'échantillonnage sans correction pour la cinquième strate à tirage complet dans l'objet de classe strata `noadjust`.

```
> data(USbanks)
> adjust <- strata.geo(x = USbanks, n = 100, Ls = 5,
  alloc = c(0.5, 0, 0.5))
> noadjust <- strata.bh(x = USbanks, bh = adjust$bh,
  n = 100, Ls = 5, alloc = c(0.5, 0, 0.5), takeall = 0,
  takeall.adjust = FALSE)
```

Les deux plans sont présentés au tableau 2. Sans la strate à tirage complet, la taille d'échantillon est $n = 99$. Dans ce

cas, la taille d'échantillon non arrondie pour la strate 5 est `noadjust$nh.noint[5]`= 25,40 pour $N_5 = 24$ unités. À noter que, lorsque n est grand ou que le CV cible est bas, il peut y avoir plusieurs strates à tirage complet.

Tableau 2
Plans stratifiés obtenus avec et sans correction automatique pour une strate à tirage complet

	n		1	2	3	4	5
		b_h	118,59	200,92	340,39	576,68	
		N_h	114	116	64	39	24
correction	100	n_h	13	20	25	18	24
non-correction	99	n_h	13	20	24	18	24

2.4 Ajout d'une strate à tirage complet

Considérons maintenant la base de données de $N = 284$ municipalités suédoises en annexe de Särndal, Swensson et Wretman (1992). Dans les instructions qui suivent, nous employons la méthode géométrique pour répartir cette population en `Ls=5` strates par la variable `REV84`, qui donne les valeurs foncières en 1984. Nous employons l'attribution de puissance avec l'exposant 0,7 et `alloc= c(0.35,0.35,0)`. L'objet `R` de classe strata `geo` contient le plan stratifié. La commande `plot(geo)` produit la figure 1. On y trouve un histogramme de la variable de stratification avec les bornes des strates et un tableau récapitulatif du plan stratifié.

```
> data(Sweden)
> geo <- strata.geo(x = Sweden$REV84, CV = 0.05, Ls = 5,
  alloc = c(0.35, 0.35, 0))
```

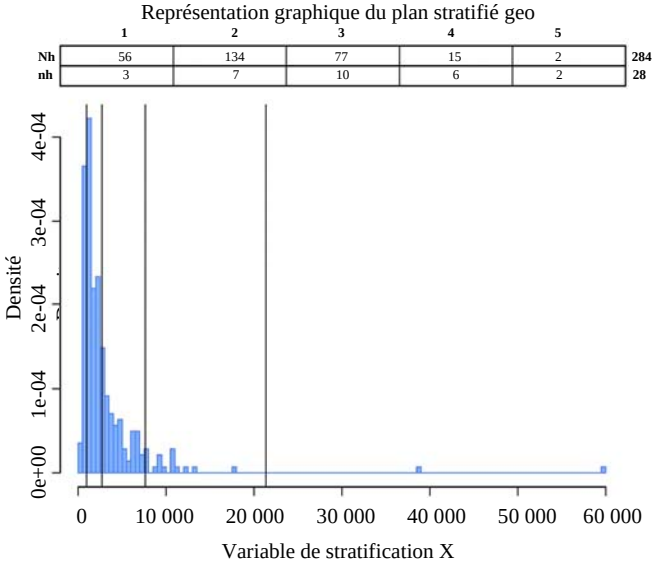

Figure 1 Graphique de l'objet en `R` `geo`

La figure 1 indique que, avec la méthode géométrique de stratification, on se trouve à mettre deux des trois valeurs extrêmes de `REV84` dans une strate à tirage complet. Le code R qui suit crée `cum`, un plan stratifié pour cette population par la méthode cum $\sqrt{f}$. Cette méthode cum $\sqrt{f}$ est d'une application difficile, puisque les classes sont de longueur `{max(REV84)–min(REV84)}/50`= 1191. On constate à la figure 1 que la plupart des classes présentent une fréquence nulle. En fait, la strate 5 comprend 43 des 50 classes. Dans ce plan, il n'y a pas de strate à tirage complet. Pour calculer les tailles d'échantillon avec une telle strate, on peut employer la fonction `strata.bh` avec les limites cum $\sqrt{f}$ stockées dans `cum$bh` en prenant la commande `takeall=1`. Nous créons ainsi le troisième plan d'échantillonnage au tableau 3. Nous créons le quatrième plan du tableau 3 `cum3` en rendant la taille d'échantillon de la strate 5 du plan cum $\sqrt{f}$ égale à sa taille de population par la commande `cum3$nh[5]<-cum3$Nh[5]`. Nous calculons enfin par la fonction `var.strata` la variance de l'estimation $\bar{y}_s$ pour la variable `REV84` à l'aide de ce quatrième plan.

```
> cum <- strata.cumrootf(x = Sweden$REV84, nclass = 50,
  CV = 0.05, Ls = 5, alloc = c(0.35, 0.35, 0))
> cum2 <- strata.bh(x = Sweden$REV84, bh = cum$bh, CV = 0.05,
  Ls = 5, takeall = 1, alloc = c(0.35, 0.35, 0))
> cum3 <- cum
> cum3$nh[5] <- cum3$Nh[5]
> cum3.var <- var.strata(cum3, y = Sweden$REV84)
```

Tableau 3
Quatre plans stratifiés de la population de municipalités suédoises

Méthode		1	2	3	4	5	n	CV
géométrique	N_h	56	134	77	15	2		
	n_h	3	7	10	6	2	28	4,83
cum $\sqrt{f}$	N_h	120	70	52	27	15		
	n_h	7	7	9	8	10	41	4,87
	$n_h^{modif\,1}$	2	2	3	2	15	24	4,44
	$n_h^{modif\,2}$	7	7	9	8	15	46	2,29

Le tableau 3 montre bien que le taux d'échantillonnage de la cinquième strate est ce qui détermine la valeur de n. Le plan cum $\sqrt{f}$ paraît moins efficace que celui de la méthode géométrique, puisque son taux d'échantillonnage est 10/15 = 67 % pour la strate 5. Avec une strate à tirage complet, on obtient une valeur de n comparable à celle du plan géométrique. La population `REV84` présente trois valeurs extrêmes relevées au tableau 1. La méthode géométrique et la méthode de la fonction cum $\sqrt{f}$ dépendent très largement de la valeur maximale de X, aussi serait-il avisé d'écarter ces valeurs extrêmes avant d'appliquer ces méthodes. C'est ce dont nous parlerons à la prochaine section.

La technique *ad hoc* présentée dans cette section pour une modification arbitraire de tailles d'échantillon est applicable dans plusieurs situations. Par exemple, si des strates ont une taille d'échantillon vallant 1, on peut majorer celle-ci à 2 pour que l'estimateur de variance soit exempt de biais.

2.5 Strate à tirage obligatoire

Il pourrait être bon dans un plan stratifié d'imposer la contrainte de l'échantillonnage d'un certain nombre d'unités

avant la construction des strates. L'argument `certain` disponible pour les quatre fonctions `strata` rend la chose possible. À titre d'exemple, reprenons la comparaison des plans présentés au tableau 3 pour la méthode de la fonction cum$\sqrt{f}$ et la méthode géométrique. On met les trois grandes municipalités de la figure 1 dans une strate à tirage obligatoire et les $N = 281$ autres municipalités dans `Ls=4` strates par ces deux méthodes de stratification. Nous donnons plus loin le code R pour l'élaboration de ces deux plans. Avec la commande `x=sort(Sweden$REV84)`, on ordonne les municipalités selon *REV*84. Ainsi, les trois grandes municipalités en question sont les éléments 282, 283 et 284 du vecteur trié. Les deux objets R de classe strata, à savoir `geo_cer` et `cum_cer`, contiennent chacun un élément `certain.info` qui caractérise la strate à tirage obligatoire.

```
> geo_cer <- strata.geo(x = sort(Sweden$REV84), CV = 0.05,
  Ls = 4, alloc = c(0.35, 0.35, 0), certain = 282:284)
> cum_cer <- strata.cumrootf(x = sort(Sweden$REV84),
  nclass = 50, CV = 0.05, Ls = 4, alloc = c(0.35, 0.35, 0),
  certain = 282:284)
> cum_cer$certain.info

  Nc     meanc
3.00 38923.67
```

Au tableau 4, le plan cum$\sqrt{f}$ est plus efficace que celui de la méthode géométrique. On a intérêt à mettre les trois grandes municipalités dans une strate à tirage obligatoire, car les tailles d'échantillon au tableau 4 sont inférieures à celles du tableau 3. Avec l'argument `certain`, on peut forcer l'inclusion de tout ensemble d'unités dans l'échantillon. On peut s'en servir pour y inclure des unités qui sont extrêmes pour une variable secondaire, différente de la variable de stratification, ou qui se sont déjà révélées hautement instables.

Tableau 4
Deux plans stratifiés visant les municipalités suédoises et élaborés avec une strate à tirage obligatoire

Méthode		1	2	3	4	5	*n*	CV
géométrique	N_h	42	116	88	35	3		
	n_h	2	5	7	7	3	24	4,71
cum $\sqrt{f}$	N_h	127	79	46	29	3		
	n_h	3	4	4	5	3	19	4,72

3. Méthode d'optimisation

Les méthodes de stratification présentées à la section 2 ne donnent pas toujours un plan stratifié optimal qui minimise la taille d'échantillon n nécessaire à l'obtention du CV cible (ou qui minimise le CV pour un n fixe). Dans cette section, nous présenterons la fonction `strata.LH` pour la détermination de plans optimaux. `LH` correspond à Lavallée et Hidiroglou (1988), des pionniers de la construction de plans optimaux pour des populations réelles. Dans un plan stratifié avec strate à tirage complet, la variance de l'estimateur par dilatation est donnée par

$$\mathrm{Var}(\bar{y}_s) = \sum_{h=1}^{L-1}\left(\frac{N_h}{N}\right)^2\left(\frac{1}{(n-N_L)a_h} - \frac{1}{N_h}\right)S_{yh}^2,$$

où $\{a_h\}$ est la règle d'attribution pour le calcul des tailles d'échantillon. Le n qui assure un CV de c est donné par

$$n = N_L + \frac{\sum_{h=1}^{L-1} N_h^2 S_{yh}^2 \,/\, (a_h N^2)}{\bar{Y}^2 c^2 + \sum_{h=1}^{L-1} N_h S_{yh}^2 \,/\, N^2}. \qquad (1)$$

Dans cette expression, on peut écrire $n = n(b_1,\ldots,b_L)$ pour bien faire voir que la valeur de n dépend des bornes des strates. Avec la fonction `strata.LH`, on vise des limites optimales b_h qui minimisent $n(b_1, \ldots, b_{L-1})$. Nous disposons de deux algorithmes de minimisation, soit celui de Sethi (1963) tel que mis en oeuvre par Lavallée et Hidiroglou (1988) avec `algo="Sethi"` et l'algorithme de recherche aléatoire de Kozak (2004) avec `algo="Kozak"`. Le second est l'algorithme par défaut. Dans cette section, $Y = X$; nous ne distinguons donc pas la variable de stratification de la variable d'enquête.

3.1 Exemple tiré de Sethi (1963) avec distribution normale

Un problème classique consiste à déterminer les limites optimales de L strates dans une population infinie à partir d'une distribution connue. Ainsi, Sethi (1963) a calculé des bornes optimales pour les distributions normale et χ^2_{30}. Pour obtenir des solutions approximatives, on peut appliquer la fonction `strata.LH` à une population Monte Carlo simulée à partir de la distribution connue sans demander de strate à tirage complet. Dans (1), on a $N_h / N^2 \approx 0$ et les limites optimales sont les mêmes pour tout CV cible c.

Dans le code R qui suit, nous simulons des populations de taille 10^5 à partir des distributions χ^2_{30} et $N(10, 1)$. Observons que le programme *stratification* exige que la variable de stratification soit positive et, par conséquent, la fonction ne serait pas applicable à une distribution normale standard. En soustrayant 10 des limites $N(10, 1)$ nous obtenons celles de la $N(0, 1)$. Les calculs utilisent la fonction `strata.LH`, avec les arguments `algo="Sethi"` et `takeall=0`, une strate à tirage complet n'est donc pas demandée.

```
> z <- rnorm(100000, 10)
> zl5 <- strata.LH(x = z, CV = 0.001, Ls = 5,
  alloc = c(0.5, 0, 0.5), takeall = 0, algo = "Sethi")
> zl5$bh - 10
[1] -1.1247340 -0.3480829 0.3297044 1.0979017

> x30 <- rchisq(100000, 30)
> xl5 <- strata.LH(x = x30, CV = 0.01, Ls = 5,
  alloc = c(0.5, 0, 0.5), takeall = 0, algo = "Sethi")
> xl5$bh
[1] 22.82148 28.12303 33.38642 40.20165
```

Au tableau 5, la concordance est bonne entre les bornes présentées au tableau 8 de Sethi (1963) et celles obtenues par simulation. On pourrait employer la même technique pour calculer les bornes optimales des strates à partir d'une distribution arbitraire comme dans Khan, Nand et Ahmad (2008).

Tableau 5
Comparaison des bornes optimales de Sethi (1963) et des bornes approximatives obtenues avec *stratification*

		Résultats de *stratification*				Résultats de Sethi			
	L	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
	2	-0,007				0,00			
b_h	3	-0,531	0,567			-0,55	0,55		
$N(0,1)$	4	-0,883	-0,008	0,864		-0,88	0,00	0,88	
	5	-1,125	-0,348	0,330	1,098	-1,11	-0,34	0,34	1,11
	2	30,674				30,6			
b_h	3	26,535	35,141			26,0	35,0		
χ^2_{30}	4	24,340	30,733	38,179		24,0	30,6	38,0	
	5	22,821	28,123	33,386	40,202	22,0	28,0	33,0	40,0

3.2 Exemple tiré de Gunning et Horgan (2004)

Les plans stratifiés construits par Lavallée et Hidiroglou (1988) ont toujours une strate à tirage complet pour une variable d'enquête asymétrique. Pour montrer que la chose n'était pas toujours obligatoire, Gunning et Horgan (2004) ont construit leurs plans stratifiés optimaux avec une strate à tirage complet pour les quatre populations au tableau 1. Les résultats de leur tableau 7 (avec de légères corrections tenant à des erreurs d'arrondis) sont reproduits au tableau 6. Si on compare les tableaux 1 et 6, on constate que les plans optimaux avec strate à tirage complet présentent des valeurs n de plus de 100 pour trois populations sur quatre. Le plan optimal est supérieur à celui de la méthode géométrique seulement pour la population Debtors. Nous présentons ci-après le code R pour exécuter l'algorithme de Sethi sur la population Debtors.

```
> pop1LH <- strata.LH(x = Debtors, CV = 0.0359, Ls = 5,
  alloc = c(0.5, 0, 0.5), takeall = 1, algo = "Sethi")
```

Au tableau 6, on s'attendrait à ce que les plans optimaux obtenus à l'aide d'un algorithme itératif aient une taille d'échantillon inférieure à celle des plans géométriques, attente déçue pour trois populations. Ce pourrait être que l'algorithme de Sethi est incapable de trouver la véritable valeur minimale de *n*. Pour le vérifier, nous avons refait les calculs avec l'argument `algo="Kozak"`. Les tailles d'échantillon n figurent à la deuxième colonne du tableau 7. Avec l'algorithme de Kozak, on trouve une valeur n inférieure à celle de l'algorithme de Sethi pour trois des quatre populations. Cela montre bien la faiblesse qu'accuse l'algorithme de Sethi dans le traitement de populations réelles. À la deuxième colonne du tableau 7, les valeurs n sont de plus de 100 pour deux des quatre populations. Dans ce cas, le plan de la méthode géométrique pourrait être meilleur, parce qu'on n'a pas besoin d'une strate à tirage complet. Nous avons réexécuté l'algorithme de Kozak sans exiger une strate à tirage complet, c'est-à-dire avec `takeall=0`. Les résultats figurent à la troisième colonne du tableau 7. Pour les populations Debtors et UScolleges, le retrait de la strate à tirage complet vient réduire la taille d'échantillon n. Cependant, pour la population UScities, l'algorithme de Kozak demeure pire que la méthode géométrique. Il ne trouve pas la véritable valeur minimale de n avec les arguments par défaut qui règlent sa recherche aléatoire. Pour mieux comprendre les résultats du tableau 7, nous présenterons maintenant plus en détail le choix de limites initiales de strates avec `strata.LH` et les paramètres qui règlent la recherche aléatoire avec `algo="Kozak"`.

Tableau 6
Plans stratifiés optimaux avec strate à tirage complet par l'algorithme de Sethi appliqué aux quatre populations du tableau 1

Population	***n***	**CV**		**1**	**2**	**3**	**4**	**5**
Debtors	93	0,0359	b_h	349,33	1 190,16	3 482,98	10 322,50	
			N_h	1 856	991	350	146	26
			n_h	13	17	17	20	26
UScities	137	0,0145	b_h	14,72	21,62	35,59	80,47	
			N_h	189	270	336	164	79
			n_h	4	8	16	30	79
UScolleges	107	0,0183	b_h	512,32	869,76	1 577,23	3 668,85	
			N_h	133	180	185	110	69
			n_h	4	6	10	18	69
USbanks	104	0,0107	b_h	99,37	129,60	181,94	317,36	
			N_h	70	66	82	65	74
			n_h	4	4	7	15	74

Tableau 7
Taille d'échantillon n pour trois plans optimaux et quatre populations

Population	**`algo=Sethi`**	**`algo=Kozak`**	**`algo=Kozak`**
	`takeall=1`	`takeall=1`	`takeall=0`
Debtors	93	92	82
UScities	137	114	123
UScolleges	107	107	95
USbanks	104	88	88

3.3 Réglage des algorithmes

Les bornes initiales par défaut pour les deux algorithmes itératifs sont les valeurs initiales arithmétiques de Gunning et Horgan (2007) avec $b_h = \min X + (\max X - \min X) \times h/L$, pour $h = 1, \ldots, L-1$. Au tableau 7, ce choix est contestable et les bornes géométriques auraient été plus proches des bornes optimales. Dans la fonction `strata.LH`, l'argument `initbh=` permet de spécifier un vecteur de $L-1$ valeurs initiales des bornes. On peut modifier le nombre maximal d'itérations avec l'élément `maxiter` de l'argument `algo.control`.

L'algorithme de Kozak a d'abord été proposé dans Kozak (2004) ; voir aussi Kozak et Verma (2006). On y procède par

une recherche aléatoire où on choisit les $L-1$ limites de strates parmi les valeurs ordonnées de X, avec élimination des doublons. Dans une itération, on prend au hasard une valeur d dans l'ensemble {-maxstep,-maxstep+1,…,maxstep} et une des $L-1$ bornes, puis on déplace la limite choisie de d positions dans le vecteur des valeurs triées de X. Si (1) diminue avec la nouvelle borne, cette dernière est conservée, sinon elle est mise de coté et les bornes sont inchangées pour cette itération. L'algorithme s'arrête lorsque les bornes n'ont pas changé pour maxstill itérations consécutives. Les valeurs par défaut sont maxstep=3 et maxstill=100. Deux exécutions consécutives de l'algorithme de Kozak pourraient donner des plans différents, parce que cet algorithme est aléatoire. La fonction strata.LH exécute l'algorithme rep fois et l'information de chaque itération est contenue dans l'élément rep.detail des objets R de classe strata. La valeur par défaut est rep=3. Si les rep exécutions mènent à des plans différents, on peut modifier les paramètres de réglage de l'algorithme. On peut aussi employer rep="change", qui exécute l'algorithme 27 fois avec des valeurs initiales et des maxstep différentes. On trouvera en annexe un autre exemple illustrant le cas où l'algorithme de Kozak n'atteint pas un minimum global.

Avec N_u valeurs uniques de X, il y a approximativement $\binom{N_u-1}{L-1}$ ensembles possibles de limites de strates. Si ce nombre est inférieur à minsol, tous les ensembles possibles de bornes sont testés par opposition à un traitement aléatoire. La valeur par défaut est minsol=1000. Les éléments maxstep, maxstill, minsol et rep appartiennent à l'argument algo.control. Au tableau 7, nous avons été incapables d'améliorer le plan stratifié issu de la méthode géométrique pour la population UScities. Nous présentons ci-après la commande pour exécuter l'algorithme de Kozak 27 fois avec divers paramètres de réglage.

```
> data(UScities)
> pop2LHrep <- strata.LH(x = UScities, CV = 0.0145, Ls = 5,
  alloc = c(0.5, 0, 0.5), takeall = 0, algo = "Kozak",
  algo.control = list(rep = "change"))
```

Cette commande s'exécute en quelques secondes et donne un plan stratifié avec $n = 100$, qui ressemble à ce que présente le tableau 1 pour la population UScities.

3.4 Plans avec taille d'échantillon préétablie *n*

Avec l'algorithme de Kozak, il est possible de trouver les bornes qui minimisent le CV de $\bar{y}_s$ pour une taille fixe d'échantillon n plutôt que de minimiser n pour un CV préétabli. À titre d'exemple, reprenons les plans stratifiés du tableau 1. Les bornes de la méthode géométrique nous servent de valeurs initiales et nous exécutons l'algorithme de Kozak par défaut. Nous présentons ci-après le code R pour la population Debtors.

```
> pop1k <- strata.LH(x = Debtors, initbh = pop1$bh, n = 100,
  Ls = 5, alloc = c(0.5, 0, 0.5), algo = "Kozak")
```

Les CV de l'estimateur $\bar{y}_s$ avec les plans stratifiés optimaux sont de 3,12 %, 1,43 %, 1,72 % et 1,04 % respectivement pour les quatre populations comparativement à des valeurs de 3,59 %, 1,45 %, 1,83 % et 1,07 % au tableau 1. Ainsi, l'algorithme itératif permet de réduire les CV.

4. Stratification avec moments anticipés

On peut tenir compte d'une différence entre la variable de stratification X et la variable d'enquête Y en créant un modèle pour la distribution conditionnelle de Y étant donné X. Dans le programme *stratification*, il existe un modèle loglinéaire où

$$Y = \exp(\alpha)X^{\beta}\exp(\sigma\epsilon),$$

ainsi qu'un modèle linéaire hétéroscédastique avec

$$Y = \alpha + \beta X + \sigma\epsilon X^{\gamma}, \qquad (2)$$

α, β, et γ sont des paramètres réels spécifiés par l'utilisateur et ϵ est une variable aléatoire $N(0, 1)$. Un modèle de remplacement aléatoire (Rivest 1999) est également disponible, et il est possible d'ajouter des taux de mortalité par strate (Baillargeon, Rivest et Ferland 2007) au modèle loglinéaire.

Dans ces modèles, la moyenne anticipée de Y pour les unités classées dans la strate h avec $X \in [b_{h-1}, b_h)$ est

$$\bar{Y}_h = \frac{1}{N_h}\sum_{b_{h-1}\le X_i<b_h} E(Y \mid X_i)$$

alors que la variance anticipée est

$$S_{yh}^2 = \frac{1}{N_h}\sum_{b_{h-1}\le X_i<b_h}\{E(Y \mid X_i) - \bar{E}(Y \mid X)_h\}^2 + \frac{1}{N_h}\sum_{b_{h-1}\le X_i<b_h}\mathrm{Var}(Y \mid X_i)$$

où $\bar{E}(Y \mid X)_h$ est la moyenne des valeurs prédites de Y pour les unités de la strate h. Dans les fonctions strata.cumrootf, strata.geo et strata.bh, ces expressions servent à évaluer les propriétés échantillonnales de $\bar{y}_s$, alors que, dans la fonction strata.LH, la minimisation de (1) se fait avec les moments anticipés. Dans strata.LH, les bornes des strates dépendent du modèle pour la relation entre X et Y ; ce n'est pas le cas pour les autres fonctions strata.

4.1 Exemple avec les municipalités suédoises *MU*284

À la section 2.5, nous avons construit deux plans stratifiés pour la population $MU284$ avec $REV84$ comme variable de stratification. Le code R qui suit permet d'examiner la performance de ces plans pour la variable $RMT85$. Le vecteur ord contient la position des statistiques d'ordre de la variable $REV84$. Ainsi, Y[ord] est le vecteur de la variable $RMT85$ ordonné selon $REV84$.

```
> data(Sweden)
> X <- Sweden$REV84
> Y <- Sweden$RMT85
> ord <- order(X)
> geo_rmt <- var.strata(geo_cer, y = Y[ord])
> cum_rmt <- var.strata(cum_cer, y = Y[ord])
> c(geo_rmt$RRMSE,cum_rmt$RRMSE)
[1] 0.06889558 0.07368794
```

À la section 2.4, les CV de l'estimateur $\overline{y}_s$ pour la variable de stratification $REV84$ sont inférieurs à 5 % dans le cas des plans cum $\sqrt{f}$ et géométrique. Pour l'estimation de la moyenne de $RMT85$, les CV sont de plus de 6 %. On voit bien qu'un calcul des tailles d'échantillon avec une variable de stratification vient sous-estimer le n nécessaire à la réalisation du CV cible pour une variable d'intérêt différente. Ces résultats sont présentés pour les deux premiers plans stratifiés au tableau 8, lequel indique également le plan optimal obtenu en appliquant l'algorithme de Kozak à la variable $REV84$ avec $Y = X$.

Tout comme Rivest (2002), nous ajustons un modèle loglinéaire pour la relation entre les deux variables. Comme on peut le voir à la figure 2, il existe des valeurs extrêmes et le code R qui suit permet d'estimer les paramètres du modèle loglinéaire en écartant les municipalités pour lesquelles le rapport X/Y est atypique. Les 18 municipalités mises de coté sont représentées par un astérisque à la figure 2. On ajuste ensuite un modèle de régression aux données restantes à l'aide de la fonction `lm`.

```
> keep <- (X/Y > quantile(X/Y,0.03)) & (X/Y < quantile(X/Y,0.97))
> reg <- lm(log(Y)[keep] ~ log(X)[keep])
> coef(reg)
(Intercept) log(X)[keep]
-3.153025   1.058355
> summary(reg)$sigma
[1] 0.25677
```

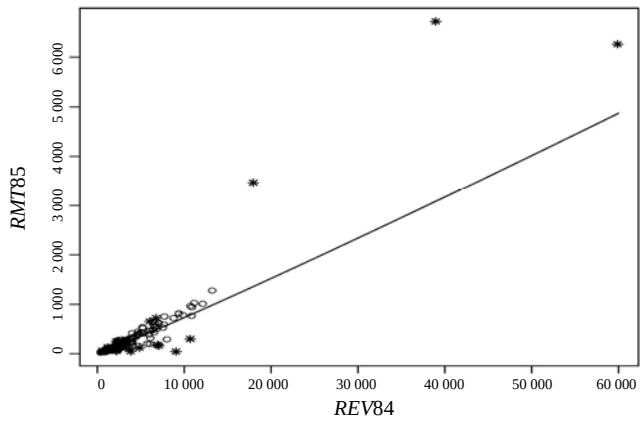

Figure 2 Graphique de dispersion de $RMT85$ par $REV84$ pour les données sur les municipalités suédoises

Le code qui suit stratifie la population $MU284$ par $REV84$ à l'aide des méthodes cum $\sqrt{f}$ et géométrique. On procède toutefois à l'allocation avec les moments anticipés calculés à l'aide du modèle de régression loglinéaire de $RMT85$ sur $REV84$. Dans ces deux plans, les strates sont les mêmes que celles qui ont été calculées précédemment. Le modèle influe seulement sur les CV prévus. Tel n'est pas le cas pour le plan optimal où les moments anticipés sont employés dans l'algorithme de stratification. L'algorithme de Kozak pourrait ne pas trouver le minimum global n dans ce cas. Nous utilisons donc les bornes calculées avec $Y = X$ comme valeurs initiales.

```
> geo_cer.m <- strata.geo(x = X[ord], CV = 0.05, Ls = 4,
  alloc = c(0.35, 0.35, 0), model = "loglinear",
  certain = (length(X) - 2):length(X), model.control =
  list(beta = 1.058355, sig2 = 0.25677^2))
> geo_cer.var <- var.strata(geo_cer.m, y = Y[ord])
> cum_cer.m <- strata.cumrootf(x = X[ord], nclass = 50,
  CV = 0.05, Ls = 4, alloc = c(0.35, 0.35, 0),
  certain = (length(X) - 2):length(X), model = "loglinear",
  model.control = list(beta = 1.058355, sig2 = 0.25677^2))
> cum_cer.var <- var.strata(cum_cer.m, y = Y[ord])
> LH <- strata.LH(x = X, CV = 0.05, Ls = 5,
  alloc = c(0.35, 0.35, 0), takeall = 1)
> LH.var <- var.strata(LH, y = Y)
> LH_m <- strata.LH(x = X, CV = 0.05, Ls = 5,
  initbh = LH$bh, alloc = c(0.35, 0.35, 0), takeall = 1,
  model = "loglinear", model.control = list(beta = 1.058355,
  sig2 = 0.25677^2))
> LH_m.var <- var.strata(LH_m, y = Y)
```

Au tableau 8, les tailles d'échantillon calculées avec les moments anticipés donnent des CV de moins de 5 % pour l'estimation de la moyenne de la variable $RMT85$. Le plan optimal LH exige un n un peu inférieur à ceux des deux autres. Si on tient compte de $Y \neq X$ quand on minimise (1), on obtient une strate à tirage complet plus grande, sa taille passant de 4 à 5.

Tableau 8
Trois plans stratifiés pour l'estimation de la moyenne de $RMT85$ avec $REV84$ comme variable de stratification

Modèle	Méthode		1	2	3	4	5	n	CV anticipé
$Y = X$	cum $\sqrt{f}$	N_h	127	79	46	29	3		
		n_h	3	4	4	5	3	19	7,37
	géométrique	N_h	42	116	88	35	3		
		n_h	2	5	7	7	3	24	6,89
	LH	N_h	120	82	45	33	4		
		n_h	3	4	4	5	4	20	7,11
loglinéaire	cum $\sqrt{f}$	N_h	127	79	46	29	3		
		n_h	6	8	9	10	3	36	4,78
	géométrique	N_h	42	116	88	35	3		
		n_h	3	8	13	13	3	40	4,74
	LH	N_h	121	81	45	32	5		
		n_h	6	7	7	9	5	34	4,90

Notons enfin que les arguments `model` et `model.control` peuvent être employés avec `var.strata`. Pour le plan géométrique considéré ici, on peut obtenir des résultats fort semblables à ceux qu'on obtient avec l'argument `y=Y`. Tel que montré ci-dessous, le modèle donne un CV de 6,894 % comparativement à 6,890 % pour la variable originale $RMT85$. Dans le cas de la méthode cum $\sqrt{f}$, le CV du modèle est de 7,282 % par rapport à 7,369 % calculé plus haut. Pour l'algorithme de Lavallée et Hidiroglou, les deux valeurs s'établissent à 7,080 % et 7,110 %.

```
> geo_rmt2 <- var.strata(geo_cer, model = "loglinear",
  model.control = list(beta = 1.058355, sig2 = 0.25677^2))
> geo_rmt2$RRMSE

[1] 0.0689368
```

4.2 Exemple tiré de Anderson, Kish et Cornell (1976) avec la distribution normale à deux variables

Dans Anderson *et coll.* (1976), on examine une stratification optimale pour Y étant donné X lorsque (X, Y) suit une distribution normale à deux variables avec corrélation ρ. Ceci correspond au modèle (2) avec $\alpha = \gamma = 0$, $\beta = \rho$, et $\sigma^2 = 1 - \rho^2$, où X a une distribution $N(0, 1)$. Pour reproduire les résultats de Anderson *et coll.* (1976), nous tirons une population de taille $N = 10^5$ d'une distribution $N(0, 1)$ et sélectionnons `model="linear"` (comme à la section 3.1, une moyenne de 10 empêche X de prendre des valeurs négatives). Dans le cas d'un modèle linéaire, seul l'algorithme de Kozak fonctionne. Ce problème étant particulier, nous réglons à 20 le paramètre `maxstep` et n'exécutons qu'une répétition (`rep=1`) de l'algorithme. En cas d'absence de strate à tirage complet, les limites optimales des strates sont indépendantes du CV comme à la section 3.1. Nous avons choisi CV = 0,01 dans les calculs.

```
> x <- rnorm(1e+05, 10)
> bi3a <- strata.LH(x = x, CV = 0.01, Ls = 3, takenone = 0,
  model = "linear",
  model.control = list(beta = 0.25, sig2 = 1 - 0.25^2,
  gamma = 0), algo.control = list(maxstep = 20, rep = 1))
> bi3a$bh – 10

[1] -0.619354 0.604198
```

Au tableau 9, les résultats du programme *stratification* sont égaux à ceux de Anderson *et coll.* (1976) jusqu'à près de deux décimales. Cela fait bien voir la polyvalence de ce programme qui peut trouver le plan stratifié optimal pour toute distribution de la variable de stratification et pour un certain nombre de modèles généraux pour la distribution conditionnelle de Y étant donné X.

Tableau 9
Comparaison des bornes optimales de strates dans Anderson *et coll.* (1976) et des bornes approximatives obtenues avec *stratification*

		Résultats de *stratification*				Résultats de Anderson *et coll.*			
L	**|ρ|**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
3	0,250	-0,619	0,604			-0,61	0,61		
	0,950	-0,591	0,568			-0,58	0,58		
	0,990	-0,571	0,549			-0,56	0,56		
4	0,250	-0,984	0,004	0,985		-0,98	0,00	0,98	
	0,950	-0,930	0,009	0,942		-0,93	0,00	0,93	
	0,990	-0,902	-0,001	0,895		-0,90	0,00	0,90	
5	0,250	-1,245	-0,377	0,387	1,251	-1,24	-0,38	0,38	1,24
	0,950	-1,187	-0,358	0,372	1,197	-1,19	-0,37	0,37	1,19
	0,990	-1,136	-0,344	0,353	1,144	-1,14	-0,35	0,35	1,14

5. Autres caractéristiques

Baillargeon et Rivest (2009) se sont attachés à d'autres aspects d'un plan stratifié, soit aux taux de non-réponse prévus par strate et à l'ajout d'une strate à tirage nul qui n'est pas échantillonnée. Dans cette section, nous examinerons brièvement comment ces aspects sont pris en charge par *stratification*. On se doit de tenir compte de la non-réponse dans le calcul de n. Une strate à tirage nul fait que $\bar{y}_s$ est entaché d'un biais, auquel cas on spécifie le degré de précision cible sous forme d'erreur quadratique moyenne relative (EQMR) plutôt que de CV. La formule (4.3) de Baillargeon et Rivest (2009) donne une généralisation de (1) comprenant ces deux caractéristiques. C'est cette formule qui est utilisée pour le calcul des tailles d'échantillon dans la procédure d'optimisation.

5.1 Non-réponse

On peut apporter une correction de non-réponse *a posteriori* en divisant les tailles d'échantillon de strates sans non-réponse par les taux de réponse. C'est ce qui est illustré par le code R qui suit où figure la variable MRTS représentative de l'Enquête mensuelle sur le commerce de détail de Statistique Canada. On apporte postérieurement les corrections de non-réponse dans la fonction `var.strata` avec l'argument `rh.postcorr=TRUE`. Une autre possibilité est de tenir compte des taux de réponse dans la répartition de l'échantillon entre les strates, ce qui peut être spécifié dans une fonction `strata` par l'argument `rh=`. Avec cette manière de procéder, on pénalise les strates à taux élevé de non-réponse, puisqu'on se retrouve normalement avec une valeur de n moindre que dans le cas des corrections *a posteriori*. C'est ce qui est illustré dans la partie cum $\sqrt{f}$ du tableau 10. Avec quatre strates et des taux de réponse respectifs de 0,8, 0,8, 0,9 et 1, la correction *a posteriori* doit viser n = 445 pour la réalisation du CV cible comparativement à n = 444 pour une attribution tenant compte de la non-réponse.

```
> data(MRTS)
> cum <- strata.cumrootf(x = MRTS, nclass = 500, CV = 0.01,
  Ls = 4, alloc = c(0.5, 0, 0.5))
> cum.var <- var.strata(cum, rh = c(0.8, 0.8, 0.9, 1))
> cum.post <- var.strata(cum, rh = c(0.8, 0.8, 0.9, 1),
  rh.postcorr = TRUE)
> cum_rh <- strata.cumrootf(x = MRTS, nclass = 500, CV = 0.01,
  Ls = 4, alloc = c(0.5, 0, 0.5), rh = c(0.8, 0.8, 0.9, 1))
```

On peut également tenir compte de la non-réponse lorsqu'on élabore un plan stratifié optimal soit *a posteriori* soit dans la construction des strates. Ces deux stratégies sont présentées pour la population MRTS dans le code R qui suit. Les taux plus élevés de non-réponse pour les petites unités pénalisent la première strate qui sera moindre en cas de prise en compte de la non-réponse dans l'algorithme de stratification, comme on peut le voir au tableau 10. En tenant compte de la non-réponse dans la construction des strates, on

obtient une valeur de n moindre que dans le traitement *a posteriori*. Au tableau 3 de Baillargeon et Rivest (2009), on trouvera d'autres exemples, incluant les moments anticipés ainsi que la non-réponse, de l'élaboration de plans stratifiés pour la population MRTS.

```
> LH <- strata.LH(x = MRTS, CV = 0.01, Ls = 4,
  alloc = c(0.5, 0, 0.5), takeall = 1)
> LH.var <- var.strata(LH, rh = c(0.8, 0.8, 0.9, 1))
> LH.post <- var.strata(LH, rh = c(0.8, 0.8, 0.9, 1),
  rh.postcorr = TRUE)
> LH_rh <- strata.LH(x = MRTS, CV = 0.01, Ls = 4,
  alloc = c(0.5, 0, 0.5), takeall = 1, rh = c(0.8, 0.8, 0.9, 1))
```

Tableau 10
Deux exemples de correction de non-réponse soit *a posteriori* (*post*) soit dans l'élaboration du plan

Méthode	rh		1	2	3	4	*n*	CV anticipé
cum $\sqrt{f}$	aucun	N_h	778	742	355	125		
		n_h	87	90	88	125	390	1,11
		n_h^{post}	109	113	98	125	445	1,00
	donné	N_h	778	742	355	125		
		n_h	105	108	106	125	444	1,00
LH	aucun	N_h	774	675	374	177		
		n_h	77	65	60	177	379	1,11
		n_h^{post}	96	81	67	177	421	1,00
	donné	N_h	675	677	449	199		
		n_h	70	69	80	199	418	1,00

5.2 Strate à tirage nul

Une strate à tirage nul qui n'est pas échantillonnée peut présenter un avantage lorsqu'une population comporte de petites unités dont les valeurs Y sont proches de 0. On mesure alors la précision de $\bar{y}_s$ par l'erreur quadratique moyenne $\mathrm{Var}(\bar{y}_s) + (T_{0y}/N)^2$, où T_{0y} est le total prévu pour Y dans la strate à tirage nul. En prenant `takenone=1` dans la fonction `strata.LH`, on se trouve à élaborer un plan optimal avec une strate à tirage nul. Baillargeon et Rivest (2009) ont démontré que l'algorithme de Sethi ne fonctionne pas en pareil cas et qu'il faut utiliser l'algorithme de Kozak. Si on fait intervenir une strate à tirage nul, on peut apporter une correction brute de biais en divisant $\bar{y}_s$ par la proportion du total de la variable X dans les strates à tirage partiel. Ainsi, la pénalisation de biais dans l'erreur quadratique moyenne pourrait être trop sévère et une autre mesure de précision comme $\mathrm{Var}(\bar{y}_s) + (p \times T_{0y}/N)^2$, pourrait être utilisée dans l'algorithme de stratification, p étant un chiffre dans l'intervalle $(0, 1)$. Cette pénalisation de biais plus légère peut s'opérer en posant `bias.penalty` égal à p. Le code R qui suit élabore trois plans stratifiés optimaux pour la population MRTS avec et sans strate à tirage nul ; la pleine pénalisation pour le biais par défaut est alors à comparer à une pénalisation plus légère avec $p = 0{,}5$.

```
> data(MRTS)
> notn <- strata.LH(x = MRTS, CV = 0.1, Ls = 3,
  alloc = c(0.5, 0, 0.5))
> tn1 <- strata.LH(x = MRTS, CV = 0.1, Ls = 3,
  alloc = c(0.5, 0, 0.5), takenone = 1)
> tn0.5 <- strata.LH(x = MRTS, CV = 0.1, Ls = 3,
  alloc = c(0.5, 0, 0.5), takenone = 1, bias.penalty = 0.5)
```

On pourra voir au tableau 11 les tailles d'échantillon n pour les trois plans. Si on prévoit une strate à tirage nul avec pleine pénalisation de biais, n baisse de 22 à 16. Dans ce plan, la strate à tirage nul rend compte de 3 % du total pour la variable X. En ramenant la pénalisation pour le biais à $p = 0{,}5$, on augmente la taille de la strate à tirage nul et diminue n. D'autres exemples cités au tableau 2 de Baillargeon et Rivest (2009) montrent que la taille de la strate à tirage nul décroît habituellement avec l'EQMR cible. Dans l'exemple MRTS, l'ajout d'une telle strate réduit largement la valeur de n, alors que le plan ne change pas dans d'autres cas.

Tableau 11
Tailles d'échantillon de trois plans stratifiés optimaux pour la population MRTS

`takenone=`	0	1	1
`bias.penalty=`	S.O.	1	0,5
n	22	16	13
% T_x	0	3	9

6. Conclusion

Le programme R *stratification* offre des méthodes souples permettant d'élaborer des plans d'échantillonnage stratifiés à l'aide d'une variable unidimensionnelle de stratification comme une mesure de taille dans une enquête auprès des entreprises. Nous disposons de plusieurs méthodes pour établir les bornes et les tailles d'échantillon des strates. Ce programme permet de considérer plusieurs caractéristiques comme les strates à tirage complet ou à tirage nul, le degré de non-correspondance entre X et Y et une non-réponse par strate.

Remerciements

Nous sommes reconnaissants envers S. Er, E. Gagnon, M. Kozak et J. Stardom pour toutes leurs observations constructives concernant ce programme de stratification, ainsi qu'à la Chaire de recherche du Canada en échantillonnage statistique et en analyse de données et au Conseil de recherches en sciences naturelles et en génie du Canada de l'aide financière qu'ils ont apportée. Ces travaux ont été financés par une bourse du U.S. National Science Foundation (SES-0751671).

7. Annexe

7.1 Description plus détaillée de l'algorithme de Kozak

Comme nous l'avons indiqué à la section 3.3, l'algorithme de Kozak effectue une recherche aléatoire des bornes. Pour être acceptée, une nouvelle borne doit non seulement faire baisser le critère d'optimisation, soit n ou

l'EQMR de $\bar{y}_s$, mais aussi faire en sorte que les strates à tirage partiel contiennent au moins `minNh` unités et que les tailles d'échantillon soient positives. La valeur par défaut de `minNh` est 2. Nous disposons également d'un algorithme de Kozak à traitement non aléatoire avec `method= "modified"` dans l'argument `algo.control`. Il essaie toutes les modifications de bornes possibles à une itération et sélectionne celle qui fait le plus baisser le critère d'optimisation. Il est plus lent que l'algorithme de Kozak sans améliorer pour autant la détection du minimum global de ce critère. Nous ne discuterons donc pas davantage de cette approche.

Pour illustrer ce qu'est le traitement complet des solutions possibles mentionné à la section 3.3, considérons l'ensemble de données `USbanks` qui contient 357 valeurs, dont seulement 200 sont uniques. Si nous voulons répartir cette population en deux strates, il n'y a que $\binom{200-1}{2-1} = 199$ solutions qui soient possibles. Avec la commande qui suit, nous pouvons énumérer toutes les solutions possibles :

```
> enum <- strata.LH(x = USbanks, CV = 0.05, Ls = 2,
  alloc = c(0.5, 0, 0.5))
```

On peut trouver ces solutions avec leur valeur correspondante de critère d'optimisation dans `enum$sol.detail`. Seules les solutions satisfaisant aux contraintes d'admissibilité précitées y sont incluses.

Dans l'exécution de l'algorithme de Kozak, les valeurs initiales des bornes pourraient ne pas respecter les contraintes d'admissibilité et l'algorithme risquerait alors de ne pas fonctionner du tout. Dans ce cas, les bornes initiales seraient remplacées par des bornes robustes. Celles-ci donnent une strate à tirage nul vide si une telle strate est demandée. Les strates à tirage complet sont les plus petites possible et les strates à tirage partiel ont à peu près le même nombre de valeurs uniques de X.

Prenons une fois de plus l'exemple de la section 3.2 avec l'ensemble de données `UScities` où l'algorithme de Kozak a trouvé un minimum local avec les arguments par défaut. S'il utilise les bornes initiales géométriques, cet algorithme converge rapidement vers ce qui semble être un minimum global.

```
> LH_init <- strata.LH(x = UScities, initbh = pop2$bh,
  n = 100, Ls = 5, alloc = c(0.5, 0, 0.5), takeall = 0,
  algo.control = list(rep = 1))
> LH_init$iter.detail

      b1    b2    b3   b4          opti   step   iter   run
1   18.5  33.5  59.5  107    0.01444981      0      0     1
2   20.5  33.5  59.5  107    0.01435576      2      2     1
3   19.5  33.5  59.5  107    0.01434272     -1     10     1
4   19.5  33.5  58.0  107    0.01432714     -1     12     1
5   19.5  31.5  58.0  107    0.01431013     -2     13     1
6   19.5  32.5  58.0  107    0.01430163      1     63     1

> LH_init$niter
[1] 163
```

L'élément `LH_init$iter.detail` donne en sortie les bornes initiales et celles des cinq itérations avec modification de bornes seulement. En tout 163 itérations ont été nécessaires pour que l'algorithme converge. Les bornes initiales géométriques sont très proches des solutions optimales. On peut également éviter un minimum local en modifiant certains des paramètres de l'algorithme. Le code R qui suit permet un accroissement des pas (`maxstep=20`) et augmente le nombre maximum d'itérations (`maxstill=1000`) ainsi que le nombre de répétitions de l'algorithme (`rep=20`).

```
> LH_param <- strata.LH(x = UScities, n = 100, Ls = 5,
  alloc = c(0.5, 0, 0.5), takeall = 0, algo.control =
  list(maxstep = 20, maxstill = 1000, rep = 20))
```

Les résultats des 20 répétitions figurent dans `LH_param$rep.detail` et sont résumés au tableau 12. La solution obtenue avec les bornes initiales géométriques se réalise 9 fois sur 20.

Tableau 12
Solutions obtenues par l'algorithme de Kozak pour 20 répétitions

CV	B1	B2	B3	B4	Fréquence
0,0143	19,50	32,50	58,00	107,00	9
0,0167	16,50	23,50	37,50	78,00	5
0,0167	15,50	22,50	35,50	73,00	6

La figure 3 montre que, par un accroissement des pas, l'algorithme peut plus facilement atteindre le minimum global (CV = 0,0143) que s'il emploie les arguments par défaut (pointillés, CV = 0,0167).

7.2 Tableau récapitulatif du programme R *stratification*

Dans cette annexe, nous résumons le programme *stratification*. Le tableau 13 énumère les cinq fonctions constitutives avec leurs arguments. Les notes qui suivent complètent le tableau.

(1) Dans la formule générale d'allocation (Hidiroglou et Srinath 1993), les tailles d'échantillon des strates sont proportionnelles à $N_h^{2q_1} \bar{Y}_h^{2q_2} S_{yh}^{2q_3}$.

(2) La valeur par défaut de `initbh` est l'ensemble de points de départ arithmétiques de Gunning et Horgan (2007) ; voir la section 3.3 à ce sujet. Avec `takenone=1` et si `initbh` est de la taille `Ls-1`, la borne initiale de la strate à tirage nul est fixée au premier percentile de X. Si ce dernier est égal à la valeur minimale de X, cette borne initiale donnerait une strate à tirage nul vide. Dans ce cas on fixe plutôt la borne initiale de la strate à tirage nul à la deuxième plus petite valeur de X.

(3) Les éléments à spécifier dans l'argument `algo.control` dépendent de l'algorithme. Le tableau qui suit indique les éléments employés par chaque algorithme avec leurs valeurs par défaut. On trouvera dans `help(strata.LH)` une description complète de chaque élément.

Algorithme	maxiter	method	minNh	maxstep	maxstill	rep	minsol
Sethi	500	-	-	-	-	-	-
Kozak initial	10 000	"original"	2	3	100	3	1 000
Kozak modifié	3 000	"modified"	2	3	-	-	1 000

(4) Les éléments de l'argument `model.control` dépendent du modèle :

- modèle loglinéaire avec mortalité (`loglinear`) :

$$Y = \begin{cases} \exp(\alpha + \texttt{beta}\,\log(X) + \texttt{epsilon}) & \text{avec probabilité } p_h \\ 0 & \text{avec probabilité } 1-p_h \end{cases}$$

où $\texttt{epsilon} \sim N(0, \texttt{sig2})$ est indépendant de X. Le paramètre p_h est spécifié par `ph`, `ptakenone` et `pcertain`.

- modèle linéaire hétéroscédastique (`linear`) :

$$Y = \texttt{beta}\,X + \texttt{epsilon}$$

où

$$\texttt{epsilon} \sim N(0, \texttt{sig2}\,X^{\texttt{gamma}}).$$

- modèle de remplacement aléatoire (`random`) :

$$Y = \begin{cases} X & \text{avec probabilité } 1 - \texttt{epsilon} \\ X\text{new} & \text{avec probabilité } \texttt{epsilon} \end{cases}$$

où Xnew est une variable aléatoire indépendante de X avec la même distribution que X.

Le tableau qui suit présente les valeurs par défaut de `model.control` selon le modèle.

Modèle	beta	sig2	ph	ptakenone	pcertain	gamma	epsilon
"loglinear"	1	0	rep(1,Ls)	1	1	-	-
"linear"	1	0	-	-	-	0	-
"random"	-	-	-	-	-	-	0

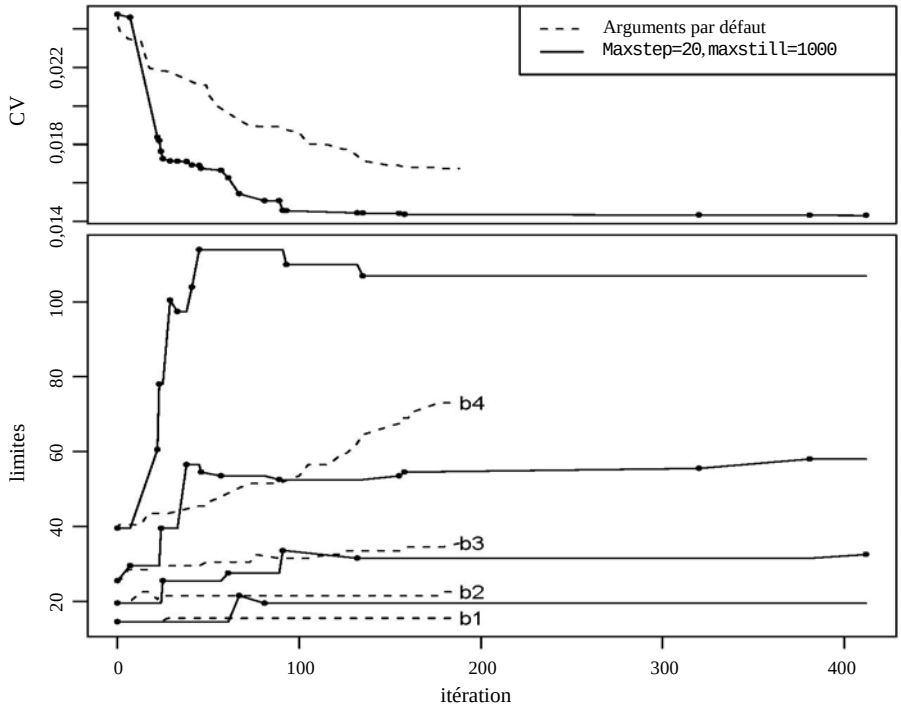

Figure 3 Succession des itérations pour deux exécutions de l'algorithme de Kozak

Tableau 13
Tableau récapitulatif du programme R *stratification*

Argument	Strata.cumrootf	Strata.geo	Strata.LH	Strata.bh	Var.strata	Description	Format	Valeur par défaut
x	•	•	•	•		variable de stratification	vecteur	aucune (x est obligatoire)
n	•	•	•	•		taille totale d'échantillon cible	scalaire	aucune (n ou CV est obligatoire)
CV	•	•	•	•		CV ou EQMR cible	scalaire	aucune (n ou CV est obligatoire)
Ls	•	•	•	•		nombre de strates échantillonnées	scalaire	3
alloc	•	•	•	•		règle d'allocation (1)	liste (q1, q2, q3) où qi ≥ 0	Neyman (q1=q3=0.5, q2=0)
certain	•	•	•	•		indices des x pour les unités à échantillonner obligatoirement	vecteur	NULL (pas de strate à tirage obligatoire)
nclass	•					nombre de classes	scalaire	*min*(10*L*, *N*)
bh				•		bornes des strates	vecteur	aucune (bh est obligatoire)
takeall.adjust				•		indicateur de correction pour strates à tirage complet	TRUE ou FALSE	FALSE (aucune correction)
takeall			•	•		nombre de strates à tirage complet	un nombre parmi {0, 1,.., Ls − 1}	0
initbh			•			bornes initiales des strates (2)	vecteur	bornes équidistantes
algo			•			identification de l'algorithme	"Kozak" ou "Sethi"	"Kozak"
algo.control			•			spécification des paramètres de l'algorithme (3)	liste (maxiter, method, minNh, maxstep, maxstill, rep)	dépend de algo
strata					•	plan stratifié	objet de classe strata	aucune (strata est obligatoire)
y					•	variable d'étude	vecteur	NULL (un model doit être fourni)
model	•	•	•	•	•	identification du modèle	"none", "loglinear", "linear"* ou "random"* →	"none" (* indisponible avec l'algo de Sethi)
model.control	•	•	•	•	•	spécification des paramètres du modèle (4)	liste (beta, sig2, ph, ptakenone, gamma, epsilon)	dépend du modèle, mais équivaut à model="none"
rh	•	•	•	•	•	taux de réponse anticipés	scalaire ou vecteur	rep(1,Ls) ou rh tiré de strata
rh.postcorr					•	indicateur de correction postérieure pour non-réponse	TRUE ou FALSE	FALSE (aucune correction)
takenone			•	•		nombre de strates à tirage nul	0 ou 1	0
bias.penalty			•	•		pénalité pour le biais	scalaire	1

Bibliographie

Anderson, D.W., Kish, L. et Cornell, R.G. (1976). Quantifying gains from stratification for optimum and approximately optimum strata using a bivariate normal model. *Journal of the American Statistical Association*, 71, 887-892.

Baillargeon, S., Rivest, L.-P. et Ferland, M. (2007). Stratification en enquêtes entreprises : une revue et quelques avancées. *Recueil de la Section des méthodes d'enquêtes, Société Statistique du Canada* (www.ssc.ca/survey/documents/SSC2007_S_Baillargeon.pdf).

Baillargeon, S., et Rivest, L.-P. (2009). A general algorithm for univariate stratification. *Revue Internationale de Statistique*, 77, 331-344.

Cochran, W.G. (1961). Comparison of methods for determining stratum boundaries. *Bulletin of the International Statistical Institute*, 32, 345-358.

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

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

Dayal, S. (1985). Allocation of sample using values of auxiliary characteristics. *Journal of Statistical Planning and Inference*, 11, 321-328.

Detlefsen, R.E., et Veum, C.S. (1991). Design issues for the Retail Trade Sample Surveys of the U.S. Bureau of the Census. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 214-219.

Gunning, P., et Horgan, J.M. (2004). Un nouvel algorithme pour la construction de bornes de stratification dans les populations asymétriques. *Techniques d'enquête*, 30, 177-185.

Gunning, P., et Horgan, J.M. (2007). Improving the Lavallée and Hidiroglou algorithm for stratification of skewed populations. *Journal of Statistical Computation and Simulation*, 77, 277-291.

Hidiroglou, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.

Hidiroglou, M.A., et Srinath, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economic Statistics*, 11, 397-405.

Khan, M.G.M., Nand, N. et Ahmad, N. (2008). Détermination des bornes optimales de strate au moyen de la programmation dynamique. *Techniques d'enquête*, 34, 227-236.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6, 797-806.

Kozak, M., et Verma, M.R. (2006). Approche de la stratification par une méthode géométrique et par optimisation : une comparaison de l'efficacité. *Techniques d'enquête*, 32, 177-183.

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

McEvoy, R.H. (1956). Variation in bank asset portfolios. *The Journal of Finance*, 11(4), 463-473.

Rivest, L.-P. (1999). Stratum jumpers: Can we avoid them?. *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association, (Alexandria, VA), 64-72.

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

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

Sethi, V.K. (1963). A note on the optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5, 20-33.

Sigman, R.S., et Monsour, N.J. (1995). Selecting samples from list frames of businesses. Dans *Business Survey Methods*, (Éds., B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.L. Colledge et P.S. Kott), 133-152.

Slanta, J., et Krenzke, T. (1996). Utilisation de la méthode de Lavallée et Hidiroglou pour le calcul des limites de stratification aux fins de l'enquête annuelle sur les dépenses en capital du Bureau of the Census. *Techniques d'enquête*, 22, 65-75.

Sweet, E.M., et Sigman R.S. (1995). Evaluation of model-assisted procedures for stratifying skewed populations using auxiliary data. *U.S. Bureau of the Census* (www.census.gov/srd/papers/pdf/sm95-22.pdf).

Techniques d'enquête, juin 2011
Vol. 37, N° 1, pp. 73-81
Statistique Canada, N° 12-001-X au catalogue

Estimation de la variance par répliques sous échantillonnage à deux phases

Jae Kwang Kim et Cindy Long Yu [1]

Résumé

Dans l'échantillonnage à deux phases pour la stratification, l'échantillon de deuxième phase est sélectionné selon un plan stratifié basé sur l'information observée sur l'échantillon de première phase. Nous élaborons un estimateur de variance corrigé du biais fondé sur une méthode de répliques qui étend la méthode de Kim, Navarro et Fuller (2006). La méthode proposée est également applicable quand la fraction d'échantillonnage de première phase n'est pas négligeable et que le tirage de l'échantillon de deuxième phase se fait par échantillonnage de Poisson avec probabilités inégales dans chaque strate. La méthode proposée peut être étendue à l'estimation de la variance pour les estimateurs par la régression à deux phases. Les résultats d'une étude par simulation limitée sont présentés.

Mots clés : Échantillonnage double ; jackknife ; estimateur par la régression ; estimateur à facteur d'extension repondéré.

1. Introduction

L'échantillonnage à deux phases, introduit pour la première fois par Neyman (1938) et parfois appelé échantillonnage double, est une méthode d'échantillonnage rentable. On y recourt habituellement quand la collecte des données sur les variables d'intérêt coûte très cher, mais qu'il est relativement peu coûteux de recueillir des données sur des variables corrélées aux variables d'intérêt. L'échantillonnage à deux phases a des applications sous diverses formes (par exemple, Rao 1973 ; Cochran 1977 ; Breidt et Fuller 1993 ; Rao et Sitter 1995 ; Hidiroglou et Särndal 1998 ; Fuller 1998 ; Hidiroglou 2001 ; Fuller 2003). L'échantillonnage à deux phases pour la stratification fait référence à la situation où l'observation provenant de l'échantillon de première phase est utilisée pour procéder à une stratification pour l'échantillonnage de deuxième phase. Si l'on sélectionne l'échantillon de première phase pour les besoins de la stratification, l'échantillonnage de deuxième phase représente un outil utile lorsqu'il n'existe pas au départ de base de sondage pour la stratification. Par exemple, dans le cas des enquêtes sur les forêts, il est très difficile et coûteux de se rendre dans les régions éloignées pour faire des déterminations sur place. Par contre, les photographies aériennes sont relativement peu coûteuses, et les déterminations ayant trait, par exemple, au type de forêt d'après les photos aériennes sont fortement corrélées à celles faites au sol et peuvent être utilisées pour stratifier l'échantillon de première phase.

L'estimation de la variance par répliques est très répandue dans le cas des enquêtes complexes. Rust et Rao (1996) et Wolter (2007) donnent deux aperçus complets de ce sujet. La méthode de répliques ne requiert pas le calcul de la dérivée partielle du développement en série de Taylor et l'utilisateur peut facilement produire des estimations de variance sans connaître le plan d'échantillonnage qui a été utilisé pour recueillir les données. De surcroît, la tendance a recourir à cette méthode s'accentue à cause des questions de confidentialité (Lu et Sitter 2006). Une fois que les poids de rééchantillonnage sont fournis, il n'est pas nécessaire que l'utilisateur dispose d'informations sur le plan de sondage, telles que les identificateurs de strate, pour analyser les données.

Deux estimateurs de la moyenne de population sous échantillonnage à deux phases sont utilisés fréquemment, à savoir l'estimateur à facteur d'extension double (DEE pour double expansion estimator) et l'estimateur à facteur d'extension repondéré (REE pour reweighted expansion estimator) qui doivent leur nom à Kott et Stukel (1997). En général le REE est plus efficace que le DEE dans la situation d'échantillonnage à deux phases pour la stratification quand les y à l'intérieur d'une strate sont homogènes. L'estimation de la variance pour l'échantillonnage à deux phases est un problème compliqué en pratique et les praticiens s'intéressent aux méthodes de répliques pour l'estimation de la variance. Rao et Shao (1992) ont proposé un estimateur de variance par le jackknife convergent pour le REE dans le contexte de l'imputation hot deck en traitant les répondants comme un échantillon de deuxième phase. Kott et Stukel (1997) ont examiné le même problème et concluent que les estimateurs de variance jackknife donnent de bons résultats pour le REE si la fraction d'échantillonnage de première phase est négligeable. La fraction d'échantillonnage, $f_1 = nN^{-1}$, est dite négligeable si f_1 converge vers zéro sous les conditions asymptotiques décrites à la section 2. Binder, Babyak, Brodeur, Hidiroglou

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, Iowa 50011, États-Unis ; Cindy Long Yu, Department of Statistics, Center for Survey Statistics et Methodology, Iowa State University, Ames, Iowa 50011, États-Unis. Courriel : cindyyu@iastate.edu.

et Jocelyn (2000) ont étudié l'estimation de la variance pour un plan d'échantillonnage à deux phases similaire en utilisant la méthode de linéarisation de Taylor. Kim et coll. (2006, KNF) offrent une étude rigoureuse de la méthode de répliques et l'envisagent pour d'autres types d'estimateurs. La méthode KNF a été élaborée principalement dans les conditions où la fraction d'échantillonnage de première phase est négligeable et l'échantillonnage de deuxième phase est un échantillonnage aléatoire stratifié. Si la fraction d'échantillonnage de première phase n'est pas négligeable, des répliques supplémentaires sont nécessaires pour obtenir des estimations de variance convergentes.

Dans le présent article, nous proposons une nouvelle méthode de répliques pour l'estimation de la variance sous échantillonnage à deux phases. La méthode proposée est une extension de la méthode KNF qui englobe la situation où la fraction d'échantillonnage de première phase n'est pas nécessairement négligeable. Contrairement à la méthode KNF, la méthode proposée ne requiert pas de répliques supplémentaires pour la correction du biais dans l'estimation de la variance, mais nécessite des corrections des poids de rééchantillonnage. En outre, la méthode proposée est applicable à l'échantillonnage de Poisson avec probabilités inégales dans les strates de deuxième phase, lequel n'a pas été discuté dans KNF. La méthode fondée sur des répliques est très facile à mettre en œuvre et peut être appliquée à divers types d'estimateurs.

Le plan de l'article est le suivant. À la section 2, nous présentons les conditions de base et à la section 3, nous décrivons la méthode proposée. À la section 4, nous étendons la méthode proposée à d'autres estimateurs sous échantillonnage à deux phases. À la section 5, nous présentons les résultats d'une étude par simulation limitée. Enfin, à la section 6, nous présentons nos conclusions.

2. Conditions de base

Pour mieux exposer la motivation de l'étude, à la présente section, nous supposons simplement que l'échantillonnage de première phase correspond au tirage d'un échantillon aléatoire simple de taille n d'une population finie de taille N et que l'échantillon de deuxième phase est un échantillon aléatoire stratifié. À la section 3, nous étendons les conditions afin d'inclure tout échantillonnage arbitrairement mesurable à la première phase et l'échantillonnage de Poisson avec probabilités inégales dans chaque strate à la deuxième phase. Au moyen de l'information obtenue auprès de l'échantillon de première phase, ce dernier est stratifié en H strates pour l'échantillonnage de deuxième phase. Dans la strate h, nous avons n_h éléments de l'échantillon de première phase et soit A_{h1} l'ensemble d'indices pour les éléments de l'échantillon de première phase dans la strate h. À la deuxième phase, un échantillon aléatoire stratifié de taille r est sélectionné en tirant un échantillon de taille $r_h (\le n_h)$ dans la strate h, où $r = \sum_{h=1}^{H} r_h$ et la fraction d'échantillonnage r_h/n_h est fixe pour chaque strate. Pour discuter formellement de la théorie asymptotique, nous supposons que nous avons une suite de populations finies, une suite d'échantillons de première phase et une suite d'échantillons de deuxième phase, comme il est décrit dans KNF. Dans ces conditions asymptotiques, nous permettons que la taille de l'échantillon de deuxième phase r tende vers l'infini à la même vitesse que la taille de l'échantillon de première phase n, c'est-à-dire $r = O(n)$ et $r^{-1} = O(n^{-1})$, et H est fixe. Donc, dans les conditions où H est fixe, $r_h^{-1} = O(n^{-1})$.

Quand la variable étudiée y_i est observée dans l'échantillon de deuxième phase, la moyenne de population de y est estimée par

$$\bar{y}_{tp} = \frac{1}{n} \sum_{h=1}^{H} \sum_{i \in A_{h2}} \frac{n_h}{r_h} y_i,$$

où A_{h2} est l'ensemble d'indices pour les éléments de l'échantillon de deuxième phase qui appartiennent à la strate h. La variance de $\bar{y}_{tp}$ peut s'écrire

$$\mathrm{Var}(\bar{y}_{tp}) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 + E\left\{\sum_{h=1}^{H} \left(\frac{n_h}{n}\right)^2 \left(\frac{1}{r_h} - \frac{1}{n_h}\right) s_{h1}^2\right\} \quad (1)$$

où $\bar{y}_1 = n^{-1} \sum_{h=1}^{H} \sum_{i \in A_{h1}} y_i$, $S^2 = (N-1)^{-1} \sum_{i=1}^{N} (y_i - \bar{y}_N)^2$, $s_{h1}^2 = (n_h - 1)^{-1} \sum_{i \in A_{h1}} (y_i - \bar{y}_{h1})^2$, et $\bar{y}_{h1} = n_h^{-1} \sum_{i \in A_{h1}} y_i$. En utilisant

$$n^{-1} S^2 \doteq E\left\{n^{-1} \sum_{h=1}^{H} w_h [(\bar{y}_{h1} - \bar{y}_1)^2 + s_{h1}^2]\right\}$$

où $w_h = n^{-1} n_h$ et $\doteq$ indique une approximation en ignorant les termes d'ordre $o(n^{-1})$, le terme de variance (1) est approximé par

$$\mathrm{Var}(\bar{y}_{tp}) \doteq E\left\{n^{-1}(1 - f_1) \sum_{h=1}^{H} w_h (\bar{y}_{h1} - \bar{y}_1)^2 + \sum_{h=1}^{H} (r_h^{-1} - n_h^{-1} f_1) w_h^2 s_{h1}^2\right\}, \quad (2)$$

où $f_1 = nN^{-1}$.

Un estimateur convergent de la variance de $\bar{y}_{tp}$ peut être dérivé de (2) en remplaçant $\bar{y}_{h1}$ et s_{h1}^2 par leurs estimations $\bar{y}_{h2} = r_h^{-1} \sum_{i \in A_{h2}} y_i$ et $s_{h2}^2 = (r_h - 1)^{-1} \sum_{i \in A_{h2}} (y_i - \bar{y}_{h2})^2$, respectivement. Autrement dit, un estimateur de variance convergent est donné par

$$\hat{V} = n^{-1}(1 - f_1) \sum_{h=1}^{H} w_h (\bar{y}_{h2} - \bar{y}_2)^2 + \sum_{h=1}^{H} (r_h^{-1} - n_h^{-1} f_1) w_h^2 s_{h2}^2, \quad (3)$$

où $\bar{y}_2 = \sum_{h=1}^{H} w_h \bar{y}_{h2}$. L'estimateur de variance (3) est un estimateur de variance linéarisé.

Kott et Stukel (1997) et KNF ont élaboré un estimateur de variance jackknife en supprimant successivement des unités de l'échantillon complet de première phase, puis en ajustant les poids. Les répliques jackknife complètes sont

$$\bar{y}_{tp}^{(k)} = \frac{1}{N}\sum_{h=1}^{H} \hat{N}_{h1}^{(k)} \bar{y}_{h2}^{(k)} \tag{4}$$

où k est l'indice de l'unité supprimée dans la réplique jackknife,

$$\frac{1}{N}\hat{N}_{h1}^{(k)} = \sum_{i\in A_{h1}} w_i^{(k)} = \begin{cases} (n-1)^{-1}(n_h - 1) & \text{si } k \in A_{h1} \\ (n-1)^{-1} n_h & \text{si } k \notin A_{h1} \end{cases}$$

et

$$\bar{y}_{h2}^{(k)} = \frac{\sum_{i\in A_{h2}} w_i^{(k)} y_i}{\sum_{i\in A_{h2}} w_i^{(k)}} = \begin{cases} (r_h - 1)^{-1}(r_h \bar{y}_{h2} - y_k) & \text{si } k \in A_{h2} \\ \bar{y}_{h2} & \text{si } k \notin A_{h2}. \end{cases} \tag{5}$$

L'estimateur de variance jackknife complet de la forme

$$\hat{V}_J = \sum_{k\in A_1} \frac{n-1}{n}(1 - f_1)(\bar{y}_{tp}^{(k)} - \bar{y}_{tp})^2, \tag{6}$$

où $\bar{y}_{tp}^{(k)}$ est défini dans (4), est asymptotiquement équivalent à

$$\hat{V}_J \doteq n^{-1}(1 - f_1)\sum_{h=1}^{H} w_h(\bar{y}_{h2} - \bar{y}_2)^2 + (1 - f_1)\sum_{h=1}^{H} r_h^{-1} w_h^2 s_{h2}^2. \tag{7}$$

Donc, si nous comparons (7) à (2), le biais de l'estimateur de variance jackknife (6) est

$$\text{Biais}(\hat{V}_J) \doteq -E\left\{ f_1 \sum_{h=1}^{H} (r_h^{-1} - n_h^{-1}) s_{h2}^2 \right\}.$$

Par conséquent, si la fraction d'échantillonnage de première phase est négligeable au sens où $f_1 \doteq 0$, le biais est négligeable, c'est-à-dire que le biais $= o(n^{-1})$. Sinon, l'estimateur de variance sous-estime la variance.

Afin d'examiner une méthode du jackknife corrigée du biais, au lieu de (5), nous considérons

$$\bar{y}_{h2}^{(k)} = \begin{cases} (r_h - \delta_h)^{-1}(r_h \bar{y}_{h2} - \delta_h y_k) & \text{si } k \in A_{h2} \\ \bar{y}_{h2} & \text{si } k \notin A_{h2}, \end{cases} \tag{8}$$

où δ_h doit être déterminé. In (5), nous avons utilisé $\delta_h = 1$. L'estimateur de variance jackknife en utilisant (8) au lieu de (5) est asymptotiquement équivalent à

$$\hat{V}_J \doteq n^{-1}(1 - f_1)\sum_{h=1}^{H} w_h(\bar{y}_{h2} - \bar{y}_2)^2 + (1 - f_1)\sum_{h=1}^{H} \frac{(r_h - 1)\delta_h^2}{(r_h - \delta_h)^2} w_h^2 s_{h2}^2.$$

Donc, le biais asymptotique est donné par

$$\text{Biais}(\hat{V}_J) \doteq E\left[\sum_{h=1}^{H} \left\{ (1 - f_1)\frac{(r_h - 1)\delta_h^2}{(r_h - \delta_h)^2} - \frac{1}{r_h}\left(1 - f_1 \frac{r_h}{n_h}\right) \right\} w_h^2 s_{h2}^2 \right].$$

Le biais asymptotique est nul si

$$\delta_h = \frac{r_h}{1 + \sqrt{r_h(r_h - 1)}/d_h}$$

où $d_h = \sqrt{(1 - f_1 r_h n_h^{-1})/(1 - f_1)}$. D'où, avec δ_h déterminé de cette façon dans (8), l'estimateur de variance jackknife résultant est approximativement sans biais sans émettre l'hypothèse que $f_1 \doteq 0$.

3. Méthode proposée

Nous étendons maintenant la méthode proposée à la section 2 à un plan d'échantillonnage de première phase plus général. Pour cela, nous devons émettre l'hypothèse que l'estimateur de variance par répliques de la forme

$$\hat{V}_1 = \sum_{k=1}^{L} c_k(\hat{\theta}^{(k)} - \hat{\theta})^2,$$

où $\hat{\theta} = \sum_{i\in A_1} w_i y_i$, et $\hat{\theta}^{(k)} = \sum_{i\in A_1} w_i^{(k)} y_i$, est convergent pour la variance de $\hat{\theta}$ sous le plan d'échantillonnage à un seul (de premier) degré. Autrement dit,

$$\frac{\hat{V}_1}{\text{Var}(\hat{\theta})} - 1 = o_p(1). \tag{9}$$

Ici L est le nombre de répliques. Pour la plupart des plans de sondage mesurables, c'est-à-dire les plans dont toutes les probabilités d'inclusion conjointe sont positives, nous pouvons construire l'estimateur de variance par répliques qui satisfait (9) même quand la fraction d'échantillonnage $f = n/N$ est grande. Par exemple, voir Fay (1984) et Flyer (1987). Brick et Morganstein (1996) décrivent l'algorithme de base pour WesVar, un logiciel disponible dans le commerce pour l'estimation de la variance par répliques dans le cas des sondages.

À la présente section, nous envisageons également un cas plus intéressant d'échantillonnage stratifié avec probabilités inégales pour la deuxième phase. Plus précisément, nous considérons que l'échantillon de deuxième phase est un échantillon de Poisson avec probabilités inégales dans les strates de deuxième phase. Fuller (1998) a également examiné l'échantillonnage de Poisson à la deuxième phase et soutenu qu'il s'agit d'une bonne approximation. Un exemple de cette situation dans le contexte des enquêtes sur les forêts est que, en plus des types de forêt, les interprètes des photos peuvent aussi déterminer, d'après les photos aériennes prises à la première phase, la densité des arbres et la hauteur des arbres, qui peuvent être utilisées pour construire les probabilités de sélection de deuxième phase dans chaque strate (type de forêt).

À la présente section, nous nous concentrons d'abord sur l'estimateur de type REE, puisqu'il est plus efficace que celui de type DEE, et nous discutons de l'extension à l'estimateur DEE à la section 4. Soit w_i, le poids d'échantillonnage de première phase et soit w_{i2}, l'inverse la probabilité conditionnelle à la deuxième phase. C'est-à-dire, $w_{i2} = \pi_{i2}^{-1}$, où $\pi_{i2} = \Pr(i \in A_{h2} \mid i \in A_{h1})$. L'estimateur de type REE peut s'écrire

$$\overline{y}_{tp} = \frac{1}{N}\sum_{h=1}^{H} \hat{N}_{h1}\, \overline{y}_{h2} \tag{10}$$

où $\hat{N}_{h1}=\sum_{i\in A_{h1}} w_i$ et $\overline{y}_{h2}=(\sum_{i\in A_{h2}} w_i\pi_{i2}^{-1})^{-1}\sum_{i\in A_{h2}} w_i\pi_{i2}^{-1}y_i$. Dans KNF, π_{i2} est supposée constante dans la strate de deuxième phase.

Nous considérons une approche fondée sur les répliques pour l'estimation de la variance de l'estimateur de type REE (10) quand π_{i2} n'est pas nécessairement constante dans la strate de deuxième phase. Nous considérons le cas particulier où le plan d'échantillonnage de deuxième phase est un échantillonnage de Poisson. En utilisant la méthode de répliques satisfaisant (9), l'estimateur de variance de type KNF peut être appliqué à l'estimation de la variance de $\overline{y}_{tp}$ dans cette situation. Autrement dit,

$$\hat{V}_{\text{KNF}} = \sum_{k=1}^{L} c_k (\overline{y}_{tp}^{(k)} - \overline{y}_{tp})^2, \tag{11}$$

où

$$\overline{y}_{tp}^{(k)} = \frac{1}{N}\sum_{h=1}^{H} \hat{N}_{h1}^{(k)}\, \overline{y}_{h2}^{(k)} \tag{12}$$

avec $\overline{y}_{h2}^{(k)} = (\sum_{i\in A_{h2}} w_i^{(k)}\pi_{i2}^{-1})^{-1}\sum_{i\in A_{h2}} w_i^{(k)}\pi_{i2}^{-1}y_i$ et $\hat{N}_{h1}^{(k)} = \sum_{i\in A_{h1}} w_i^{(k)}$, et c_k est un facteur associé à la réplique k déterminé par la méthode de répliques. Sous échantillonnage de Poisson à la deuxième phase, nous avons le biais asymptotique suivant :

$$\text{Biais}(\hat{V}_{\text{KNF}}) = -\frac{1}{N^2}\sum_{h=1}^{H}\sum_{i\in U_h} \pi_{i2}^{-1}(1-\pi_{i2})(y_i - \overline{Y}_h)^2, \tag{13}$$

où U_h est l'ensemble d'indices des éléments de population dans la strate h et $\overline{Y}_h = N_h^{-1}\sum_{i\in U_h} y_i$. Une esquisse de preuve de (13) est présentée à l'annexe A.

Un estimateur asymptotiquement sans biais du biais (13) est donné par

$$\hat{V}_{\text{biais}} = -\frac{1}{N^2}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i\, \pi_{i2}^{-2}(1-\pi_{i2})(y_i - \overline{y}_{h2})^2. \tag{14}$$

Le biais est négligeable si $n/N \doteq 0$. Donc, nous pouvons sans risque ignorer le biais de l'estimateur de variance de type KNF quand la fraction d'échantillonnage de première phase est négligeable. Si cette fraction n/N n'est pas négligeable, le biais peut être arbitrairement grand. KNF discutent aussi d'une méthode de rééchantillonnage corrigée du biais en utilisant des répliques supplémentaires, ce qui peut donner lieu à un grand nombre de répliques. Créer des répliques supplémentaires pour corriger le biais est parfois fastidieux dans le cas d'enquêtes à grande échelle.

Nous considérons une autre option d'estimateur de variance par répliques corrigé du biais qui ne nécessite pas la création de répliques supplémentaires. Pour élaborer un estimateur de variance corrigé du biais fondé sur une méthode de répliques, définissons une variable aléatoire

$$\delta_{ki} \overset{\text{indep}}{\sim} \text{Bernoulli}(p_k), \tag{15}$$

où p_k doit être déterminée. Soit

$$\hat{V}_{\text{KNF}}^{*} = \sum_{k=1}^{L} c_k (\overline{y}_{tp}^{*(k)} - \overline{y}_{tp})^2 \tag{16}$$

où

$$\overline{y}_{tp}^{*(k)} = \frac{1}{N}\sum_{h=1}^{H} \hat{N}_{h1}^{(k)}\, \overline{y}_{h2}^{*(k)} \tag{17}$$

avec $\hat{N}_{h1}^{(k)} = \sum_{i\in A_{h1}} w_i^{(k)}$,

$$\overline{y}_{h2}^{*(k)} = \frac{\sum_{i\in A_{h2}} w_i^{(k)} M_{i2}^{(k)} \pi_{i2}^{-1} y_i}{\sum_{i\in A_{h2}} w_i^{(k)} M_{i2}^{(k)} \pi_{i2}^{-1}} \tag{18}$$

avec

$$M_{i2}^{(k)} = 1 + (\delta_{ki} - p_k) b_i \tag{19}$$

et b_i doit également être déterminé. Par construction, $E_*(\delta_{ki} - p_k) = 0$, où E_* indique que l'espérance est considérée par rapport au mécanisme précisé dans (15). Donc, les répliques (18) créent une variation supplémentaire dans les poids de rééchantillonnage, la variation supplémentaire dans (18) provenant de la distribution (15). Un choix approprié de p_i et b_i peut rendre convergent l'estimateur de variance résultant.

Sous les conditions de régularité discutées dans KNF, nous avons

$$E_*(\hat{V}^*_{\mathrm{KNF}}) = \hat{V}_{\mathrm{KNF}} + N^{-2}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i^2\, b_i^2\, \pi_{i2}^{-2}\, u\,(y_i - \bar{y}_{h2})^2 + o_p\,(n^{-1}), \tag{20}$$

où $u = \sum_{k=1}^{L} c_k\, p_k\,(1-p_k)$. Une esquisse de preuve de (20) est présentée à l'annexe B. Si b_i est déterminé par

$$b_i = \sqrt{(1-\pi_{i2})\, w_i^{-1} u^{-1}}, \tag{21}$$

l'estimateur de variance (16) est convergent, parce que le deuxième terme de (20) annule $\hat{V}_{\mathrm{bias}}$ dans (14). Il en est ainsi même quand la fraction d'échantillonnage de première phase n/N n'est pas négligeable. Pour garantir que les poids de rééchantillonnage soient non négatifs dans (18), nous exigeons que b_i dans (19) soit ≤ 1. Si nous posons $p_k = 0{,}5$, alors

$$b_i = \sqrt{\frac{4\,(1-\pi_{i2})\, w_i^{-1}}{\sum_{k=1}^{L} c_k}},$$

qui est inférieur ou égal à 1 si $\sum_{k=1}^{L} c_k \geq 4$. En fait, les p_k peuvent être égales à n'importe quel nombre compris entre 0 et 1, à condition que le b_i résultant dans (21), soit inférieur ou égal à 1.

4. Extensions

À la présente section, nous considérons l'extension de la méthode de répliques proposée à d'autres types d'estimateurs à deux phases que le REE donné en (10).

4.1 Estimateur à facteur d'extension double

En échantillonnage à deux phases, on se sert également de l'estimateur à facteur d'extension double (*double expansion estimator*), qui doit son nom à Kott et Stukel (1997). L'estimateur à facteur d'extension double (DEE) possède la forme simple

$$\bar{y}_{\mathrm{DEE}} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i\, \pi_{i2}^{-1}\, y_i. \tag{22}$$

Quand l'échantillon de deuxième phase est un échantillon aléatoire stratifié, $\pi_{i2} = r_h/n_h$ et la méthode KNF peut être appliquée en utilisant la réplique

$$\bar{y}^{(k)}_{\mathrm{DEE}} = \frac{1}{N}\sum_{h=1}^{H}\left(\frac{\sum_{i\in A_{h1}} w_i^{(k)}\, w_i^{-1}}{\sum_{i\in A_{h2}} w_i^{(k)}\, w_i^{-1}}\right)\sum_{i\in A_{h2}} w_i^{(k)}\, y_i.$$

L'estimateur de variance KNF pour l'estimateur DEE est convergent quand la fraction d'échantillonnage de première phase est négligeable. Si elle ne l'est pas, nous pouvons utiliser la méthode de répliques proposée à la section 3. La méthode de répliques proposée pour l'estimateur DEE crée les répliques

$$\bar{y}^{*(k)}_{\mathrm{DEE}} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i^{(k)}\, w_{i2}^{*(k)}\, y_i, \tag{23}$$

où

$$w_{i2}^{*(k)} = M_{i2}^{(k)}\frac{\sum_{i\in A_{h1}} w_i^{(k)}\, w_i^{-1}}{\sum_{i\in A_{h2}} w_i^{(k)}\, w_i^{-1}\, M_{i2}^{(k)}},$$

et $M_{i2}^{(k)}$ est le facteur de répliques défini en (19). Le biais de l'estimateur de variance par répliques en utilisant la réplique (23) est négligeable si les répliques sont construites de façon à satisfaire (21).

Si l'échantillon de deuxième phase est obtenu par échantillonnage avec probabilités inégales dans chaque strate, la méthode de répliques telle que (23) ne peut pas être appliquée directement. L'estimateur DEE donné par (22) est généralement moins efficace que l'estimateur REE donné par (10). Notons que ce dernier peut être exprimé sous la forme

$$\bar{y}_{\mathrm{REE}} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i\, w_{i2}^{*}\, y_i, \tag{24}$$

où

$$w_{i2}^{*} = \pi_{i2}^{-1}\frac{\sum_{i\in A_{h1}} w_i}{\sum_{i\in A_{h2}} w_i\, \pi_{i2}^{-1}}. \tag{25}$$

Les répliques (17) peuvent s'écrire

$$\bar{y}^{*(k)}_{\mathrm{REE}} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i^{(k)}\, w_{i2}^{*(k)}\, y_i, \tag{26}$$

où

$$w_{i2}^{*(k)} = M_{i2}^{(k)}\pi_{i2}^{-1}\frac{\sum_{i\in A_{h1}} w_i^{(k)}}{\sum_{i\in A_{h2}} w_i^{(k)} M_{i2}^{(k)}\pi_{i2}^{-1}} \tag{27}$$

et $M_{i2}^{(k)}$ est défini dans (19).

4.2 Estimateur par la régression

En échantillonnage à deux phases, les variables auxiliaires observées dans l'échantillon de première phase peuvent être utilisées à l'étape de l'estimation. L'estimateur par la régression à deux phases du total de population peut s'écrire sous la forme

$$\hat{Y}_{t,\,\mathrm{REG}} = \hat{\mathbf{T}}'_{x,1}\,\hat{\boldsymbol{\beta}}_2 \tag{28}$$

où $\hat{\mathbf{T}}_{x,1} = \sum_{i\in A_1} w_i\,\mathbf{x}_i$ est le vecteur des totaux de population de la variable de contrôle $\mathbf{x}_i$ estimés sur l'échantillon de première phase, $\hat{\boldsymbol{\beta}}_2 = (\sum_{i\in A_2} w_i w_{i2}^{*}\mathbf{x}_i\mathbf{x}_i')^{-1}\sum_{i\in A_2} w_i w_{i2}^{*}\mathbf{x}_i y_i$ est un vecteur de coefficients de régression estimés sur l'échantillon de deuxième phase et w_{i2}^{*} est donné par (25). Notons que l'estimateur par la régression donné par (28) peut

intégrer l'échantillonnage stratifié dans la deuxième phase si $\mathbf{x}_i$ englobe le vecteur des indicateurs de strate.

En utilisant les arguments de la section 3, nous pouvons construire la k^{e} réplique pour $\hat{Y}_{t,\text{REG}}$ par

$$\hat{Y}_{t,\text{REG}}^{(k)} = \hat{\mathbf{T}}_{x,1}^{(k)\prime}\hat{\boldsymbol{\beta}}_2^{(k)}, \tag{29}$$

où

$$\hat{\mathbf{T}}_{x,1}^{(k)} = \sum_{i\in A_1} w_i^{(k)}\mathbf{x}_i$$

$$\hat{\boldsymbol{\beta}}_2^{(k)} = \left(\sum_{i\in A_2} w_i^{(k)} w_{i2}^{*(k)}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i\in A_2} w_i^{(k)}\, w_{i2}^{*(k)}\,\mathbf{x}_i\, y_i$$

et $w_{i2}^{*(k)}$ est défini dans (27).

La méthode de répliques (29) peut être appliquée directement à l'estimateur par calage à deux phases dont ont discuté Hidiroglou et Särndal (1998). Si $H = 1$, alors la réplique de $\hat{\boldsymbol{\beta}}_2$ donnée par (29) se réduit à

$$\hat{\boldsymbol{\beta}}_2^{(k)} = \left(\sum_{i\in A_2} w_i^{(k)} M_{i2}^{(k)}\pi_{i2}^{-1}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i\in A_2} w_i^{(k)}\, M_{i2}^{(k)}\,\pi_{i2}^{-1}\,\mathbf{x}_i\, y_i.$$

5. Étude par simulation

Afin d'étudier la performance en échantillon fini des estimateurs proposés, nous avons procédé à une étude par simulation limitée. Dans la simulation, nous avons commencé par générer une population finie artificielle de taille $N = 1\,000$ avec cinq variables $(z_i, q_i, x_i, y_i, u_i)$, où les éléments de population sont générés indépendamment à partir de $z_i \sim \exp(1) + 2$; $q_i \sim \chi^2(1) + 2$; $x_i \sim N(2, 1)$; $u_i \sim \text{Unif}\{1, 2, 3, 4\}$, où $\text{Unif}\{1, ..., G\}$ désigne une loi uniforme discrète avec support $\{1, ..., G\}$; et

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 q_i + e_i$$

avec $(\beta_0, \beta_1, \beta_2, \beta_3) = (0, 2, 1, 1)$ et $e_i \sim N(0, 1)$. The variables z_i, q_i, x_i, u_i, et e_i sont mutuellement indépendantes. Nous avons défini la strate pour l'échantillon de deuxième phase en utilisant la variable u_i. Nous nous sommes servis de la variable x_i pour calculer l'estimateur par la régression à deux phases (28) avec $\mathbf{x}_i = (1, x_i)'$, de la variable z_i comme mesure de taille pour l'échantillonnage avec probabilités inégales dans l'échantillonnage de première phase, et de la variable q_i comme mesure de taille pour l'échantillonnage avec probabilités inégales dans l'échantillonnage de deuxième phase.

Pour obtenir les échantillons sélectionnés avec probabilités inégales de cette étude par simulation, nous avons utilisé l'échantillonnage de Poisson ou l'échantillonnage de Rao-Sampford (Rao 1965 et Sampford 1967), avec des probabilités de sélection proportionnelles à la mesure de la variable de taille. Notons que la taille d'échantillon finale est aléatoire sous l'échantillonnage de Poisson, mais est fixe sous l'échantillonnage de Rao-Sampford.

Nous avons employé pour la simulation une structure factorielle $2 \times 3 \times 2$ avec trois facteurs, qui sont :

1. Échantillonnage pour l'échantillon de première phase (2) : Échantillonnage aléatoire simple de taille $n = 200$ contre échantillonnage de Rao-Sampford de taille $n = 200$ en utilisant z_i comme mesure de taille.
2. Échantillonnage pour l'échantillon de deuxième phase (3) : Échantillonnage aléatoire stratifié de taille $r_h = 25$, échantillonnage de Poisson stratifié avec taille d'échantillon prévue $r_h = 25$ en utilisant q_i comme mesure de taille pour l'échantillonnage avec probabilités inégales, et échantillonnage de Rao-Sampford stratifié de taille $r_h = 25$ en utilisant q_i comme mesure de taille pour l'échantillonnage avec probabilités inégales.
3. Méthodes d'estimation de la variance (2) : Estimateur KNF (11) sans répliques supplémentaires contre estimateur de variance proposé en utilisant (16), calculé en se basant sur la méthode du jackknife.

À partir de la population finie générée ci-dessus, nous avons produit $B = 5\,000$ échantillons Monte Carlo indépendants pour la simulation. Dans le cas des plans de sondage avec échantillonnage de Rao-Sampford dans la première phase, nous avons utilisé la méthode d'estimation jackknife de la variance proposée par Berger (2007), qui donne un estimateur convergent de la variance d'échantillonnage de première phase. Le paramètre d'intérêt est la moyenne de population de la variable y. Pour chaque échantillon Monte Carlo, nous avons calculé deux estimateurs ponctuels, à savoir le REE donné par (24) et l'estimateur par la régression (REG) donné par (28) en utilisant la variable auxiliaire $(1, x_i)$. Nous avons calculé les biais relatifs des estimateurs de variance en divisant le biais Monte Carlo de l'estimateur de variance par la variance Monte Carlo de l'estimateur ponctuel.

Le tableau 1 donne la moyenne et la variance des deux estimateurs ponctuels. Pour l'estimation ponctuelle, l'estimateur par la régression est significativement plus efficace que l'estimateur REE pour cette population, car la variable auxiliaire x est corrélée avec la variable étudiée y. La variance asymptotique théorique de l'estimateur par la régression sous échantillonnage aléatoire simple à la première phase et sous échantillonnage aléatoire stratifié à la deuxième phase est approximativement égale à

$$\left(\frac{1}{200} - \frac{1}{1\,000}\right)8 + \left(\frac{1}{100} - \frac{1}{200}\right)4 = 0{,}052$$

et la variance asymptotique théorique de l'estimateur REE pour le même plan d'échantillonnage est, approximativement, $(1/100 - 1/1\,000)\,8 = 0{,}072$, qui concorde avec les

résultats numériques du tableau 1. L'échantillonnage de Rao-Sampford à la deuxième phase est légèrement plus efficace que l'échantillonnage de Poisson en raison de la taille d'échantillon fixe dans le premier cas.

Le tableau 2 donne le biais relatif (BR) et le coefficient de variation (CV) des deux estimateurs de variance. Nous avons calculé les biais relatifs des estimateurs de variance en divisant le biais Monte Carlo de l'estimateur de variance par la variance Monte Carlo de l'estimateur ponctuel. Pour calculer le coefficient de variation de l'estimateur de variance, nous avons divisé l'erreur-type Monte Carlo de l'estimateur de variance par la moyenne Monte Carlo de l'estimateur de variance.

Dans cette simulation, comme la fraction d'échantillonnage de première phase n'est pas négligeable ($n/N = 0{,}2$), l'estimateur de variance KNF sans répliques supplémentaires sous-estime la variance réelle et l'estimateur de variance proposé estime la variance avec un biais plus petit, inférieur à 3 % en valeur absolue dans tous les cas, ce qui est en harmonie avec la théorie des sections 3 et 4. La valeur absolue du biais relatif dans l'estimateur de variance KNF est grande car, même si la variance due à $\hat{\mathbf{T}}_{x1}$ est estimée avec convergence dans (29), la variance due à $\hat{\boldsymbol{\beta}}_2$ est sous-estimée sans les répliques supplémentaires. Le biais relatif dans notre estimeur proposé est réduit parce que les répliques (18) créent une variation supplémentaire dans les poids de rééchantillonnage par l'ajout d'une perturbation δ_k tirée d'une loi choisie convenablement. Le CV de l'estimateur de variance proposé est un peu plus grand que celui de l'estimateur KNF, parce que le premier comporte un caractère aléatoire supplémentaire dû à la génération de δ_{ki} d'après (15).

Tableau 1
Moyenne et variance des estimateurs ponctuels (5 000 échantillons)

Estimateur	Échantillonnage de première phase	Échantillonnage de deuxième phase	Moyenne	Variance
REE	EAS	EAS Str.	10,0	0,0749
		Pois. Str.	10,0	0,0784
		RS Str.	10,0	0,0754
	RS	EAS Str.	10,0	0,0768
		Pois. Str.	10,0	0,0827
		RS Str.	10,0	0,0781
REG	EAS	EAS Str.	10,0	0,0540
		Pois. Str.	10,0	0,0510
		RS Str.	10,0	0,0495
	RS	EAS Str.	10,0	0,0551
		Pois. Str.	10,0	0,0531
		RS Str.	10,0	0,0515

REE : Estimateur à facteur d'extension repondéré (23)
REG : Estimateur par la régression (27)
EAS : Échantillonnage aléatoire simple
RS : Échantillonnage de Rao-Sampford
EAS Str. : Échantillonnage aléatoire simple stratifié
Pois. Str. : Échantillon de Poisson stratifié
RS Str. : Échantillonnage de Rao-Sampford stratifié.

Tableau 2
Biais relatif (BR) et coefficient de variation (CV) pour les estimateurs de variance (5 000 échantillons)

Méthode	Estimateur	Échantionnage de première phase	Échantillonnage de deuxième phase	BR (%)	CV(%)
KNF	REE	EAS	EAS Str.	-11,25	18,22
			Pois. Str.	-9,56	18,67
			RS Str.	-7,75	15,35
		RS	EAS Str.	-8,05	18,61
			Pois. Str.	-9,03	20,84
			RS Str.	-5,73	17,27
	REG	EAS	EAS Str.	-6,76	22,32
			Pois. Str.	-6,06	15,81
			RS Str.	-3,26	12,82
		RS	EAS Str.	-4,17	21,74
			Pois. Str.	-3,64	16,92
			RS Str.	-3,20	13,78
Nouveau	REE	EAS	EAS Str.	0,09	18,23
			Pois. Str.	-1,23	19,70
			RS Str.	-0,04	16,06
		RS	EAS Str.	0,78	19,78
			Pois. Str.	-2,07	21,26
			RS Str.	1,00	17,67
	REG	EAS	EAS Str.	-0,61	22,00
			Pois. Str.	-0,57	16,55
			RS Str.	-0,08	13,36
		RS	EAS Str.	0,67	22,86
			Pois. Str.	-0,01	16,97
			RS Str.	0,59	14,02

KNF : Estimateur de variance de Kim et coll. (2006) sans répliques supplémentaires pour la correction du biais
Nouveau : Estimateur de variance proposé (16)
REE : Estimateur à facteur d'extension repondéré (23)
REG : Estimateur par la régression (27)
SRS : Échantillonnage aléatoire simple
RS : Échantillonnage de Rao-Sampford
EAS Str. : Échantillonnage aléatoire simple stratifié
Pois. Str. : Échantillonnage de Poisson stratifié
RS Str. : Échantillonnage de Rao-Sampford stratifié.

6. Conclusion

L'estimation de la variance par répliques sous échantillonnage à deux phases représente un problème pratique important dans les sondages et la méthode KNF est un outil utile à cet égard. Dans le présent article, nous proposons un méthode qui est une extension de la méthode KNF en ce sens qu'elle s'applique directement quand la fraction d'échantillonnage n'est pas négligeable, sans accroître le nombre de répliques. La méthode proposée s'applique aussi à l'échantillonnage de Poisson avec probabilités inégales dans chaque strate à l'étape de l'échantillonnage de deuxième phase. La théorie a été élaborée uniquement sous échantillonnage de Poisson à la deuxième phase, mais les résultats de simulation présentés à la section 5 indiquent que la méthode proposée donne des résultats raisonnablement bons pour d'autres plans d'échantillonnage avec probabilités inégales, tels que l'échantillonnage de Rao-Sampford.

Puisque la méthode de répliques proposée fournit des estimateurs de variance convergents pour les moyennes de population, elle peut être appliquée facilement à d'autres paramètres de population finie qui sont des fonctions lisses des moyennes de population.

Dans le cas de certaines grandes enquêtes, le nombre de répliques peut être assez élevé, parce que la méthode utilise le même nombre de répliques pour l'échantillon de première phase. Si l'on souhaite réduire le nombre de répliques, la méthode de Fuller (1998) ou celle de Kim et Sitter (2003) est une option. Un examen plus approfondi suivant cette piste sera le sujet d'une prochaine étude.

Remerciements

Les travaux de recherche ont été financés par l'entente de coopération n° 68-3A75-4-122 conclue entre le Natural Resources Conservation Service de l'USDA et le Center for Survey Statistics and Methodology de l'Iowa State University. Les auteurs remercient Wayne Fuller et deux examinateurs anonymes de leurs commentaires constructifs.

Annexe

A. Preuve de (13)

Soit $\mathbf{a} = (a_1, ..., a_N)$ où a_i est la version étendue de l'indicateur d'échantillonnage de deuxième phase discutée dans Kim et coll. (2006). C'est-à-dire, $a_i = 1$, si l'unité i est sélectionnée dans l'échantillon de deuxième phase après avoir été sélectionnée dans l'échantillon de première phase et $a_i = 0$ autrement.

En vertu de l'hypothèse (9), conditionnellement à $\mathbf{a}$, nous avons

$$\sum_{k=1}^{L} c_k (\bar{y}_{h2}^{(k)} - \bar{y}_{h2})^2 = \mathrm{Var}(\bar{y}_{h2} \mid \mathbf{a}) + o_p(n^{-1}).$$

Donc, le biais de $\sum_{k=1}^{L} c_k (\bar{y}_{h2}^{(k)} - \bar{y}_{h2})^2$ en tant qu'estimateur de $\mathrm{Var}(\bar{y}_{h2})$ est égal, en ignorant les termes d'ordre $o(n^{-1})$, à

$$E\{\mathrm{Var}(\bar{y}_{h2} \mid \mathbf{a})\} - \mathrm{Var}(\bar{y}_{h2}) = \mathrm{Var}\{E(\bar{y}_{h2} \mid \mathbf{a})\}.$$

En utilisant la définition étendue de a_i, nous avons

$$E(\bar{y}_{h2} \mid \mathbf{a}) = \frac{\sum_{i \in U_h} \pi_{i2}^{-1} a_i y_i}{\sum_{i \in U_h} \pi_{i2}^{-2} a_i}$$

et, en vertu de l'hypothèse d'échantillonnage de Poisson des a_i,

$$\mathrm{Var}\left(\frac{\sum_{i \in U_h} \pi_{i2}^{-1} a_i y_i}{\sum_{i \in U_h} \pi_{i2}^{-1} a_i} \right) = N_h^{-2} \sum_{i \in U_h} \pi_{i2}^{-1} (1 - \pi_{i2}) (y_i - \bar{Y}_h)^2 + o(N^{-1}). \quad \text{(A.1)}$$

Donc, le biais de l'estimateur de variance KNF est de la forme (13) sous l'hypothèse d'échantillonnage de Poisson de a_i.

B. Preuve de (20)

Pour chaque k,

$$\bar{y}_{tp}^{*(k)} - \bar{y}_{tp} = \bar{y}_{tp}^{*(k)} - \bar{y}_{tp}^{(k)} + \bar{y}_{tp}^{(k)} - \bar{y}_{tp},$$

où $\bar{y}_{tp}^{(k)}$ est défini par (12). Donc,

$$\hat{V}_{\mathrm{KNF}}^{*} = \sum_{k=1}^{L} c_k (\bar{y}_{tp}^{*(k)} - \bar{y}_{tp})^2 = \sum_{k=1}^{L} c_k (\bar{y}_{tp}^{(k)} - \bar{y}_{tp})^2 + 2 \sum_{k=1}^{L} c_k (\bar{y}_{tp}^{(k)} - \bar{y}_{tp})(\bar{y}_{tp}^{*(k)} - \bar{y}_{tp}^{(k)}) + \sum_{k=1}^{L} c_k (\bar{y}_{tp}^{*(k)} - \bar{y}_{tp}^{(k)})^2. \quad \text{(B.1)}$$

En vertu de la construction de $\bar{y}_{tp}^{*(k)}$, nous avons

$$E_*(\bar{y}_{tp}^{*(k)}) = \bar{y}_{tp}^{(k)} + o_p(n^{-1}). \quad \text{(B.2)}$$

En outre, en écrivant $q_{ki} = M_{i2}^{(k)} - 1$, nous avons $q_{ki} = O_p(n^{-1/2})$ et nous pouvons appliquer un développement en série de Taylor pour obtenir

$$\bar{y}_{h2}^{*(k)} = \bar{y}_{h2}^{(k)} + \frac{\sum_{i \in A_{h2}} w_i^{(k)} \pi_{i2}^{-1} q_{ki} (y_i - \bar{y}_{h2}^{(k)})}{\sum_{i \in A_{h2}} w_i^{(k)} \pi_{i2}^{-1}} + o_p(n^{-1}). \quad \text{(B.3)}$$

En outre, comme

$$\frac{1}{N_h} \sum_{i \in A_{h2}} w_i^{(k)} \pi_{i2}^{-1} z_i - \frac{1}{N_h} \sum_{i \in A_{h2}} w_i \pi_{i2}^{-1} z_i = O_p(n^{-1})$$

pour toute variable z dont les quatrièmes moments sont bornés, nous pouvons montrer que (B.3) se réduit à

$$\bar{y}_{h2}^{*(k)} = \bar{y}_{h2}^{(k)} + \frac{\sum_{i \in A_{h2}} w_i \pi_{i2}^{-1} q_{ki} (y_i - \bar{y}_{h2})}{\sum_{i \in A_{h2}} w_i \pi_{i2}^{-1}} + o_p(n^{-1}).$$

D'où, nous pouvons écrire

$$\sum_{k=1}^{L} c_k (\bar{y}_{tp}^{*(k)} - \bar{y}_{tp}^{(k)})^2 = \sum_{k=1}^{L} c_k \left\{ N^{-1} \sum_{h=1}^{H} \sum_{i \in A_{h2}} w_i \pi_{i2}^{-1} q_{ki} (y_i - \bar{y}_{h2}) \right\}^2 + o_p(n^{-1}). \quad \text{(B.4)}$$

En insérant (B.2) et (B.4) dans (B.1), nous avons

$$E_*(\hat{V}^*_{\mathrm{KNF}}) = \hat{V}_{\mathrm{KNF}} + \frac{1}{N^2}\sum_{k=1}^{L} c_k \sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i^2 E_*(q_{ki}^2)\pi_{i2}^{-2}(y_i - \overline{y}_{h2})^2 + o_p(n^{-1}),$$

et, comme $E_*(q_{ki}^2) = p_k(1-p_k)\,b_i^2$, nous avons (20).

Bibliographie

Berger, Y.G. (2007). A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika*, 94, 953-964.

Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M. et Jocelyn, W. (2000). Variance estimation for two-phase stratified sampling. *The Canadian Journal of Statistics*, 28, 751-764.

Breidt, F.J. et Fuller, W.A. (1993). Regression weighting for multipurpose samplings. *Sankhyā*, B, 55, 297-309.

Brick, J.M., et Morganstein, D. (1996). WesVarPC: Software for computing variance estimates from complex designs. *Proceedings of the 1996 Annual Research Conference*, U.S. Bureau of the Census, 861-866.

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

Fay, R.E. (1984). Some properties of estimates of variance based on replication methods. *Proceedings of the Survey Research Method Section*, American Statistical Association, 495-500.

Flyer, P. (1987). Finite population correction for replication estimates of variance. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 732-736.

Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica*, 8, 1153-1164.

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

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.

Kim, J.K., Navarro, A. et Fuller, W.A. (2006). Replicate variance estimation after multi-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.

Kim, J.K., et Sitter, R.R. (2003). Efficient variance estimation for two-phase sampling. *Statistica Sinica*, 13, 641-653.

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

Lu, W., et Sitter, R.R. (2006). Risque de divulgation et estimation de la variance. *Recueil de la série des symposiums internationaux de Statistique Canada*, 11-522-XIF.

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

Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.

Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.

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

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

Rust, K.F., et Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

Sampford, M.R. (1967). On sampling without replacement with unequal probability of selection. *Biometrika*, 54, 499-513.

Wolter, K. (2007). *Introduction to Variance Estimation.* 2ième Édition, New York : Springer.

ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Techniques d'enquête, juin 2011
Vol. 37, N° 1, pp. 83-102
Statistique Canada, N° 12-001-X au catalogue

Rentabilité des enquêtes avec échantillonnage en grappes répétées

Stanislav Kolenikov et Gustavo Angeles [1]

Résumé

Nous analysons l'efficacité statistique et économique de diverses enquêtes avec échantillonnage en grappes pour lesquelles la collecte des données est effectuée à deux périodes, ou vagues, consécutives. Dans le cas d'un plan à échantillons indépendants, un échantillon en grappes est tiré de manière indépendante à chacune des deux vagues. Dans le cas d'un plan à panel de grappes, les mêmes grappes sont utilisées aux deux vagues, mais le tirage des échantillons dans les grappes est effectué indépendamment aux deux périodes. Dans un plan à panel d'unités d'observation, les grappes ainsi que les unités d'observation sont retenues d'une vague de collecte des données à l'autre. En supposant que la structure de la population est simple, nous calculons les variances sous le plan ainsi que les coûts des enquêtes réalisées selon ces divers types de plan. Nous considérons d'abord l'estimation de la variation de la moyenne de population entre deux périodes et nous déterminons les répartitions d'échantillon optimales pour les trois plans étudiés. Nous proposons ensuite un cadre de maximisation de l'utilité emprunté à la microéconomie en vue d'illustrer une approche possible pour choisir le plan dans laquelle nous nous efforçons d'optimiser simultanément plusieurs variances. La prise en compte simultanée de plusieurs moyennes et de leurs variances a tendance à faire pencher la préférence du plan à panel d'unités d'observation vers les plans à panel de grappes et à échantillons indépendants plus simples si le mode de collecte de données par panel est trop coûteux. Nous présentons des exemples numériques qui illustrent comment un concepteur d'enquête pourrait choisir le plan efficace sachant les paramètres de population et les coûts de collecte des données.

Mots clés : Étude longitudinale ; échantillon en grappes ; DHS ; NHIS.

1. Introduction

Afin d'analyser la dynamique des phénomènes sociaux, comportementaux ou de santé des populations, les chercheurs et les responsables de l'élaboration des politiques doivent obtenir des renseignements sur les caractéristiques de la population à de multiples périodes. Les enquêtes à plan de sondage complexe sont la source d'information utilisée le plus fréquemment pour les grandes populations, telles que celle d'un pays dans son ensemble. Outre les aspects habituellement pris en considération dans les enquêtes ponctuelles, comme la stratification et la mise en grappes, d'autres éléments peuvent être importants dans les enquêtes dont les données sont recueillies pour deux périodes ou plus. Dans ce genre d'enquête, le coût total et l'erreur d'enquête totale sont influencés par le chevauchement entre les échantillons consécutifs, l'érosion (informative) de l'échantillon, les effets de la durée de la présence dans l'échantillon ou du conditionnement, et d'autres facteurs dynamiques.

En vue d'estimer les variations au moyen d'enquêtes répétées, il est souvent souhaitable que la corrélation temporelle entre les unités observées soient élevée, ce qui peut être réalisé en administrant le questionnaire de l'enquête aux mêmes unités d'échantillonnage et/ou d'observation. Dans les enquêtes longitudinales, une visite est faite aux mêmes unités d'observation (personnes, ménages) pendant plusieurs périodes, éventuellement pendant un nombre indéfini de périodes (Panel Study of Income Dynamics (PSID) des États-Unis, British Household Panel Study (BHPS) et d'autres). Un recueil de renseignements sur les études longitudinales peut être consulté sur le site Web de l'Institute for Social and Economics Research, à l'adresse http://iser.essex.ac.uk/ulsc/keeptrack/index.php. Dans les enquêtes à panel rotatif, les unités d'observation sont recrutées dans l'échantillon pour quelques périodes, puis sortent de l'échantillon et sont de nouveau étudiées à une période ultérieure. La Current Population Survey (CPS) des États-Unis (Binder et Hidiroglou 1988, Eckler 1955, Rao et Graham 1964) et un certain nombre d'enquêtes environnementales (Fuller 1999, McDonald 2003, Scott 1998) sont des exemples d'enquête à panel rotatif. Une autre option consiste à utiliser les mêmes unités primaires d'échantillonnage (UPE) aux différentes vagues, mais à sélectionner indépendamment les unités d'observation (unités secondaires d'échantillonnage, USE). Les enquêtes dont les données sont recueillies de cette façon comprennent les Demographic and Health Surveys (DHS) internationales et la National Health Interview Survey (NHIS) des États-Unis.

Nous nous concentrerons sur les enquêtes dont les données sont recueillies à deux périodes, ou vagues, en utilisant un plan d'échantillonnage en grappes à deux degrés à chaque vague de collecte des données. Nous considérons trois plans de sondage possibles, qui se distinguent par l'importance et la profondeur du chevauchement des unités d'échantillonnage au cours du temps. Le concepteur de l'enquête peut se limiter à réduire tout effet éventuel du

1. Stanislav Kolenikov, Department of Statistics, 146 Middlebush Hall, University of Missouri, Columbia, MO 65211-6100, U.S.A. Courriel : kolenikovs@missouri.edu; Gustavo Angeles, Associate Director of the Center for Evaluation Research, National Institute of Public Health, Mexico, Mexico. Courriel : gangeles@insp.mx.

chevauchement des échantillons, et tirer deux échantillons indépendants à deux périodes distinctes. Nous donnons à ce plan le nom de plan à *échantillons indépendants*. Il pourrait aussi juger utile de recycler les UPE d'une vague à l'autre. S'il estime qu'il est difficile de suivre les USE d'une vague à l'autre, les sous-échantillons dans les grappes peuvent être tirés indépendamment aux deux vagues de collecte des données. Nous donnerons à ce type de plan le nom de plan à *panel de grappes*. Si une très grande précision est essentielle, le plan entièrement longitudinal visera à repérer tous les individus qui ont répondu à la première vague et à leur demander de participer à une deuxième interview. Afin de distinguer ce plan de celui à panel de grappes, nous le nommerons plan à *panel d'unités d'observation*.

Un aspect particulier qui nous semble est important dans la gestion des enquêtes, mais qui n'est pas souvent abordé dans la littérature existante, est le coût de mise en œuvre (Groves 1989). Les modèles de coût classiques, tel celui utilisé pour établir le plan de répartition optimale de l'échantillon de Neyman-Tchuprow (Neyman 1938), peuvent être étendus afin d'y inclure des termes de coût pour la première visite à la grappe et à l'unité d'observation finale, ainsi que pour les visites subséquentes. En ce qui concerne les grappes, le coût est vraisemblablement plus faible à la deuxième visite. Il n'est plus nécessaire de créer de nouvelles cartes ni d'établir des listes. Les mêmes intervieweurs peuvent être choisis pour effectuer les interviews lors des vagues subséquentes de collecte des données. La coopération avec les dirigeants communautaires, parfois importante, par exemple chez certaines sociétés traditionnelles, a déjà été établie. L'effet de la collecte des données par panel au niveau individuel est moins évident. Un ménage interviewé lors des vagues précédentes qui a déménagé doit éventuellement être dépisté dans une région géographique différente, ce qui accroît le coût (moyen) d'interview du panel. La probabilité que des circonstances de ce genre se présentent augmente dans le cas typique d'intervalles plus longs entre les vagues de l'enquête dans les pays en développement : les intervalles entre les vagues de la DHS sont habituellement de cinq à sept ans. En revanche, si un mode d'interview moins coûteux peut être utilisé après la première vague (par exemple, une interview téléphonique au lieu d'une visite sur place), le coût de l'interview du panel diminue.

Dans le présent article, nous tenons compte de considérations statistiques ainsi qu'économiques en vue de choisir le plan de sondage approprié et ses paramètres. Nous supposons que le concepteur de l'enquête souhaite estimer la variation de la moyenne de population entre deux périodes, et/ou les moyennes proprement dites. À la section 2, nous présentons une population sommaire et calculons les variances sous le plan des moyennes, ainsi que et la différence entre ces moyennes, pour les trois plans d'échantillonnage étudiés.

Afin d'intégrer les aspects économiques de la collecte des données, à la section 3, nous présentons un modèle de coût relativement simple pour une enquête avec échantillonnage en grappes répétées. Nous définissons et résolvons les problèmes d'optimisation pour obtenir les tailles d'échantillon optimales pour les trois plans pris en considération. En introduisant dans le modèle les estimations des paramètres statistiques (variances et autocorrélations) et des composantes de coût (coûts au niveau de la grappe et au niveau individuel), le concepteur d'enquête peut comparer les valeurs numériques des variances pour choisir le meilleur plan. À la section 4, nous illustrons cette approche et montrons que chacun des plans pris en considération peut être le meilleur, selon les valeurs des paramètres. Les résultats intuitifs (par exemple, le coût plus élevé de la collecte des données et les autocorrélations plus faibles des caractéristiques observées rendent les modes de collecte des données par panel moins intéressants) sont assortis d'une justification analytique et d'un support quantitatif.

Les sections 2 à 4 traitent de l'efficacité de l'estimation de la différence entre les moyennes seulement, mais des objectifs de collecte des données plus réalistes comprendraient l'estimation de caractéristiques contemporaines et de leurs variances. À cette fin, à la section 5, nous présentons un cadre de maximisation de l'utilité qui décrit le choix de scénarios d'échantillonnage qui s'offre au concepteur de l'enquête. Ce cadre fournit une fonction objectif agrégée qui combine plusieurs critères d'élaboration du plan. Les résultats sont de nouveau ceux attendus : si les modes de collecte des données par panel plus coûteux produisent des tailles d'échantillon plus petites, les estimations des moyennes sont moins efficaces que dans le cas de plans d'échantillonnage plus simples. Le seul moyen de justifier cette perte d'efficacité est d'obtenir une amélioration radicale de l'estimation de la différence, qui ne peut se produire que si les autocorrélations sont plus fortes. Les effets de ce genre sont également illustrés à la section 5. À la section 7, nous présentons nos conclusions. Les preuves sont données en annexe.

2. Variances sous le plan

Posons que la population est constituée de N grappes, ou UPE, aux deux périodes et que chaque grappe est constituée de M individus, ou USE. De cette population est tiré un EAS de $1 < n_t \leq N$ grappes au temps $t = 1, 2$, et un EAS de $1 < m_t \leq M$ individus dans chaque grappe présente dans l'échantillon au temps t. Soit l'indice i désignant les UPE et l'indice j désignant les USE. Donc, la mesure typique sera désignée Y_{tij} dans la population et

y_{tij} dans l'échantillon. Les totaux de population $T[\cdot]$ et leurs estimations $t[\cdot]$ peuvent être calculés comme il suit :

total de grappe :

$$T_{ti\cdot}[Y] = \sum_{j=1}^{M} Y_{tij},\ t_{ti\cdot}[y] = \frac{M}{m}\sum_{j=1}^{M} y_{tij},$$

total de population :

$$T_{t\cdot\cdot}[Y] = \sum_{i=1}^{N} Y_{ti\cdot},\ t_{t\cdot\cdot}[y] = \frac{N}{n}\sum_{i=1}^{N} t_{ti\cdot}[y]. \quad (2.1)$$

Les moyennes pour les unités d'observation sont

$$\bar{Y}_{ti\cdot} = \frac{1}{M}\sum_{i=1}^{M} Y_{tij} = \frac{T_{ti\cdot}[Y]}{T_{ti\cdot}[1]},\ \bar{y}_{ti\cdot} = \frac{1}{m}\sum_{i=1}^{m} y_{tij} = \frac{t_{ti\cdot}[y]}{t_{ti\cdot}[1]},$$

$$\bar{Y}_{t\cdot\cdot} = \frac{T_{t\cdot\cdot}[Y]}{T_{t\cdot\cdot}[1]} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} Y_{tij}}{NM},\ \bar{y}_{t\cdot\cdot} = \frac{t_{t\cdot\cdot}[y]}{t_{t\cdot\cdot}[1]} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{tij}}{nm}. \quad (2.2)$$

La variance de Y et ses composantes intragrappe (indiquée par l'indice *w* pour *within*) et intergrappe (indiquée par *b* pour *between-cluster*) sont

$$S_t^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{tij} - \bar{Y}_{t\cdot\cdot})^2}{N\bar{M} - 1}, \quad (2.3)$$

$$S_{twi}^2 = \frac{\sum_{j=1}^{M}(Y_{tij} - \bar{Y}_{ti\cdot})^2}{M - 1},\ \bar{S}_{tw}^2 = \frac{1}{N}\sum_{i=1}^{N} S_{twi}^2, \quad (2.4)$$

$$S_{tb}^2 = \frac{\sum_{i=1}^{N}(\bar{Y}_{ti\cdot} - \bar{Y}_{t\cdot\cdot})^2}{N - 1}. \quad (2.5)$$

La caractéristique d'intérêt principal est la variation des moyennes,

$$D = \bar{Y}_{2\cdot\cdot} - \bar{Y}_{1\cdot\cdot}, \quad (2.6)$$

estimée par

$$d = \bar{y}_{2\cdot\cdot} - \bar{y}_{1\cdot\cdot}. \quad (2.7)$$

Une propriété de cet estimateur intéressante pour les analystes et les utilisateurs des données est sa cohérence interne : l'estimateur de la différence est égal à la différence des estimateurs. Si les échantillons de périodes consécutives ne se chevauchent que partiellement, l'estimateur composite ou l'estimateur des MCG (Fuller 1999, Hansen, Hurwitz et Madow 1953, Patterson 1950, Rao et Graham 1964, Wolter 2007) sont plus efficaces.

Dans la suite de l'exposé, nous supposons que toutes les procédures d'échantillonnage correspondent à l'échantillonnage aléatoire simple sans remise. Pour la moyenne contemporaine, la variance est donnée par (Cochran 1977, Th. 10.1) :

$$\mathrm{V}[\bar{y}_{t\cdot\cdot}] = \left(1 - \frac{n}{N}\right)\frac{S_{tb}^2}{n} + \left(1 - \frac{m}{M}\right)\frac{\bar{S}_{tw}^2}{nm}. \quad (2.8)$$

Pour la simplicité et la clarté de l'exposé, nous formulerons souvent l'hypothèse de conditions symétriques :

$$S_{1wi}^2 = S_{2wi}^2 = S_{wi}^2,\ \bar{S}_{1w}^2 = \bar{S}_{2w}^2 = \bar{S}_w^2,\ S_{1b}^2 = S_{2b}^2 = S_b^2. \quad (2.9)$$

Les dérivations analytiques sont possibles sans ces hypothèses, mais deviennent extrêmement lourdes. En outre, il n'est pas raisonnable de penser que le concepteur d'enquête pourrait connaître les caractéristiques de la future population. Donc (2.9) devrait être considérée comme un modèle de travail raisonnable.

2.1 Plan à échantillons indépendants

Proposition 1. *Soit le tirage sans remise de n_1 sur N grappes et de m_1 sur M unités d'observation dans les grappes au temps $t = 1$. Soit le tirage sans remise de n_2 sur N grappes et de m_2 sur M unités d'observation dans les grappes au temps $t = 2$, ce tirage étant indépendant de celui effectué au temps $t = 1$. Alors*

$$\mathrm{V}_{\iota}(d) = \left(1 - \frac{n_1}{N}\right)\frac{S_{1b}^2}{n_1} + \left(1 - \frac{n_2}{N}\right)\frac{S_{2b}^2}{n_2} + \left(1 - \frac{m_1}{M}\right)\frac{\bar{S}_{1w}^2}{n_1 m_1} + \left(1 - \frac{m_2}{M}\right)\frac{\bar{S}_{2w}^2}{n_2 m_2}. \quad (2.10)$$

Le résultat découle immédiatement de (2.8) en raison de l'indépendance des deux échantillons. L'indice inférieur ι de la variance désigne le « plan à échantillons indépendants ». Sous les conditions de symétrie de (2.9), si les tailles d'échantillon sont les mêmes au deux périodes, $n_1 = n_2 = n$ et $m_1 = m_2 = m$, alors

$$\mathrm{V}_{e,\iota}[d] = 2\left(1 - \frac{n}{N}\right)\frac{S_b^2}{n} + 2\left(1 - \frac{m}{M}\right)\frac{\bar{S}_w^2}{nm}, \quad (2.11)$$

où l'indice inférieur e, ι signifie « variances égales, plan à échantillons indépendants ».

2.2 Plan à panel de grappes

Proposition 2. *Soit n sur N grappes échantillonnées sans remise à la première période et utilisées aux deux périodes. Soit m sur M unités d'observation échantillonnées sans remise indépendamment aux deux périodes. Alors*

$$\mathrm{V}_c[d] = \left(1 - \frac{n}{N}\right)\frac{S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}} S_{1b} S_{2b}}{n} + \left(1 - \frac{m}{M}\right)\frac{\bar{S}_{1w}^2 + \bar{S}_{2w}^2}{nm},$$

$$\rho^{\mathrm{I}} = \frac{1}{S_{1b} S_{2b}(N-1)}\sum_{i=1}^{N}(\bar{Y}_{1i\cdot} - \bar{Y}_{1\cdot\cdot})(\bar{Y}_{2i\cdot} - \bar{Y}_{2\cdot\cdot}). \quad (2.12)$$

Ici, l'indice inférieur c désigne le plan à panel de grappes (en anglais cluster) » et ρ^{I} est la corrélation intertemporelle, ou autocorrélation, des moyennes de grappe. L'indice supérieur I désigne le premier degré d'échantillonnage. Si ρ^{I} est positive, le plan à panel de grappes est plus efficace que le plan à échantillons indépendants pour des valeurs fixes de n et m. Sous les conditions de symétrie,

$$V_{e,c}[d] = 2\left(1-\frac{n}{N}\right)\frac{S_b^2(1-\rho^{I})}{n} + 2\left(1-\frac{m}{M}\right)\frac{\overline{S}_w^2}{mn}, \quad (2.13)$$

où l'indice inférieur e, c signifie « variances égales, plan à panel de grappes ».

2.3 Plan à panel d'unités d'observation

Proposition 3. *Soit n sur N grappes et m sur M unités d'observation échantillonnées sans remise à la première période et utilisées aux deux périodes. Alors*

$$V_o[d] = \left(1-\frac{n}{N}\right)\frac{S_{1b}^2 + S_{2b}^2 - 2\rho^{I}S_{1b}S_{2b}}{n} + \left(1-\frac{m}{M}\right)\frac{\overline{S}_{1w}^2 + \overline{S}_{2w}^2 - 2\rho^{II}\overline{S}_{1w}\overline{S}_{2w}}{nm},$$

$$\rho^{II} = \frac{1}{\overline{S}_{1w}\overline{S}_{2w}N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{1ij}-\overline{Y}_{1i\cdot})(Y_{2ij}-\overline{Y}_{2i\cdot}). \quad (2.14)$$

L'indice inférieur o désigne le « plan à panel d'unités d'observation ». Sous l'hypothèse des conditions symétriques,

$$V_{e,o}[d] = 2\left(1-\frac{n}{N}\right)\frac{(1-\rho^{I})S_b^2}{n} + 2\left(1-\frac{m}{M}\right)\frac{(1-\rho^{II})\overline{S}_w^2}{nm},$$

$$\rho^{II} = \frac{1}{\overline{S}_w^2 N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{1ij}-\overline{Y}_{1i\cdot})(Y_{2ij}-\overline{Y}_{2i\cdot}) \quad (2.15)$$

où l'indice inférieur correspondant e, o signifie « variances égales, plan à panel d'unités d'observation ».

Ici, ρ^{II} est la corrélation intertemporelle, ou autocorrélation, des observations individuelles dans les grappes. L'indice supérieur II désigne le deuxième degré d'échantillonnage. Si ρ^{II} est positive, le plan à panel d'unités d'observation est plus efficace que le plan à panel de grappes pour des valeurs fixes de n et m.

Comment sont reliées les deux autocorrélations qui figurent dans (2.15) ? Conceptuellement, nous pouvons imaginer n'importe quel nombre de relations possibles entre elles. Introduisons un modèle de superpopulation

$$Y_{tij} = \mu_t + a_{ti} + \varepsilon_{tij}, \quad E_\xi[a_{ti}] = 0, \quad E_\xi[\varepsilon_{tij}] = 0, \quad (2.16)$$

dans lequel les termes a_{ti} et ε_{sij} sont indépendants l'un de l'autre pour tout $s, t = 1, 2$. L'indice inférieur ξ désigne les espérances sous le modèle de superpopulation. Le cas où $\rho^{I} = 0$ et $\rho^{II} = 1$ s'observe quand les variations des moyennes de grappes ont lieu indépendamment entre les grappes ($E_\xi[a_{1i}a_{2i}] = 0$), mais que les individus gardent leur position à l'intérieur de la grappe, $\varepsilon_{1ij} = \varepsilon_{2ij}$. Le cas où $\rho^{I} = 1$ et $\rho^{II} = 0$ a lieu quand les effets aléatoires de grappes sont les mêmes aux deux périodes, $a_{1i} = a_{2i}$, tandis que les effets aléatoires individuels ne sont pas corrélés ($E_\xi[\varepsilon_{1ij}\varepsilon_{2ij}] = 0$). Ni l'une ni l'autre de ces situations n'est entièrement réaliste. Cependant, on peut sans doute s'attendre à ce que la dynamique individuelle soit une source plus importante de variation au cours du temps que la dynamique de grappes, de sorte que les relations $\rho^{II} \geq \rho^{I} \geq 0$ sont les plus plausibles. Nous examinerons dans les exemples numériques des sections 4 et 5 la mesure dans laquelle le choix du meilleur plan est sensible à la relation entre les deux corrélations.

3. Coûts pour des échantillons en grappes répétés

À la présente section, nous analysons la rentabilité des échantillons en grappes lorsque l'on veut estimer la différence entre deux moyennes d'échantillon pour deux périodes différentes.

Une discussion des coûts de l'échantillonnage en grappes est donnée dans Kish (1995, section 8.3B), Thompson (1992, section 12.5), ainsi que Lehtonen et Pahkinen (2004). Un exposé mathématique plus détaillé, avec les formules de variance corrigées pour les populations finies, peut être consulté dans Hansen et coll. (1953, volume II, section 6.11).

3.1 Notation et modèles de coûts

Supposons la structure de coût suivante, qui est une extension de celle de Kish (1995) pour les enquêtes répétées :

- c_1^{I} est le coût au niveau de la grappe au temps $t = 1$ pour les grappes qui sont utilisées *à la première vague seulement* ;
- c_2^{I} est le coût au niveau de la grappe pour une *nouvelle* grappe au temps $t = 2$;
- c_{12}^{I} est le coût au niveau de la grappe pour les grappes dans lesquelles les données sont recueillies aux deux périodes $t = 1$ et $t = 2$ (coût du panel d'UPE) ;
- c_1^{II} est le coût au niveau individuel au temps $t = 1$ pour les individus qui sont observés *à la première vague seulement* ;
- c_2^{II} est le coût au niveau individuel au temps $t = 2$ pour les individus qui sont observés *à la deuxième vague seulement* ;

- c_{12}^{II} est le coût au niveau individuel si l'unité est observée aux deux périodes dans le plan à panel d'unités d'observation (coût du panel d'USE) ;
- C_0 est le budget total alloué au travail sur le terrain aux deux périodes.

Les indices supérieurs en chiffres romains désignent le degré d'échantillonnage. Les indices inférieurs en caractères arabes correspondent à la vague à la quelle l'échantillon est tiré. Les coûts au niveau de la grappe comprennent le coût de l'échantillonnage des grappes, l'obtention des cartes des UPE, la collecte de données sur la collectivité, la formation locale des intervieweurs, *etc*. Les coûts au niveau individuel sont principalement ceux des interviews sur place auprès des unités d'observation finales. Le coût total C_0 est conçu comme le coût variable de l'enquête qui est directement relié au nombre d'unités échantillonnées. Le coût fixe, tel que le coût de la préparation du questionnaire et d'autres coûts au niveau organisationnel, ne fait pas partie de C_0.

3.2 Plan à échantillons indépendants

Pour le plan à échantillons indépendants, la contrainte budgétaire est données par

$$C_0 = c_1^{\mathrm{I}} n_1 + c_1^{\mathrm{II}} n_1 m_1 + c_2^{\mathrm{I}} n_2 + c_2^{\mathrm{II}} n_2 m_2. \qquad (3.1)$$

Les deux premiers termes sont les coûts de la première vague de collecte des données, et les deux derniers, ceux de la deuxième vague.

Proposition 4. *Si les paramètres de configuration de l'enquête sont les mêmes aux deux périodes* :

$$c_1^{\mathrm{I}} = c_2^{\mathrm{I}} = c^{\mathrm{I}},\ c_1^{\mathrm{II}} = c_2^{\mathrm{II}} = c^{\mathrm{II}}, \qquad (3.2)$$

alors les tailles optimales d'échantillon et les variances résultantes sont données par

$$m = \sqrt{\frac{c^{\mathrm{I}}}{c^{\mathrm{II}}} \frac{\overline{S}_w^2}{S_b^2 - \overline{S}_w^2 / M}},$$

$$n = \frac{C_0}{2\{c^{\mathrm{I}} + [c^{\mathrm{I}} c^{\mathrm{II}} \overline{S}_w^2 / (S_b^2 - \overline{S}_w^2 / M)]^{1/2}\}},$$

$$\mathrm{V}_{e,\iota}[d] = \frac{4\left[c^{\mathrm{I}} + \sqrt{c^{\mathrm{I}} c^{\mathrm{II}} \overline{S}_w^2 / (S_b^2 - \overline{S}_w^2 / M)}\right]}{C_0} \times \left[S_b^2 + \left(\sqrt{\frac{c^{\mathrm{II}}}{c^{\mathrm{I}}} \frac{S_b^2 - \overline{S}_w^2 / M}{\overline{S}_w^2}} - \frac{1}{M}\right) \overline{S}_w^2\right] - \frac{2}{N} S_b^2. \qquad (3.3)$$

Dans les équations (3.3), les tailles d'échantillon n et m sont traitées comme des variables continues. En pratique, le nombre entier le plus proche doit être utilisé, une valeur d'au moins 2 étant nécessaire pour estimer la composante de variance appropriée, et les valeurs maximales étant N et M, respectivement.

Le nombre d'unités d'observation échantillonnées dans une grappe dépend uniquement des coûts relatifs au niveau de la grappe et au niveau de l'unité d'observation, $c^{\mathrm{I}} / c^{\mathrm{II}}$, et des variances relatives $S_b^2 / \overline{S}_w^2$, ou, ce qui est équivalent, la corrélation intragrappe. Des coûts d'interview c^{II} plus élevés empêchent le concepteur de l'enquête d'utiliser un plus grand nombre d'unités d'observation : une augmentation de c^{II} entraîne une diminution de m ainsi que de n. Un coût au niveau de la grappe plus élevé donne lieu à une redistribution des unités échantillonnés : quand c^{I} augmente, n diminue tandis que m augmente. Une plus grande variance intragrappe $\overline{S}_w^2$ nécessite le tirage d'un plus grand nombre d'unités d'observation m dans une grappe pour maintenir la précision globale. Une plus grande variance intergrappes S_b^2 nécessite l'échantillonnage d'un plus grand nombre n de grappes. Enfin, le budget total de l'enquête C_0 affecte le nombre de grappes n, mais non la taille m, du sous-échantillon. Par conséquent, la variance de d est inversement proportionnelle à C_0.

Une situation non symétrique peut être traitée comme un sous-produit des conditions de premier ordre dérivées dans la preuve (voir l'annexe). Cependant, aucune solution analytique n'est disponible dans ce cas.

3.3 Plan à panel de grappes

Pour le plan à panel de grappes, la contrainte budgétaire est donnée par

$$C_0 = c_{12}^{\mathrm{I}} n + c_1^{\mathrm{II}} n m_1 + c_2^{\mathrm{II}} n m_2. \qquad (3.4)$$

Le premier terme est le coût au niveau de la grappe associé au plan d'échantillonnage, et les deux autres termes sont les coûts de la collecte des données au niveau individuel aux première et deuxième vagues, respectivement.

Proposition 5. *Pour le plan à panel de grappes, les tailles d'échantillon sont données par*

$$m_1 = 2C_0 \Big/ c_1^{\mathrm{II}} \left(1 + \frac{\overline{S}_{2w}^2}{\kappa \overline{S}_{1w}^2} + \sqrt{D}\right),$$

$$m_2 = \kappa m_1,$$

$$n = \frac{C_0}{c_{12}^{\mathrm{I}} + c_1^{\mathrm{II}} m_1 + c_2^{\mathrm{II}} m_2},$$

$$\kappa = \sqrt{\frac{c_1^{\mathrm{II}}}{c_2^{\mathrm{II}}} \frac{\overline{S}_{2w}^2}{\overline{S}_{1w}^2}}, \qquad (3.5)$$

à condition que

$$D = \left(1 + \frac{\bar{S}_{2w}^2}{\kappa \bar{S}_{1w}^2}\right)^2 + 8\frac{(1-\rho^{\mathrm{I}})S_b^2 C_0}{\bar{S}_{1w}^2 c_1^{\mathrm{II}}} - 4\frac{C_0}{c_1^{\mathrm{II}} M}\left(1 + \frac{\bar{S}_{2w}^2}{\kappa \bar{S}_{1w}^2}\right) \geq 0.$$

La variance de l'estimateur de la différence est obtenue en entrant ces expressions dans (2.13). Sous les hypothèses de conditions symétriques aux deux vagues de l'enquête (2.9) et (3.2),

$$D = 4 - 8\frac{C_0}{Mc^{\mathrm{II}}} + 8\frac{(1-\rho^{\mathrm{I}})S_b^2 C_0}{\bar{S}_w^2 c^{\mathrm{II}}},$$

$$m_1 = m_2 = m = \frac{C_0}{c^{\mathrm{II}} + \sqrt{(c^{\mathrm{II}})^2 - \frac{2c^{\mathrm{II}}C_0}{M} + \frac{2(1-\rho^{\mathrm{I}})S_b^2 C_0 c^{\mathrm{II}}}{\bar{S}_w^2}}},$$

$$n = \frac{C_0}{c_{12}^{\mathrm{I}} + 2c^{\mathrm{II}}m} = \frac{C_0}{c_{12}^{\mathrm{I}} + 2C_0\left/\left[1 + \sqrt{1 - \frac{2c^{\mathrm{II}}C_0}{Mc^{\mathrm{II}}} + \frac{2(1-\rho^{\mathrm{I}})S_b^2 C_0}{\bar{S}_w^2 c^{\mathrm{II}}}}\right]\right.},$$

et $\mathrm{V}_{e,c}[d]$ peut être calculée à partir de (2.13).

Fait intéressant, le nombre d'USE dépend des coûts au niveau des USE c^{II}, mais non des coûts au niveau des UPE c_{12}^{I}. Une augmentation de la corrélation intragrappe, ou de S_b^2, ou une diminution de S_w^2, entraîne comme cela est à prévoir une diminution du nombre optimal d'USE et une augmentation du nombre optimal d'UPE. La dépendance des paramètres du plan à l'égard du budget de l'enquête C_0 n'est pas négligeable. Pour de très petites enquêtes, le nombre d'unités par grappe est proportionnel à C_0, tandis que le nombre de grappes n'est pas affecté par C_0. En effet, si la caractéristique étudiée présente une forte corrélation entre les périodes d'observation, il est préférable d'obtenir des estimations exactes des moyennes de grappe, desquelles découlerait une bonne exactitude de l'estimateur de la différence globale. Autrement dit, le premier terme de (2.13) est relativement petit en raison du coefficient de corrélation positif ρ^{I}, et le deuxième terme est inversement proportionnel à C_0. Pour les grandes enquêtes, $D \propto C_0$, de sorte que le nombre d'unités par grappe ainsi que le nombre de grappes sont proportionnels à $\sqrt{C_0}$. Le premier terme de (2.13) est alors inversement proportionnel à $\sqrt{C_0}$, et le deuxième terme est inversement proportionnel à C_0. Une augmentation du budget de l'enquête affectera tous les termes, quoique dans une mesure différente.

3.4 Plan à panel d'unités d'observation

Pour le plan à panel d'unités d'observation, la contrainte budgétaire est donnée par

$$C_0 = c_{12}^{\mathrm{I}} n + c_{12}^{\mathrm{II}} nm. \qquad (3.6)$$

Le premier terme est le coût au niveau de la grappe et le deuxième terme, le coût des interviews individuelles.

Proposition 6. *Pour le plan à panel d'unités d'observation, les tailles d'échantillon optimales sont données par*

$$m = \sqrt{\frac{c_{12}^{\mathrm{I}}}{c_{12}^{\mathrm{II}}}\frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2}{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}},$$

$$n = \frac{C_0}{c_{12}^{\mathrm{I}} + \sqrt{\frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2 c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}}}{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}}}. \qquad (3.7)$$

La variance sous le plan de l'estimateur de la différence résultant est

$$\mathrm{V}_{e,o}[d] = \frac{2}{C_0}\left\{(1-\rho^{\mathrm{I}})S_b^2 c_{12}^{\mathrm{I}} + (1-\rho^{\mathrm{II}})\bar{S}_w^2\sqrt{c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}}\frac{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2/M}{(1-\rho^{\mathrm{II}})\bar{S}_w^2}} + \left[(1-\rho^{\mathrm{I}})S_b^2 - \frac{1}{M}(1-\rho^{\mathrm{II}})\bar{S}_w^2\right] \times \sqrt{\frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2 c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}}}{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2/M}} + (1-\rho^{\mathrm{II}})\bar{S}_w^2\left(c_{12}^{\mathrm{II}} - \frac{c_{12}^{\mathrm{I}}}{M}\right)\right\} - \frac{2(1-\rho^{\mathrm{I}})S_b^2}{N}. \qquad (3.8)$$

Les expressions des tailles d'échantillon (3.7) ressemblent à celles données à l'équation (3.3) pour le plan à échantillons indépendants, avec le coût de la collecte des données dans une seule vague remplacé par le coût de la collecte de données de panel, et les composantes de la variance S_b^2 et $\bar{S}_w^2$ remplacées par $(1-\rho^{\mathrm{I}})S_b^2$ et $(1-\rho^{\mathrm{II}})\bar{S}_w^2$. La taille de l'échantillon de deuxième degré m dépend uniquement du coût relatif aux niveaux de la grappe et de l'unité d'observation et du ratio des composantes de la variance augmentées des autocorrélations. Donc, comme dans le cas du plan à échantillons indépendants, la relation entre la taille de l'échantillon et la portée de l'enquête n'a lieu que par la voie de $n \propto C_0$, et la variance de la différence diminue de manière inversement proportionnelle à C_0.

En étendant les relations entre les formes fonctionnelles des équations (3.3) et (3.8), nous pouvons établir les relations générales entre les deux plans :

Proposition 7. *Si* $M \gg 1$ *et* $N \gg 1$, *alors* $\mathrm{V}_{e,\iota}[d] \gtrless \mathrm{V}_{e,o}[d]$ *si*

$$2\left(\sqrt{c^{\mathrm{I}}S_b^2} + \sqrt{c^{\mathrm{II}}\bar{S}_w^2}\right)^2 \gtrless \left[\sqrt{c_{12}^{\mathrm{I}}(1-\rho^{\mathrm{I}})S_b^2} + \sqrt{c_{12}^{\mathrm{II}}(1-\rho^{\mathrm{II}})\bar{S}_w^2}\right]^2. \qquad (3.9)$$

Malheureusement, la variance sous le plan à panel de grappes que l'on peut obtenir en combinant les résultats de la proposition 5 avec (2.13) ne permet pas une comparaison aussi claire.

4. Exemple numérique

Afin d'illustrer comment les caractéristiques de la population (variances et autocorrélations) et le processus de collecte des données (coûts) influent sur le choix du plan le plus efficace, nous considérons un exemple numérique. Choisissons la configuration de base avec des conditions symétriques et posons que les valeurs des paramètres sont :

$$\begin{aligned} &N = 10\,000, \quad M = 1\,000, \quad S_b = 100, \\ &S_w = 400, \quad \rho^{\mathrm{I}} = 0{,}1, \quad \rho^{\mathrm{II}} = 0{,}35, \\ &c_1^{\mathrm{II}} = c_2^{\mathrm{II}} = 1, \quad c_{12}^{\mathrm{II}} = 3, \quad c_1^{\mathrm{I}} = c_2^{\mathrm{I}} = 10, \\ &c_{12}^{\mathrm{I}} = 18, \quad C_0 = 20\,000. \end{aligned} \tag{4.1}$$

La structure de coût implique que le coût de la collecte de l'information initiale pour une grappe est le coût de dix interviews, tandis que le coût du suivi de la même grappe est celui de huit interviews seulement. Par ailleurs, obtenir la deuxième interview auprès de la même unité coûte deux fois plus cher qu'obtenir la première.

Étant donné ces paramètres, les tailles d'échantillon et les variances sous le plan sont :

$$\begin{aligned} &m_{e,\iota} = 12, \qquad m_{e,c} = 12, \qquad m_{e,o} = 8, \\ &n_{e,\iota} = 455, \qquad n_{e,c} = 476, \qquad n_{e,o} = 476, \\ &m_{e,\iota} n_{e,\iota} = 5\,460, \quad m_{e,c} n_{e,c} = 5\,712, \quad m_{e,o} n_{e,o} = 3\,808, \\ &\mathrm{V}_{e,\iota}[d] = 99{,}86, \quad \mathrm{V}_{e,c}[d] = 91{,}37, \quad \mathrm{V}_{e,o}[d] = 90{,}20. \end{aligned} \tag{4.2}$$

Le plan à panel d'unités d'observation est 1,2 % plus efficace que le plan à panel de grappes, et 10,7 % plus efficace que le plan à échantillons indépendants. Cependant, il produit une taille d'échantillon appréciablement plus petite, égale aux deux tiers seulement de celle de l'échantillon sous le plan à panel de grappes et à 70 % de la taille de l'échantillon sous le plan à échantillons indépendants.

Naturellement, ces résultats sont fortement liés aux paramètres de la population et à la structure de coût. Pouvons-nous décrire des profils généraux de changement des variances et, donc, de l'efficacité relative des divers plans quand ces paramètres varient ? Dans (4.2), les variances sont dérivées de 13 paramètres donnés dans (4.1), et il est difficile de formuler des énoncés pertinents au sujet de tous ces paramètres simultanément. Plus bas, nous essayerons de donner des coupes transversales bidimensionnelles de cet espace à 13 dimensions et de présenter des illustrations graphiques de la variabilité des variances sous le plan, et donc les domaines d'optimalité de chaque plan, en faisant varier deux paramètres à la fois. Nous fournissons les graphiques des variances pour les plans concernés (habituellement, le plan à panel de grappes est représenté par des courbes en pointillé ; le plan à panel d'unités d'observation, par des courbes en tirets et le plan à échantillons indépendants, par des courbes en tirets et points. Dans la plupart des graphiques, le plan indépendant n'est pas affecté par les variations des paramètres qui définissent les axes des graphiques et est donc ignoré). Nous montrons également l'efficacité relative des divers plans en représentant les domaines de l'espace des paramètres en jaune/gris clair si le plan à échantillons indépendants est le plus efficace, en vert/gris moyen si le plan à panel de grappes est le plus efficace, et en violet/gris foncé si le plan à panel d'unités d'observation est le plus efficace (le code R utilisé pour générer les graphiques est disponible au http://web.missouri.edu/~kolenikovs/SMJ2011/).

La figure 1 illustre comment les variances sous le plan, et donc le plan le plus efficace, varient en fonction des coûts des panels d'UPE et d'USE, c_{12}^{I} et c_{12}^{II}. Manifestement, ces variations n'ont pas d'incidence sur la variance du plan à échantillons indépendants, qui sert de référence. En outre, les variations de c_{12}^{II} n'influencent pas la performance du plan à panel de grappes, qui correspond aux droites d'iso-variance verticales en pointillé dans le graphique de gauche. Les courbes descendantes en tirets sont les courbes d'iso-variance pour le plan à panel d'unités d'observation. Il convient de souligner que le coin inférieur gauche du graphique correspond à la situation de gratuité dans laquelle la deuxième vague de collecte des données ne coûte rien : les coûts de panel sont égaux aux coûts de la période unique, $c_{12}^{\mathrm{I}} = c_1^{\mathrm{I}}$, $c_{12}^{\mathrm{II}} = c_1^{\mathrm{II}}$. Quand les coûts de la collecte des données de panel sont extrêmement élevés (coin supérieur droit du graphique), le plan à échantillons indépendants est le plus efficace. Le point où les trois plans ont la même variance correspond à $c_{12}^{\mathrm{I}} = 22$, $c_{12}^{\mathrm{II}} = 3{,}05$; autrement dit, le coût de la deuxième interview est 2,05 fois plus élevé que celui de la première, et les coûts au niveau de la grappe sont 20 % plus élevés à la deuxième vague qu'à la première. Néanmoins, une autocorrélation positive justifie la réduction de la taille de l'échantillon du plan à panel d'unités d'observation comparativement au plan à échantillons indépendants. Si le coût du panel au niveau de la grappe est plus faible, mais que le coût de la deuxième interview est plus élevé, le plan à panel de grappes est le plus efficace. Si les deuxièmes interviews sont plus coûteuses, le plan le plus efficace est le plan à panel d'unités d'observation. Ce dernier domaine comprend notre cas de référence avec $c_{12}^{\mathrm{I}} = 18$ et $c^{\mathrm{II}} = 3$.

La figure 2 montre les variations des variances sous le plan associées aux variations des autocorrélations ρ^{I}, ρ^{II}. La variance sous le plan à échantillons indépendants n'est pas affectée par cette variation, et celle sous le plan à panel de grappes n'est pas affectée par les variations de ρ^{II}. Le plan à panel d'unités d'observation est plus efficace quand l'autocorrélation des USE est plus forte, $\rho^{\mathrm{II}} > 0{,}34$. Sinon, le plan à panel de grappes donne une variance plus faible.

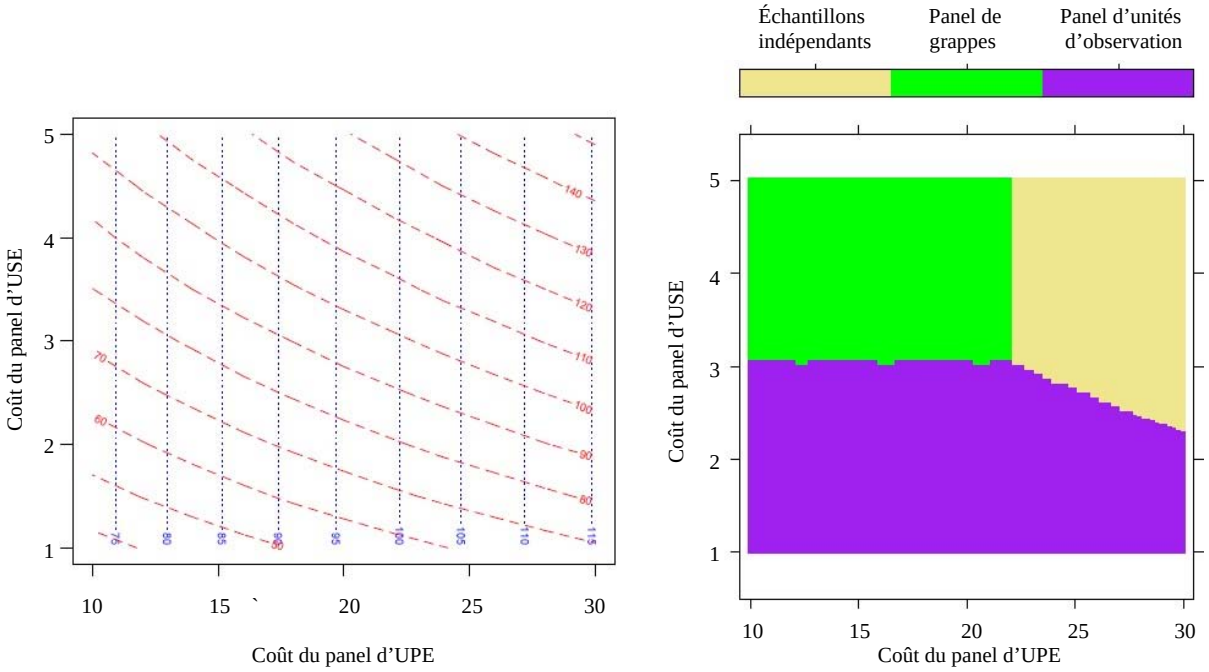

Figure 1 Variances sous le plan en fonction des coûts de collecte des données c_{12}^{I}, c_{12}^{II}. À gauche : lignes de contour de $V_{e,c}[d]$ (pointillé) et $V_{e,o}[d]$ (tiret) ; $V_{e,\iota} = 99,86$; à droite : domaines d'optimalité des trois plans

À la figure 3, nous examinons l'effet du coût au niveau de la grappe et de l'autocorrélation sur le choix du plan. Les combinaisons d'un coût élevé de collecte des données à la deuxième vague et d'une faible autocorrélation des UPE, qui figurent dans le coin supérieur gauche du graphique, font du plan à échantillons indépendants celui qui est le plus intéressant. Autrement, le plan à panel d'unités d'observation est celui qui convient le mieux. Notons que les lignes de contour pour les plans à panel d'unités d'observation sont très proches les unes des autres et que les différences de variance entre les deux plans sont inférieures à 2 % dans l'espace entier des paramètres de ce graphique.

À la figure 4, nous examinons l'effet du coût au niveau de l'unité d'observation et de l'autocorrélation sur le choix du plan. Ni la variance sous le plan à échantillons indépendants ni celle sous le plan à panel de grappes n'est affectée par la variation des paramètres présentés dans ce graphique. La variance sous le plan à échantillons indépendants est égale à 99,86, tandis que celle sous le plan à panel de grappes est de 91,37, de sorte que nous comparons le plan à panel d'unités d'observation à ce dernier seulement. Une forte autocorrélation ($\rho^{II} \geq 0,6$) peut justifier le coût très élevé de la deuxième interview (jusqu'à quatre fois plus élevé que celui de la première), mais dans le coin supérieur gauche du graphique, qui correspond aux faibles autocorrélations et aux coûts de panel élevés, le plan à panel de grappes est celui qui donne les meilleurs résultats.

La figure 5 relie les variances sous le plan aux coûts d'enquête au niveau de la grappe. L'axe horizontal donne le coût à la première période, c_1^I, et l'axe vertical, le coût supplémentaire à la deuxième période, quand les données sont recueillies selon un mode de panel, $c_{12}^I - c_1^I$. Pour le plan à échantillons indépendants, nous ignorons l'axe vertical, car ce paramètre ne figure pas dans le plan. En outre, en raison de (4.1), $c_1^I = c_2^I$. Le plan à panel d'unités d'observation est uniformément meilleur que le plan à panel de grappes pour toutes les combinaisons de paramètres sur ce graphique, quoique l'écart entre les variances ne dépasse pas 2 %. Dans le coin supérieur gauche, le coût additionnel du mode de collecte des données par panel est prohibitif, et le plan à échantillons indépendants est celui qui donne les meilleurs résultats.

La figure 6 illustre la relation entre le plan le plus efficace et le budget total de l'enquête, ainsi que le coût du mode panel de collecte des données au niveau de la grappe. Pour $C_0 > 10\,000$, le plan à panel d'unités d'observation est le meilleur si $c_{12}^I < 22,7$, c'est-à-dire si le coût additionnel du mode de collecte par panel des données au niveau de la grappe n'excède pas 127 % du coût initial de collecte au niveau de la grappe à la première vague. Curieusement, pour certaines configurations isolées des paramètres dans de petites enquêtes, le plan à panel de grappes peut donner de meilleurs résultats que le plan à panel d'unités d'observation qui domine le reste du graphique. La différence entre les variances sous le plan pour le plan à panel de grappes et le plan à panel d'unités d'observation est inférieure à 4 % pour toutes les combinaisons de paramètres dans ce graphique.

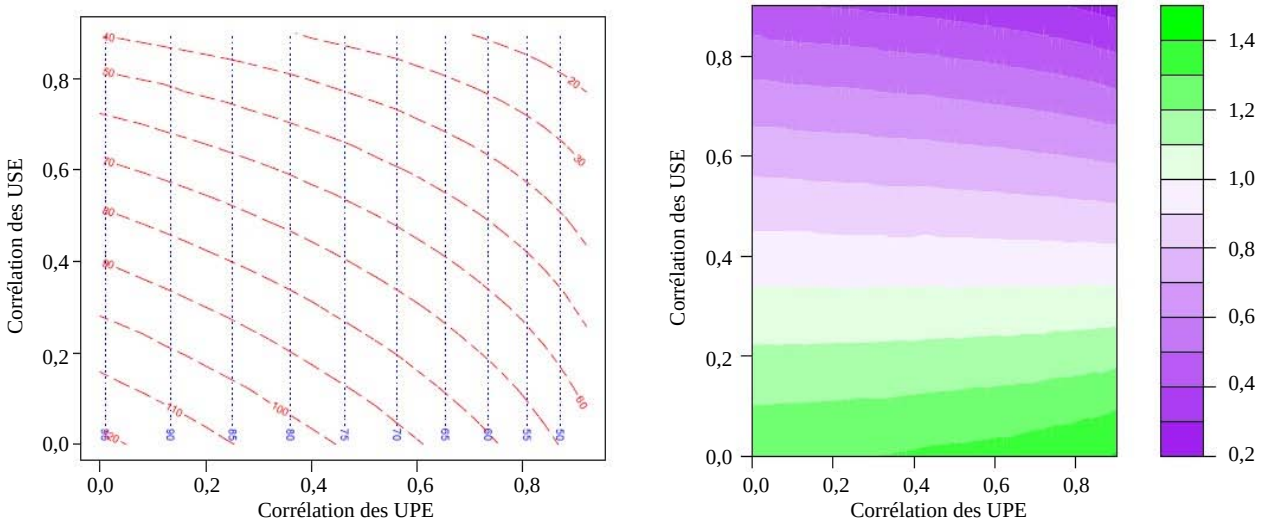

Figure 2 Variances sous le plan en fonction des corrélations de population ρ^{I}, ρ^{II}. À gauche : Lignes de contour de $V_{e,c}[d]$ (pointillé) et de $V_{e,o}[d]$ (tiret) ; $V_{e,\iota} = 99{,}86$; à droite : ratio $V_{e,o}[d] / V_{e,c}[d]$

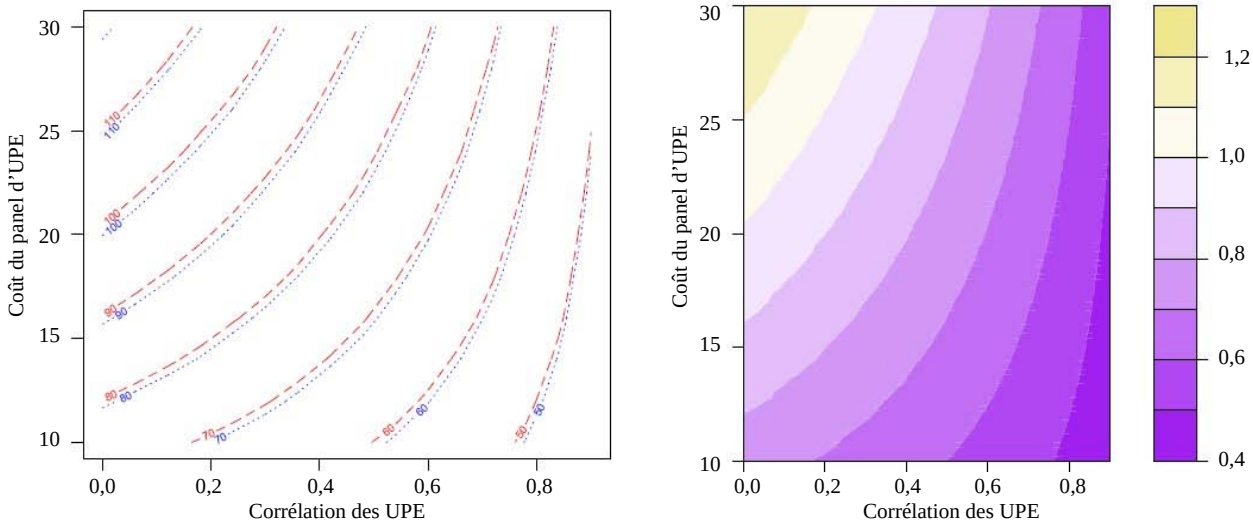

Figure 3 Variances sous le plan en fonction de l'autocorrélation au niveau de la grappe ρ^{I} et du coût c_{12}^{I}. À gauche : lignes de contour de $V_{e,c}[d]$ (pointillé) et de $V_{e,o}[d]$ (tiret) ; $V_{e,\iota} = 99{,}86$; à droite : ratio $V_{e,o}[d] / V_{e,\iota}[d]$

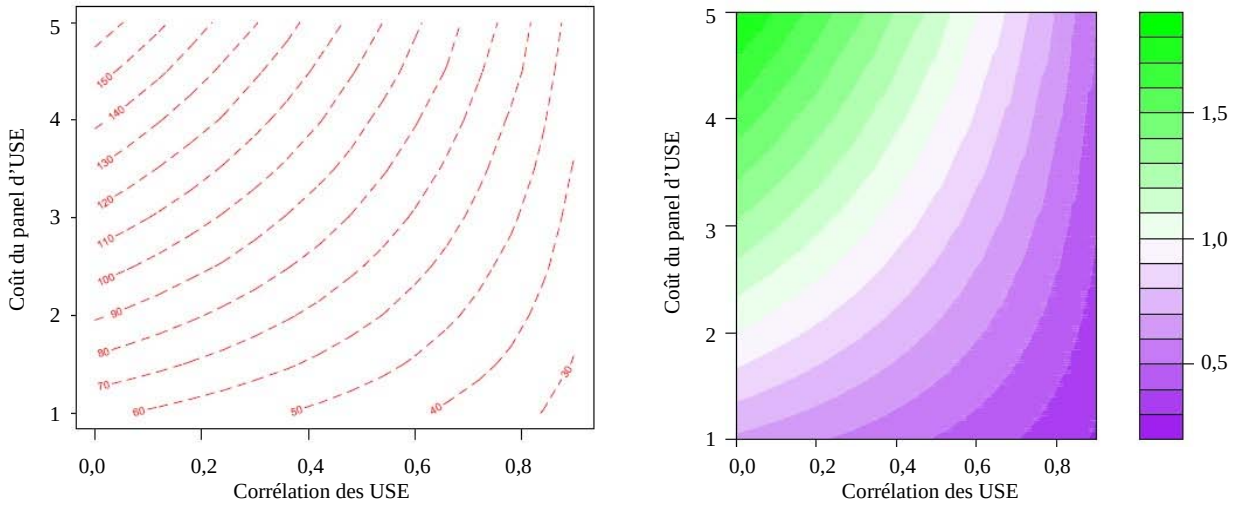

Figure 4 Variances sous le plan en fonction de l'autocorrélation au niveau des unités d'observation ρ^{II} et du coût c_{12}^{II}. À gauche : lignes de contour de $V_{e,o}[d]$ (tiret) ; $V_{e,\iota} = 99{,}86$; $V_{e,c}[d] = 91{,}37$; à droite : ratio $V_{e,o}[d] / V_{e,c}[d]$

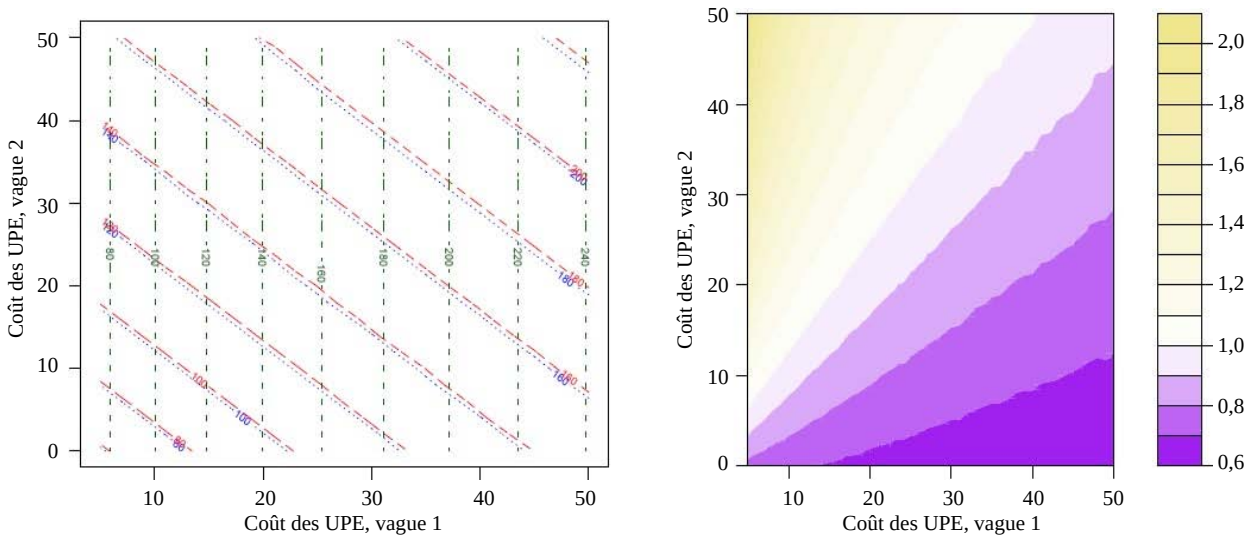

Figure 5 Variances sous le plan en fonction des coûts au niveau de la grappe à la première vague, c_1^{I}, et à la deuxième vague, $c_{12}^{\mathrm{I}} - c_1^{\mathrm{I}}$. À gauche : lignes de contour de $\mathrm{V}_{e,c}[d]$ (pointillé), $\mathrm{V}_{e,o}[d]$ (tiret) et $\mathrm{V}_{e,\iota}[d]$ (tiret et point) ; à droite : ratio $\mathrm{V}_{e,o}[d] / \mathrm{V}_{e,\iota}[d]$

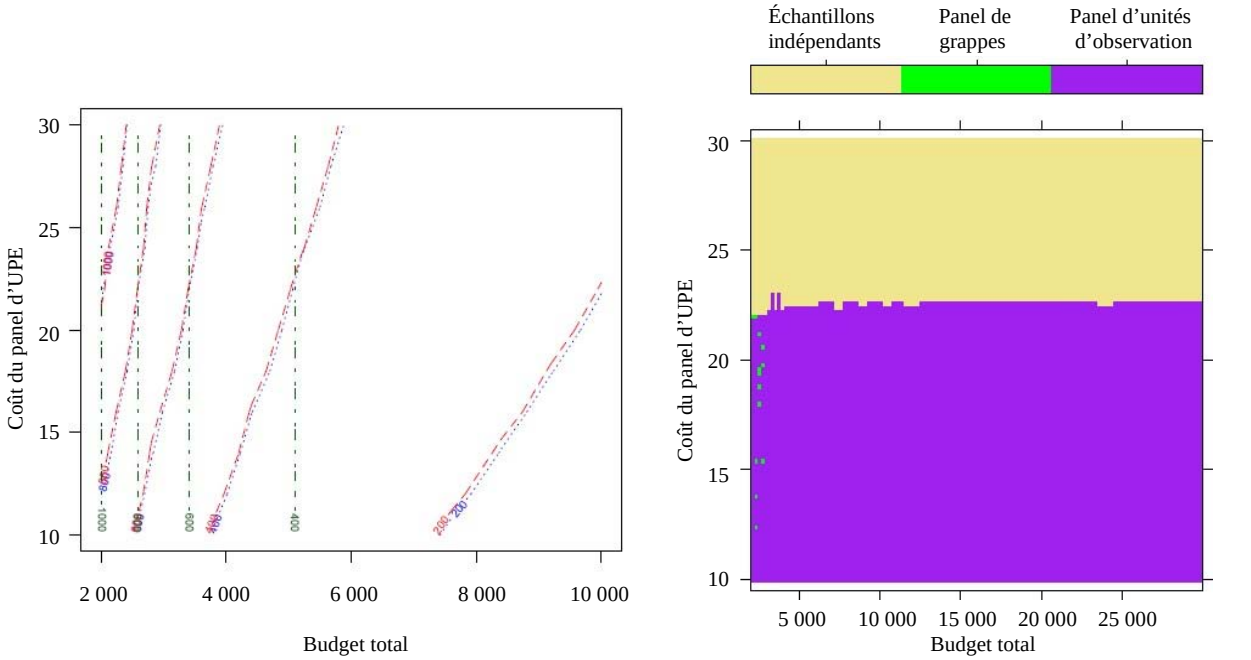

Figure 6 Variances sous le plan en fonction du budget total C_0 et du coût du panel d'UPE c_{12}^{II}. À gauche : lignes de contour de $\mathrm{V}_{e,c}[d]$ (pointillé), $\mathrm{V}_{e,o}[d]$ (tiret) et $\mathrm{V}_{e,\iota}[d]$ (tiret et point) ; à droite : domaines d'optimalité des trois plans

Dans l'ensemble, cet exemple numérique montre que, selon les paramètres de la population et les coûts de la collecte des données, chacun des trois plans peut être le plus efficace. De faibles corrélations et des coûts élevés à la deuxième vague ont tendance à favoriser le plan à échantillons indépendants. Vu que les six paramètres de population initiaux et les cinq paramètres de coût pourraient ne pas être représentatifs d'un grand nombre d'enquêtes répétées, une analyse de sensibilité semblable à celle présentée ici pourrait être nécessaire pour toute enquête particulière qu'un statisticien doit concevoir.

5. Plan de sondage avec critères multiples

Jusqu'à présent, nous avons limité notre analyse à l'estimation de la différence entre les moyennes d'une seule variable à deux vagues de collecte des données. La plupart des enquêtes à grande échelle sont conçues pour recueillir

des données sur plusieurs caractéristiques et de nombreux utilisateurs s'intéressent à la production d'estimations contemporaines. Afin de tenir compte des contraintes d'exactitude associées à ces diverses variables et estimations, le concepteur d'enquête doit avoir à l'esprit plusieurs variances quand il choisit le plan à mettre en œuvre. Il s'agit d'un problème d'optimisation à critères multiples et aucun plan individuel ne sera le meilleur pour tous les problèmes d'estimation possibles. Dans le présent contexte, le plan à panel d'unités d'observation peut donner de bonnes estimations de la variation quand l'autocorrélation des UPE ainsi que celle des USE est élevée, mais risque de produire une petite taille d'échantillon si le coût du suivi des UPE ainsi que des USE est élevé. Des estimations d'une grande précision pour n'importe quelle période pourraient être obtenues en passant au plan à panel de grappes, voire même au plan à échantillons indépendants.

Dans cette situation, il est possible de comparer les différents plans en utilisant l'argument microéconomique classique de maximisation de l'utilité sous les contraintes budgétaires (Mas-Colell, Whinston et Green 1995). Dans le contexte des plans de sondage, l'utilité pour le concepteur d'enquête augmente avec la précision des estimations d'après les données de l'enquête, ou ce qui est équivalent, diminue quand les variances sous le plan augmentent. Une forme fonctionnelle simple est donnée par la fonction d'utilité de Cobb-Douglas :

$$U(\text{plan}) = \mathrm{V}_{\text{plan}}^{-\alpha_1}[\bar{y}_{1\cdot\cdot}]\,\mathrm{V}_{\text{plan}}^{-\alpha_2}[\bar{y}_{2\cdot\cdot}]\,\mathrm{V}_{\text{plan}}^{-\alpha_3}[d]. \qquad (5.1)$$

Ici, α_1, α_2 et α_3 sont des constantes positives qui décrivent les poids relatifs des trois variances sous le plan dans le processus décisionnel. Dans (5.2), les variances $\mathrm{V}[\bar{y}_1]$ et $\mathrm{V}[\bar{y}_2]$ sont les variances des moyennes dans les enquêtes avec échantillonnage en grappes données par (2.8). La variance de l'estimateur de la différence correspond à (2.10), (2.12) ou (2.14), selon le plan choisi. Le problème que doit alors résoudre le concepteur d'enquête consiste à maximiser (5.1) sous les contraintes budgétaires particulières au plan (3.1), (3.4) ou (3.6). La maximisation est effectuée sur les paramètres du plan (mode de collecte des données, nombre de grappes à chaque période, nombre d'unités d'observation à chaque période), sachant les caractéristiques de la population (variances et autocorrélations) et le processus de collecte des données (coûts).

Supposons que la précision des trois estimations $\bar{y}_1$, $\bar{y}_2$ et d ont chacune la même importance pour le décideur, de sorte que $\alpha_1 = \alpha_2 = \alpha_3$. Afin d'obtenir une fonction objectif mesurée dans les unités de variance et sur la même échelle que les variances, il est commode de définir une variance multicritères

$$\mathrm{V}_{\text{plan}} = (\mathrm{V}_{\text{plan}}[\bar{y}_{1\cdot\cdot}]\,\mathrm{V}_{\text{plan}}[\bar{y}_{2\cdot\cdot}]\,\mathrm{V}_{\text{plan}}[d])^{1/3}, \qquad (5.2)$$

et d'exprimer le problème d'optimisation comme étant la maximisation de cette expression.

La caractérisation analytique du plan qui optimise (5.2) devient relativement lourde. À la place, nous utilisons l'exemple numérique de la section précédente pour démontrer comment la prise en compte des autres objectifs de conception de l'enquête affecte le choix du plan. Nous devrions nous attendre à ce que, dans le cas des plans pour lesquels le coût du suivi est plus élevé ($c_{12}^{\mathrm{I}} \geq c_1^{\mathrm{I}} + c_2^{\mathrm{I}}$, $c_{12}^{\mathrm{II}} \geq c_1^{\mathrm{II}} + c_2^{\mathrm{II}}$), les plans plus simples soient sélectionnés plus fréquemment : le plan à panel de grappes peut être préféré au plan à panel d'unités d'observation, et le plan à échantillons indépendants peut être préféré au plan à panel de grappes. Pour les conditions de base (4.1), nous avons

$$\mathrm{V}_{e,\iota}[\bar{y}] = 49{,}93,\ \mathrm{V}_{e,c}[\bar{y}] = 47{,}68,\ \mathrm{V}_{e,o}[\bar{y}] = 61{,}69,$$
$$\mathrm{V}_{e,\iota} = 62{,}91, \quad \mathrm{V}_{e,c} = 59{,}23, \quad \mathrm{V}_{e,o} = 70{,}02,$$

où les indices temporels de $y_{t\cdot\cdot}$ sont omis. Le plan à panel d'unités d'observation est assez inefficace pour l'estimation de la moyenne propre à la période, car un plus petit nombre d'unités sont échantillonnées. En revanche, le plan à panel de grappes est le plus efficace, suivi de près par le plan à échantillons indépendants.

Les figures 7 à 12 sont la réplique des figures 1 à 6, respectivement. Puisque le meilleur plan en ce qui concerne V est maintenant le plan à panel de grappes, la plupart de ces graphiques indiquent la préférence pour ce plan. La figure 7 montre que, si l'on tient compte des variances des moyennes contemporaines, les plans à échantillons indépendants et à panel de grappes plus simple sont préférés pour une fraction plus importante de configuration des paramètres, et occupent une part plus grande du graphique dans la figure 1. Le point où les trois plans sont équivalents correspond à $c_{12}^{\mathrm{I}} = 20{,}6$, $c_{12}^{\mathrm{II}} = 2{,}27$, qui est plus proche de l'origine qu'à la figure 1, dans laquelle seule la variance de la différence était prise en compte.

La figure 8 révèle que le plan à panel d'unités d'observation n'est justifié que si les autocorrélations sont toutes deux supérieures à 0,6 (pour les valeurs données des variances de population et des coûts). Rappelons qu'à la figure 2, le plan à panel d'unités d'observation avait la préférence quand $\rho^{\mathrm{II}} > 0{,}34$, le choix étant peu dépendant de ρ^{I}.

La figure 9 montre comment les corrélations et les coûts au niveau des UPE influencent le choix du plan. Le plan à panel d'unités d'observation est moins efficace que celui à panel de grappes pour toutes les combinaisons de paramètres dans ce graphique. Donc, le choix se fait entre le plan à échantillons indépendants et le plan à panel de grappes. Évidemment, si la collecte des données en mode de panel est coûteuse, le plan à échantillons indépendants sera préféré au plan à panel de grappes. Fait intéressant, la préférence pour un plan particulier n'est pas monotone en

ρ_{12}^{I}. Pour des valeurs de $\rho_{12}^{I} > 0,7$, la composante $V[d]$ dans (5.2) produit des plans contenant si peu de grappes que cela affecte suffisamment fort $V[\bar{y}]$ pour nuire à la fonction objectif complète. À cette valeur de l'autocorrélation du panel, le coût de panel maximal auquel le plan à panel de grappes demeure le plus efficace est $c_{12}^{I} = 24,4$, ce qui signifie que le coût au niveau de la grappe est 44 % plus élevé à la deuxième vague qu'à la première.

La figure 10 montre qu'une plus forte autocorrélation des mesures des USE peut justifier une augmentation modérée du coût associé à la collecte des données. Le coût plus élevé pour lequel le plan à panel d'unités d'observation demeure le plus efficace est $c_{12}^{II} = 2,75$, avec $\rho^{II} = 0,78$; cela signifie que le coût de la deuxième interview peut être 75 % plus élevé que celui de la première.

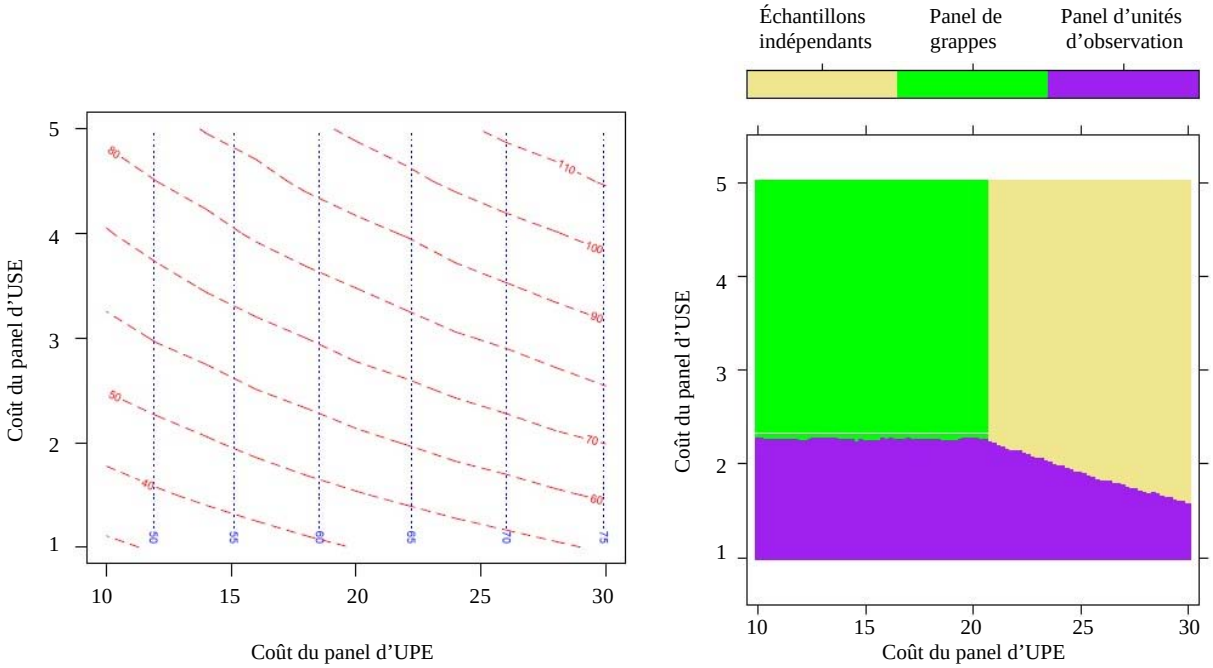

Figure 7 Variances sous le plan en fonction des coûts de collecte des données c_{12}^{I}, c_{12}^{II}. À gauche : lignes de contour de $V_{e,c}$ (pointillé) et de $V_{e,o}$ (tiret) ; $V_{e,\iota} = 62,91$; à droite : domaines d'optimalité des trois plans

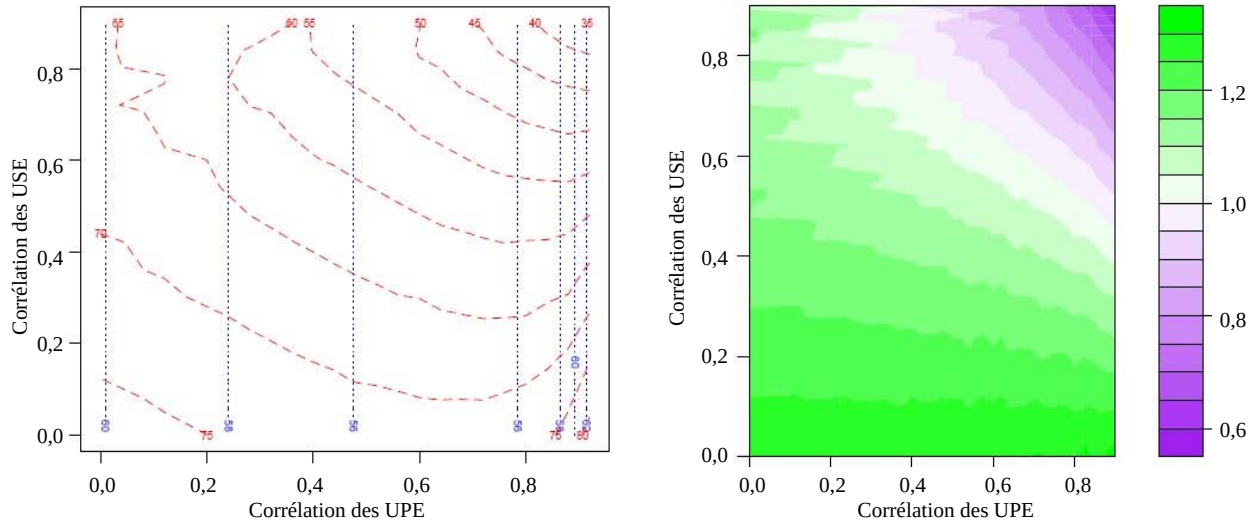

Figure 8 Variances sous le plan en fonction des autocorrélations ρ^{I}, ρ^{II}. À gauche : lignes de contour de $V_{e,c}$ (pointillé) et de $V_{e,o}$ (tiret) ; $V_{e,\iota} = 62,91$; à droite : ratio $V_{e,o}/V_{e,c}$

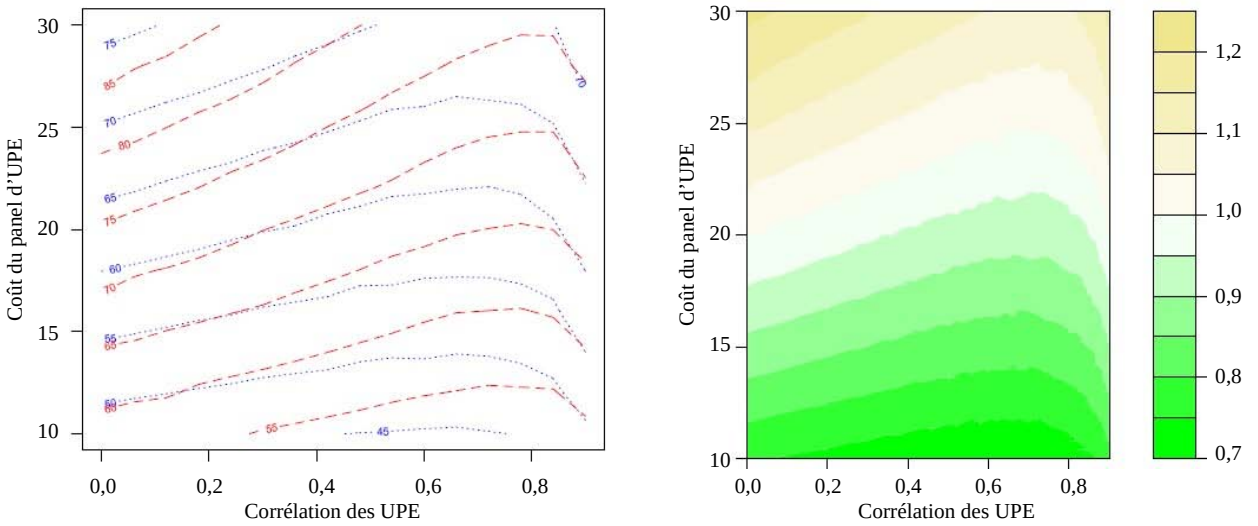

Figure 9 Variances sous le plan en fonction de l'autocorrélation au niveau de la grappe ρ^{I} et du coût c_{12}^{I}. à gauche : lignes de contour de $V_{e,c}$ (pointillé) et de $V_{e,o}$ (tiret) ; $V_{e,\iota} = 62,91$; à droite : ratio $V_{e,c}/V_{e,\iota}$

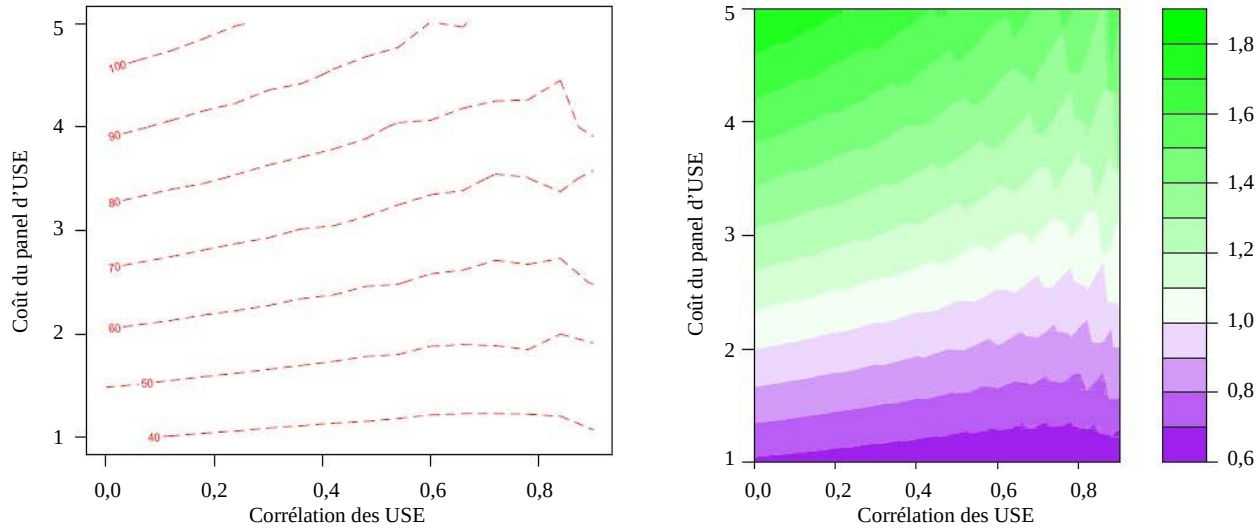

Figure 10 Variances sous le plan en fonction de l'autocorrélation au niveau des unités d'observations ρ^{II} et du coût c_{12}^{II}. À gauche : lignes de contour de $V_{e,o}$ (tiret) ; $V_{e,\iota} = 62,91$; $V_{e,c} = 59,23$; à droite : ratio $V_{e,o}/V_{e,c}$

La figure 11 est l'équivalent de la figure 5. La partie de gauche montre que le plan à panel d'unités d'observation est moins efficace que le plan à panel de grappes. La partie de droite montre que, si le coût au niveau de la grappe à la deuxième vague dépasse de plus de 15 unités ce coût à la première vague, le plan à échantillons indépendants est plus efficace que le plan à panel de grappes.

Enfin, la figure 12 donne les variances en fonction du budget total de l'enquête et du coût de la collecte de données par panel. Le graphique ne révèle qu'une très légère dépendance à l'égard de C_0, et le plan à échantillons indépendants est celui qui a la préférence si le mode de collecte par panel est trop coûteux, à savoir si le coût au niveau de la grappe à la deuxième vague dépasse de plus de 107 % le coût à la première vague.

Comme nous l'avions conjecturé au début de la présente section, l'intégration des variances des moyennes contemporaines dans la fonction objectif d'optimisation du plan déplace les préférences du concepteur d'enquête vers des plans plus simples permettant d'échantillonner un plus grand nombre d'unités d'observation finales. Le plan à panel d'unités d'observation ne se justifie maintenant que si les autocorrélations des UPE ainsi que des USE sont élevées, et que les coûts de collecte par panel sont raisonnablement faibles. En outre, le plan à panel de grappes n'est généralement justifié qu'en cas de réduction du coût au niveau de la grappe à la deuxième vague de l'enquête.

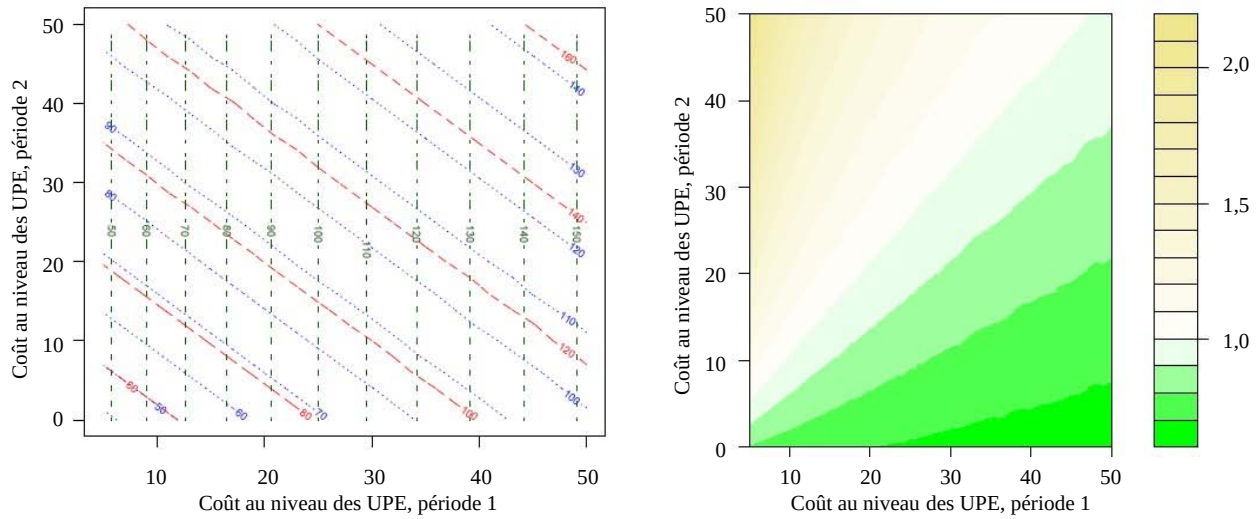

Figure 11 Variances sous le plan en fonction des coûts de collecte des données $c_1^{\mathrm{I}}, c_{12}^{\mathrm{I}}$. À gauche : lignes de contour de $\mathrm{V}_{e,c}$ (pointillé), $\mathrm{V}_{e,o}$ (tiret) et $\mathrm{V}_{e,\iota}$ (tiret et point) ; à droite : ratio $\mathrm{V}_{e,c}/\mathrm{V}_{e,\iota}$

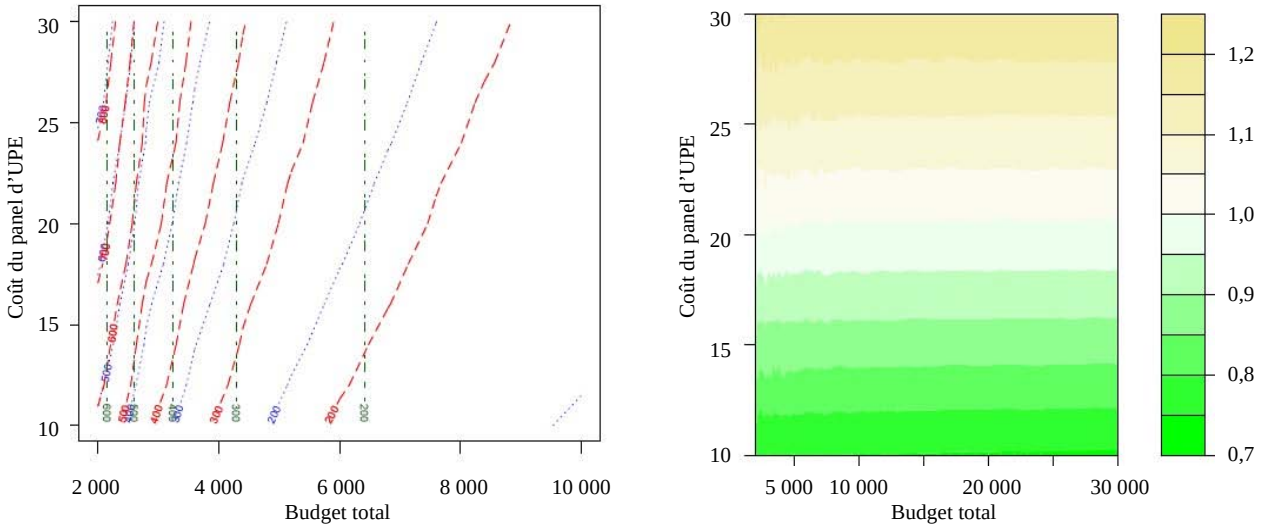

Figure 12 Variances sous le plan en fonction du budget total C_0 et du coût du panel d'UPE c_{12}^{II}. À gauche : lignes de contour de $\mathrm{V}_{e,c}$ (pointillé), $\mathrm{V}_{e,o}$ (tiret) et $\mathrm{V}_{e,\iota}$ (tiret et point) ; à droite : domaines d'optimalité des trois plans

6. Extension à des vagues multiples

Si l'enquête à concevoir doit comporter plus de deux vagues de collecte des données, le concepteur de l'enquête pourrait étendre le cadre du problème de maximisation de l'utilité (5.1) en tenant compte des considérations qui suivent.

1. Un plus grand nombre de cibles d'inférence. Les variances possibles que le concepteur d'enquête doit éventuellement prendre en considération peuvent maintenant inclure les variances contemporaines $\mathrm{V}[\bar{y}_1], \mathrm{V}[\bar{y}_2], \ldots, \mathrm{V}[\bar{y}_T]$; les différences consécutives $\mathrm{V}[\bar{y}_2-\bar{y}_1], \ldots, \mathrm{V}[\bar{y}_T-\bar{y}_{T-1}]$ ou les estimateurs composite/MCG de la variation entre deux périodes adjacentes ; d'autres contrastes $\mathrm{V}[\sum_t c_t \bar{y}_t]$, $\sum c_t = 0$; la variance des taux de croissance linéaires provenant de la régression de $\bar{y}_t$ sur t, estimée par les MCO ou les MCG, *etc*.
2. La possibilité d'une actualisation. En économie, on spécifie d'habitude les contraintes budgétaires faisant référence à l'avenir sous la forme $\sum_t x_t \delta^t$ où x_t est le montant dépensé au temps t, et $\delta < 1$ est le facteur d'actualisation associé aux taux d'intérêt. L'actualisation peut également être pertinente pour la fonction d'utilité, et les variances sous le plan qui se situe plus loin dans l'avenir peuvent avoir moins de poids dans le problème d'optimisation.
3. Des formes fonctionnelles inconnues pour les processus chronologiques associés à la variable d'intérêt. Le concepteur d'enquête doit avoir une bonne idée

de la structure de covariance de la série chronologique d'observations individuelles ainsi que de moyennes de grappes. Il est probable que les résultats seront sensibles au choix d'un modèle particulier. Dans la présente analyse, le problème est plus simple, car il suffit d'avoir un seul paramètre de corrélation pour chaque niveau. Le concepteur d'enquête pourrait devoir introduire plus de paramètres dans le modèle et, conséquemment, étudier la sensibilité du plan choisi à ces divers paramètres.

La complexité du problème, tel qu'il est exposé ci-dessus, peut devenir très rapidement difficile à maîtriser. Nous nous abstenons donc de présenter un traitement plus détaillé dans le présent article.

7. Discussion

Le présent article décrit l'analyse de diverses options de mise en œuvre d'enquêtes à échantillonnage en grappes répétées. Nous fournissons une expression analytique pour les variances sous le plan de l'estimateur par simple différence pour trois plans de sondage fréquemment utilisés (les plans à échantillons indépendants, à panel de grappes et à panel d'unités d'observation). Nous dérivons également les tailles d'échantillon optimales pour l'estimation de la différence entre deux vagues de collecte des données.

Le concepteur d'enquête qui sait que la caractéristique d'intérêt persistera dans une certaine mesure au cours du temps choisira vraisemblablement l'un des plans à collecte par panel, à condition que les coûts des nouvelles visites aux grappes et/ou aux unités d'observation ne soient pas prohibitifs. La comparaison analytique entre les plans à échantillons indépendants et à panel d'unités d'observation est possible et est donnée par la proposition 7. Il convient de préciser que la variance sous le plan de la différence est d'ordre $< O(C_0^{-1})$ pour le plan à échantillons indépendants ainsi que pour le plan à panel d'unités d'observation, et qu'elle est d'ordre $O(C_0^{-1/2})$ pour le plan à panel de grappes, où C_0 est le budget total de l'enquête. Le plan à panel de grappes n'est donc viable que pour les petites enquêtes, tandis que le plan à échantillons indépendants ou à panel d'unités d'observation sera vraisemblablement choisi pour les enquêtes à grande échelle.

La structure de coût considérée à la section 3 est assez simpliste. Par exemple, les coûts d'échantillonnage de deuxième degré à la deuxième période pourraient différer selon que les unités sont échantillonnées à partir de nouvelles grappes ou à partir de grappes réutilisées. En outre, le coût peut dépendre de la taille de grappe M_i, car plus de temps et plus de ressources pourraient être nécessaires pour obtenir les cartes et recueillir les données au niveau de la grappe dans le cas des grandes grappes. Au départ, nous avions l'intention de considérer des situations dans lesquelles le coût du panel d'USE était plus de deux fois plus élevé que le coût des interviews individuelles. Cependant, comme l'a suggéré l'un des examinateurs, ce coût pourrait être plus faible si les interviews de suivi sont effectuées selon un mode moins coûteux, tel qu'une interview téléphonique ou une enquête par la poste avec questionnaire à remplir soi-même au lieu d'une interview sur place. Le cas échéant, le plan à panel d'unités d'observation semble être le plus rentable des trois.

La structure de population est également sursimplifiée. Nous avons supposé que les grappes sont de taille équilibrée constante. Aucune unité de la population n'en sort et aucune nouvelle unité n'y apparaît. Dans de nombreuses situations pratiques, ces hypothèses sont assez contraignantes. Si la population évolue entre deux vagues de collecte des données, le concepteur d'enquête souhaitera inclure de nouvelles grappes à la deuxième vague, en utilisant les algorithmes de Ernst (1999). Les nouvelles grappes sont placées dans une strate distincte, d'où est tiré un échantillon en grappes. Dans le cas de la NHIS, cette approche est mise en œuvre en se servant comme base de sondage d'une liste de « permis de bâtir ». En outre, les effets de mesure dynamiques, tels que le conditionnement et le temps passé dans l'échantillon, donnent lieu à un biais de rotation de l'échantillon, de sorte qu'il pourrait être avantageux de prévoir au moins une certaine rotation des UPE.

En particulier, pour les études réalisées dans le cadre des DHS, le premier argument (couverture) est vraisemblablement plus important que le second (durée de la présence dans l'échantillon), étant donné l'intervalle de temps important entre les cycles de l'enquête (environ cinq ans). La non-réponse et la perte de couverture pourraient sans doute être ajoutées toutes deux au cadre courant comme sources de biais, entraînant ainsi l'optimisation de la moyenne quadratique de l'erreur totale d'enquête plutôt que de la variance sous le plan. Toutefois, la formulation de modèles convaincants de ce genre de biais pourrait être difficile.

Un autre problème qui se poserait si les grappes étaient de tailles différentes est celui de la plus grande gamme de plans applicables. Dans le présent article, nous avons supposé que l'on procédait à un échantillonnage aléatoire simple sans remise aux deux degrés d'échantillonnage. D'autres plans, tels que l'échantillonnage avec probabilités proportionnelles à la taille (PPT), peuvent aussi être utilisés. Pour d'autres plans que l'échantillonnage aléatoire simple, il faudrait se servir de l'estimateur d'Horvitz-Thompson et de l'estimateur de sa variance (Särndal, Swensson et Wretman 1992, Thompson 1997). Les calculs analytiques deviennent compliqués, mais des démonstrations numériques pratiques

semblables à celles de nos sections 4 et 5 peuvent encore être effectuées. Si la taille des grappes varie au cours du temps, l'obtention du plan optimal devient une cible mouvante et les plans qui sont optimaux pour les « anciennes » mesures de taille perdront de leur efficacité avec les « nouvelles ».

Dans des ébauches antérieures du présent article, nous avons analysé des plans intermédiaires dans lesquels une fraction non négligeable des unités est retenue et d'autres unités sont échantillonnés indépendamment. Le problème peut alors être considéré comme une minimisation de la variance sous des contraintes d'inégalité appliquées au degré de chevauchement $0 \leq \pi^{\mathrm{I}} \leq 1$, $0 \leq \pi^{\mathrm{II}} \leq 1$. La théorie générale de l'optimisation sous contraintes non linéaires fait en sorte qu'à condition que la variance de la variation de la moyenne de population D soit monotone en π^{I} et π^{II}, la solution optimale sera atteinte à l'un des sommets de l'espace des paramètres. Cela justifie notre intérêt pour les trois plans examinés dans le présent article. Ces plans correspondent aux sommets de l'espace des paramètres de coordonnées $(0, 0)$, $(1, 0)$ et $(1, 1)$ pour les plans à échantillons indépendants, à panel de grappes et à panel d'unités d'observation, respectivement. Le point $(0, 1)$ correspond à un plan impossible avec chevauchement complet des unités individuelles sans aucun chevauchement des grappes. Des calculs fastidieux montrent qu'il est possible de satisfaire les conditions de premier ordre dans certains cas intermédiaires également, mais ils correspondent à des maxima locaux de la variance. Bien que ces résultats puissent aussi être intéressants (en ce sens qu'ils fournissent une borne supérieure pour les variances sous le plan), nous ne les avons pas pris en considération dans le présent article. Dans les cas plus compliqués d'optimisation à critères multiples de la section 5, la monotonicité n'est pas nécessairement vérifiée et d'autres plans que les trois cas extrêmes considérés dans l'article peuvent produire les valeurs optimales de la fonction objectif (5.2).

Les conditions d'égalité des variances (2.9) peuvent être relâchées au prix de la production d'expressions considérablement plus compliquées. Si les tailles d'échantillon sont maintenues fixes entre les deux vagues de l'enquête, les changements qui suivent peuvent être nécessaires dans toutes les formules pertinentes. Dans les expressions qui ne font pas intervenir les autocorrélations,

$$2S_b^2 \mapsto S_{1b}^2 + S_{2b}^2, \quad 2S_w^2 \mapsto \overline{S}_{1w}^2 + \overline{S}_{2w}^2, \qquad (7.1)$$

tandis que dans les expressions dans lesquelles interviennent les autocorrélations,

$$\begin{aligned} 2(1-\rho^{\mathrm{I}})S_b^2 &\mapsto S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}} S_{1b} S_{2b}, \\ 2S_w^2(1-\rho^{\mathrm{II}}) &\mapsto \overline{S}_{1w}^2 + \overline{S}_{2w}^2 - 2\rho^{\mathrm{I}} \overline{S}_{1w} \overline{S}_{2w}. \end{aligned} \qquad (7.2)$$

Qualitativement, les résultats seront les mêmes.

Le cadre multicritères de la section 5 permet d'attribuer des poids d'importance différents aux diverses variances d'intérêt. Des valeurs relativement grandes de α_1, α_2 correspondent à l'importance plus grande des moyennes contemporaines, tandis que de plus grandes valeurs de α_3 correspondent à l'importance plus grande de l'estimation de la variation. Le problème original d'optimisation du plan pour $\mathrm{V}[d]$ peut être considéré dans le contexte de (5.1) en posant que $\alpha_1 = \alpha_2 = 0$, $\alpha_3 = 1$. Ce cadre peut aussi être élargi afin d'inclure les plans destinés à mesurer plusieurs variables. Un défi supplémentaire associé à ce genre de configuration est que les autocorrélations peuvent différer d'une variable à l'autre. Certaines caractéristiques individuelles sont constantes au cours du temps (race, sexe), tandis que d'autres évoluent lentement (logement, dépenses, préférences politiques), et d'autres encore évoluent plus rapidement (revenus ou comportements).

Le présent article traite de trois plans d'échantillonnage et d'un estimateur particulier de la variation, à savoir la différence entre les estimations de la moyenne à deux périodes dans le temps. D'autres options existent en ce qui concerne tant les plans que les estimateurs. Par exemple, dans les plans à échantillon rotatif, une fraction des unités de la première vague est retenue et certaines nouvelles unités sont recrutées. Pour ce genre de plans, l'estimation composite (Hansen et coll. 1953, Patterson 1950, Rao et Graham 1964, Wolter 2007) dans laquelle sont pondérées différemment les contributions des unités indépendantes (celles retirées de l'échantillon après la première vague et celles nouvellement recrutées pour la deuxième vague) et les contributions des unités du panel (utilisées aux deux vagues) produirait des estimations plus efficaces. En général, ce sont des considérations non liées à l'échantillonnage, telles qu'une diminution du fardeau de réponse et la détérioration de la représentativité de l'échantillon en raison de l'évolution de la population, qui motive l'utilisation de ce genre de plans. Ces considérations peuvent être prises en compte dans le modèle de coût (par exemple, un plus grand nombre de rappels nécessaires pour convaincre une unité de répondre) ou dans le modèle de l'erreur totale d'enquête (en introduisant le biais de non-réponse ou de sous-dénombrement et en considérant l'erreur quadratique moyenne d'une estimation plutôt que sa variance sous le plan).

Remerciements

Les auteurs remercient Chris Skinner et John Eltinge de leurs discussions utiles, William Kalsbeek, de ses suggestions aux premières étapes de la rédaction de l'article, ainsi que le rédacteur associé et deux examinateurs de leurs commentaires. Nash Herndon et Oksana Loginova ont amélioré la rédaction. Un appui financier partiel a été fourni par

la U.S. Agency for International Development dans le cadre du Measure Evaluation Project du Carolina Population Center de l'Université de la Caroline du Nord à Chapel Hill, aux termes de l'entente de coopération GPO-A-00-03-00003-00. Les auteurs remercient aussi de leurs commentaires constructifs les participants aux Joint Statistical Meetings (2005) et au XXIII[e] Symposium international sur les questions de méthodologie de Statistique Canada (2007).

Annexe

Dans les preuves qui suivent, les espérances, les variances et les covariances sont calculées par rapport au plan de sondage correspondant. Le premier degré de sélection sera désigné par l'indice supérieur I. Le deuxième degré de sélection sera désigné par l'indice supérieur II.

Preuve de la proposition 2. Désignons l'échantillon d'UPE par $\mathcal{S}^{\mathrm{I}}$, l'échantillon d'USE à la première période par $\mathcal{S}_{i1}^{\mathrm{II}}$, et l'échantillon d'USE à la deuxième période par $\mathcal{S}_{i2}^{\mathrm{II}}$. Alors

$$d = \bar{y}_{2..} - \bar{y}_{1..} = \frac{1}{mn}\sum_{i\in\mathcal{S}^{\mathrm{I}}}\left(\sum_{j\in\mathcal{S}_{i2}^{\mathrm{II}}} y_{2ij} - \sum_{j\in\mathcal{S}_{i1}^{\mathrm{II}}} y_{1ij}\right).$$

En désignant les espérances par rapport au premier degré d'échantillonnage par $\mathrm{E_I}$, et celles par rapport au deuxième degré d'échantillonnage par $\mathrm{E_{II}}$, nous avons que la variance sous le plan de d est égale à

$$\begin{aligned}
\mathrm{V}[d] &= \mathrm{E_I V_{II}}[d \mid \mathcal{S}^{\mathrm{I}}] + \mathrm{V_I E_{II}}[d \mid \mathcal{S}^{\mathrm{I}}] \\
&= \frac{1}{m^2n^2}\mathrm{E_I}\left\{\sum_{i\in\mathcal{S}^{\mathrm{I}}}\mathrm{V_{II}}\left[\sum_{j\in\mathcal{S}_{i2}^{\mathrm{II}}} y_{2ij} - \sum_{j\in\mathcal{S}_{i1}^{\mathrm{II}}} y_{1ij}\right]\right\} \\
&\quad + \frac{1}{m^2n^2}\mathrm{V_I}\left\{\sum_{i\in\mathcal{S}^{\mathrm{I}}}\mathrm{E_{II}}\left[\sum_{j\in\mathcal{S}_{i2}^{\mathrm{II}}} y_{2ij} - \sum_{j\in\mathcal{S}_{i1}^{\mathrm{II}}} y_{1ij}\right]\right\} \\
&= \frac{1}{m^2n^2}\mathrm{E_I}\left\{\sum_{i\in\mathcal{S}^{\mathrm{I}}}\mathrm{V_{II}}\left[\sum_{j\in\mathcal{S}_{i2}^{\mathrm{II}}} y_{2ij}\right] + \mathrm{V_{II}}\left[\sum_{j\in\mathcal{S}_{i1}^{\mathrm{II}}} y_{1ij}\right]\right\} \\
&\quad + \frac{1}{m^2n^2}\mathrm{V_I}\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}} m\bar{Y}_{2i\cdot} - m\bar{Y}_{1i\cdot}\right] \\
&= \frac{1}{m^2n^2}\mathrm{E_I}\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}}\left(1-\frac{m}{M}\right)mS_{2wi}^2 + \left(1-\frac{m}{M}\right)mS_{1wi}^2\right] \\
&\quad + \frac{1}{n^2}\left(1-\frac{n}{N}\right)n(S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}}S_{1b}S_{2b}) \\
&= \frac{1}{m^2n^2}nm\left(1-\frac{m}{M}\right)m(S_{2w}^2 + S_{1w}^2) \\
&\quad + \frac{1}{n}\left(1-\frac{n}{N}\right)(S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}}S_{1b}S_{2b}) \\
&= \left(1-\frac{n}{N}\right)\frac{2S_b^2(1-\rho^{\mathrm{I}})}{n} + \left(1-\frac{m}{M}\right)\frac{2S_w^2}{mn},
\end{aligned}$$

où la dernière égalité repose sur l'hypothèse des conditions symétriques (2.9).

Preuve de la proposition 3. Désignons l'échantillon d'UPE par $\mathcal{S}^{\mathrm{I}}$, et l'échantillon d'USE par $\mathcal{S}_i^{\mathrm{II}}$. Alors

$$d = \bar{y}_{2..} - \bar{y}_{1..} = \frac{1}{mn}\sum_{i\in\mathcal{S}^{\mathrm{I}}}\sum_{j\in\mathcal{S}_i^{\mathrm{II}}}(y_{2ij} - y_{1ij}).$$

En désignant les espérances par rapport au premier degré d'échantillonnage par $\mathrm{E_I}$, et celles par rapport au deuxième degré d'échantillonnage par $\mathrm{E_{II}}$, nous avons que la variance sous le plan de d est égale à

$$\begin{aligned}
\mathrm{V}[d] &= \mathrm{E_I V_{II}}[d \mid \mathcal{S}^{\mathrm{I}}] + \mathrm{V_I E_{II}}[d \mid \mathcal{S}^{\mathrm{I}}] \\
&= \frac{1}{m^2n^2}\mathrm{E_I}\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}}\mathrm{V_{II}}\sum_{j\in\mathcal{S}_i^{\mathrm{II}}}(y_{2ij} - y_{1ij})\right] \\
&\quad + \frac{1}{m^2n^2}\mathrm{V_I}\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}}\mathrm{E_{II}}\sum_{j\in\mathcal{S}_i^{\mathrm{II}}}(y_{2ij} - y_{1ij})\right] \\
&= \frac{1}{m^2n^2}\mathrm{E_I}m\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}}\left(1-\frac{m}{M}\right)(S_{2wi}^2 + S_{1wi}^2 - 2S_{2wi}S_{1wi}\rho^{\mathrm{II}})\right] \\
&\quad + \frac{1}{m^2n^2}\mathrm{V_I}\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}} m(\bar{Y}_{2ij} - \bar{Y}_{1ij})\right] \\
&= \frac{1}{mn^2}n\left(1-\frac{m}{M}\right)(\bar{S}_{1w}^2 + \bar{S}_{2w}^2 - 2\rho^{\mathrm{II}}\bar{S}_{1w}\bar{S}_{2w}) \\
&\quad + \frac{1}{n^2}n\left(1-\frac{n}{N}\right)(S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}}S_{1b}S_{2b}) \\
&= \left(1-\frac{n}{N}\right)\frac{S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}}S_{1b}S_{2b}}{n} \\
&\quad + \left(1-\frac{m}{M}\right)\frac{\bar{S}_{1w}^2 + \bar{S}_{2w}^2 - 2\rho^{\mathrm{II}}\bar{S}_{1w}\bar{S}_{2w}}{mn} \\
&= 2\left(1-\frac{n}{N}\right)\frac{S_b^2(1-\rho^{\mathrm{I}})}{n} \\
&\quad + 2\left(1-\frac{m}{M}\right)\frac{S_w^2(1-\rho^{\mathrm{II}})}{mn},
\end{aligned}$$

avec la dernière égalité vérifiée sous les conditions de symétrie

Preuve de la proposition 4. La fonction lagrangienne de minimisation de (2.11) sous la contrainte (3.1) est

$$\begin{aligned}
L(n_1, m_1, n_2, m_2, \lambda) = &\left(1-\frac{n_1}{N}\right)\frac{S_b^2}{n_1} + \left(1-\frac{n_2}{N}\right)\frac{S_b^2}{n_2} \\
&+ \left(1-\frac{m_1}{M}\right)\frac{\bar{S}_w^2}{n_1m_1} + \left(1-\frac{m_2}{M}\right)\frac{\bar{S}_w^2}{n_2m_2} \\
&- \lambda(c_1^{\mathrm{I}}n_1 + c_1^{\mathrm{II}}n_1m_1 + c_2^{\mathrm{I}}n_2 + c_2^{\mathrm{II}}n_2m_2 - C_0).
\end{aligned}$$

La résolution des conditions de premier ordre de cette fonction lagrangienne donne

$$-\lambda = \frac{m_1 S_b^2 + \left(1 - \frac{m_1}{M}\right)\bar{S}_w^2}{n_1^2 m_1 (c_1^{\mathrm{I}} + c_1^{\mathrm{II}} m_1)} = \frac{m_2 S_b^2 + \left(1 - \frac{m_2}{M}\right)\bar{S}_w^2}{n_2^2 m_2 (c_2^{\mathrm{I}} + c_1^{\mathrm{II}} m_2)} = \frac{\bar{S}_w^2}{m_1^2 n_1^2 c_1^{\mathrm{II}}} = \frac{\bar{S}_w^2}{m_2^2 n_2^2 c_2^{\mathrm{II}}}.$$

En utilisant ces conditions, nous avons

$$m^2 n^2 c^{\mathrm{II}} \left[m S_b^2 + \left(1 - \frac{m}{M}\right)\bar{S}_w^2 \right] = n^2 m (c^{\mathrm{I}} + c^{\mathrm{II}} m)\bar{S}_w^2,$$

qui peut s'écrire sous la forme

$$\begin{aligned} 0 &= (c^{\mathrm{I}} + c^{\mathrm{II}} m)\bar{S}_w^2 - m c^{\mathrm{II}} \left[m S_b^2 + \left(1 - \frac{m}{M}\right)\bar{S}_w^2 \right] \\ &= (c^{\mathrm{I}} + c^{\mathrm{II}} m) M \bar{S}_w^2 - m c^{\mathrm{II}} [M m S_b^2 + (M - m)\bar{S}_w^2] \\ &= c^{\mathrm{I}} M \bar{S}_w^2 + m c^{\mathrm{II}} M \bar{S}_w^2 - m^2 c^{\mathrm{II}} M S_b^2 - m c^{\mathrm{II}} M \bar{S}_w^2 + m^2 c^{\mathrm{II}} \bar{S}_w^2 \\ &= c^{\mathrm{I}} M \bar{S}_w^2 + m^2 c^{\mathrm{II}} (\bar{S}_w^2 - M S_b^2). \end{aligned}$$

D'où,

$$m = \sqrt{\frac{c^{\mathrm{I}}}{c^{\mathrm{II}}} \frac{\bar{S}_w^2}{S_b^2 - \bar{S}_w^2 / M}}.$$

D'après le budget de l'enquête (3.1), nous trouvons que le nombre de grappes est

$$n = \frac{C_0}{2(c^{\mathrm{I}} + m c^{\mathrm{II}})} = \frac{C_0}{2\{c^{\mathrm{I}} + [c^{\mathrm{I}} c^{\mathrm{II}} \bar{S}_w^2 / (S_b^2 - \bar{S}_w^2 / M)]^{1/2}\}}.$$

En introduisant ces expressions dans (2.11) et en utilisant les relations d'égalité (2.9), nous obtenons la variance de l'estimateur sous la forme

$$\begin{aligned} \mathrm{V}_{e,\iota}[d] &= 2\left(1 - \frac{n}{N}\right)\frac{S_b^2}{n} + 2\left(1 - \frac{m}{M}\right)\frac{\bar{S}_w^2}{mn} \\ &= 2\left[1 - \frac{C_0}{2(c^{\mathrm{I}} + m c^{\mathrm{II}})N}\right]\frac{2(c^{\mathrm{I}} + m c^{\mathrm{II}})S_b^2}{C_0} \\ &\quad + 4\left(1 - \frac{m}{M}\right)\frac{\bar{S}_w^2 (c^{\mathrm{I}} + m c^{\mathrm{II}})}{m C_0} \\ &= 2\left[\frac{2(c^{\mathrm{I}} + m c^{\mathrm{II}})}{C_0} - \frac{1}{N}\right] S_b^2 \\ &\quad + 4\left(\frac{1}{m} - \frac{1}{M}\right)\frac{\bar{S}_w^2 (c^{\mathrm{I}} + m c^{\mathrm{II}})}{C_0} \\ &= \frac{4(c^{\mathrm{I}} + m c^{\mathrm{II}})}{C_0}\left[S_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^2\right] - \frac{2}{N} S_b^2 \\ &= \frac{4\left[c^{\mathrm{I}} + \sqrt{c^{\mathrm{I}} c^{\mathrm{II}} \bar{S}_w^2 / (S_b^2 - \bar{S}_w^2 / M)}\right]}{C_0} \\ &\quad \times \left[S_b^2 + \left(\sqrt{\frac{c^{\mathrm{II}}}{c^{\mathrm{I}}} \frac{S_b^2 - \bar{S}_w^2 / M}{\bar{S}_w^2}} - \frac{1}{M}\right)\bar{S}_w^2\right] - \frac{2}{N} S_b^2. \end{aligned}$$

Preuve de la proposition 5. La fonction lagrangienne de minimisation de (2.13) sous la contrainte (3.4) est

$$\begin{aligned} L(n, m_1, m_2, \lambda) &= 2\left(1 - \frac{n}{N}\right)\frac{(1 - \rho^{\mathrm{I}})S_b^2}{n} \\ &\quad + \left(1 - \frac{m_1}{M}\right)\frac{\bar{S}_{1w}^2}{n m_1} + \left(1 - \frac{m_2}{M}\right)\frac{\bar{S}_{2w}^2}{n m_2} \\ &\quad - \lambda(c_{12}^{\mathrm{I}} n + c_1^{\mathrm{II}} n m_1 + c_2^{\mathrm{II}} n m_2 - C_0). \end{aligned}$$

Les conditions de premier ordre sont :

$$\begin{aligned} \frac{\partial L}{\partial n} &= -2\frac{(1 - \rho^{\mathrm{I}})S_b^2}{n^2} - \left(1 - \frac{m_1}{M}\right)\frac{\bar{S}_{1w}^2}{n^2 m_1} \\ &\quad - \left(1 - \frac{m_2}{M}\right)\frac{\bar{S}_{2w}^2}{n^2 m_2} - \lambda(c_{12}^{\mathrm{I}} + c_1^{\mathrm{II}} m_1 + c_2^{\mathrm{II}} m_2), \\ \frac{\partial L}{\partial m_1} &= -\frac{\bar{S}_{1w}^2}{n m_1^2} - \lambda c_1^{\mathrm{II}} n, \\ \frac{\partial L}{\partial m_2} &= -\frac{\bar{S}_{2w}^2}{n m_2^2} - \lambda c_2^{\mathrm{II}} n, \\ \frac{\partial L}{\partial \lambda} &= c_{12}^{\mathrm{I}} n + c_1^{\mathrm{II}} n m_1 + c_2^{\mathrm{II}} n m_2 - C_0 = 0. \end{aligned}$$

En exprimant $-\lambda n$ d'après ces conditions, nous obtenons :

$$\begin{aligned} -\lambda n &= \frac{\bar{S}_{1w}^2}{m_1^2 n c_1^{\mathrm{II}}} = \frac{\bar{S}_{2w}^2}{m_2^2 n c_2^{\mathrm{II}}} = 2(1 - \rho^{\mathrm{I}})\frac{S_b^2}{C_0} \\ &\quad + \left(1 - \frac{m_1}{M}\right)\frac{\bar{S}_{1w}^2}{m_1 C_0} + \left(1 - \frac{m_2}{M}\right)\frac{\bar{S}_{2w}^2}{m_2 C_0}. \end{aligned}$$

Alors

$$\begin{aligned} &\frac{1}{m_2} = \frac{1}{m_1}\sqrt{\frac{c_2^{\mathrm{II}}}{c_1^{\mathrm{II}}}\frac{\bar{S}_{1w}^2}{\bar{S}_{2w}^2}} \equiv \frac{1}{\kappa m_1}, \\ &\frac{1}{m_1^2}\frac{(c_{12}^{\mathrm{I}} + c_1^{\mathrm{II}} m_1 + \kappa c_2^{\mathrm{II}} m_1)\bar{S}_{1w}^2}{c_1^{\mathrm{II}}} = 2(1 - \rho^{\mathrm{I}})S_b^2 \\ &\qquad + \left(\frac{1}{m_1} - \frac{1}{M}\right)\bar{S}_{1w}^2 + \left(\frac{1}{\kappa m_1} - \frac{1}{\kappa M}\right)\bar{S}_{2w}^2, \\ &0 = [2(1 - \rho^{\mathrm{I}})S_b^2 \kappa c_1^{\mathrm{II}} - \bar{S}_{1w}^2 \kappa c_1^{\mathrm{II}} / M - \bar{S}_{2w}^2 c_1^{\mathrm{II}} / M] m_1^2 \\ &\qquad + [\bar{S}_{1w}^2 \kappa c_1^{\mathrm{II}} + \bar{S}_{2w}^2 c_1^{\mathrm{II}} - c_1^{\mathrm{II}} \bar{S}_{1w}^2 \kappa - \kappa^2 c_2^{\mathrm{II}} \bar{S}_{1w}^2] m_1 - c_{12}^{\mathrm{I}} \bar{S}_{1w}^2 \kappa, \\ &D = [\bar{S}_{1w}^2 \kappa c_1^{\mathrm{II}} + \bar{S}_{2w}^2 c_1^{\mathrm{II}} - c_1^{\mathrm{II}} \bar{S}_{1w}^2 \kappa - \kappa^2 c_2^{\mathrm{II}} \bar{S}_{1w}^2]^2 \\ &\qquad + 4[2(1 - \rho^{\mathrm{I}})S_b^2 \kappa - \bar{S}_{1w}^2 \kappa / M - \bar{S}_{2w}^2 / M] c_1^{\mathrm{II}} c_{12}^{\mathrm{I}} \bar{S}_{1w}^2 \geq 0, \\ &m_1 = \frac{c_1^{\mathrm{II}} \bar{S}_{1w}^2 \kappa + \kappa^2 c_2^{\mathrm{II}} \bar{S}_{1w}^2 - \bar{S}_{1w}^2 \kappa c_1^{\mathrm{II}} - \bar{S}_{2w}^2 c_1^{\mathrm{II}} \pm \sqrt{D}}{4(1 - \rho^{\mathrm{I}})S_b^2 \kappa c_1^{\mathrm{II}} - 2\bar{S}_{1w}^2 \kappa c_1^{\mathrm{II}} / M - 2\bar{S}_{2w}^2 c_1^{\mathrm{II}} / M}. \end{aligned}$$

La solution pour $-\sqrt{D}$ donne une valeur négative de m_1, et doit être écartée.

Les autres caractéristiques du plan sont

$$m_2 = \kappa m_1, \quad n = \frac{C_0}{c_{12}^{\mathrm{I}} + m_1 c_1^{\mathrm{II}} + m_2 c_1^{\mathrm{II}}}, \quad \kappa = \sqrt{\frac{c_1^{\mathrm{II}}}{c_2^{\mathrm{II}}}\frac{\bar{S}_{2w}^2}{\bar{S}_{1w}^2}}.$$

La variance de l'estimateur de la différence peut être calculée en utilisant (2.15).

Sous les conditions de symétrie, $\kappa = 1$, et

$$D = 4[2(1-\rho^{\mathrm{I}})S_b^2 - 2\bar{S}_w^2/M]\,c^{\mathrm{II}}c_{12}^{\mathrm{I}}\bar{S}_w^2$$

est non négative, à moins que l'expression entre les crochets soit négative (ce qui ne peut se produire que si ρ^{I} est grande et que M est petite. Dans ce cas, une solution de coin $m = M$ est réalisée). En outre,

$$m = m_1 = m_2 = \sqrt{\frac{\bar{S}_w^2 c_{12}^{\mathrm{I}}}{2[(1-\rho^{\mathrm{I}})S_b^2 - \bar{S}_w^2/M]c^{\mathrm{II}}}},$$

$$n = \frac{C_0}{c_{12}^{\mathrm{I}} + 2mc^{\mathrm{II}}} = \frac{C_0}{c_{12}^{\mathrm{I}} + \sqrt{\dfrac{2\bar{S}_w^2 c_{12}^{\mathrm{I}} c^{\mathrm{II}}}{(1-\rho^{\mathrm{I}})S_b^2 - \bar{S}_w^2/M}}},$$

$$\begin{aligned}
&\mathrm{V}_{e,o}[d] \\
&= 2\left(1-\frac{n}{N}\right)\frac{(1-\rho^{\mathrm{I}})S_b^2}{n} + 2\left(1-\frac{m}{M}\right)\frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2}{nm} \\
&= \frac{2}{n}\left[(1-\rho^{\mathrm{I}})S_b^2 + 2\left(1-\frac{m}{M}\right)\frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2}{m}\right] - \frac{2(1-\rho^{\mathrm{I}})S_b^2}{N} \\
&= \frac{2}{C_0}(c_{12}^{\mathrm{I}} + 2mc^{\mathrm{II}}) \\
&\times\left[(1-\rho^{\mathrm{I}})S_b^2 + 2\left(\frac{1}{m}-\frac{1}{M}\right)(1-\rho^{\mathrm{II}})\bar{S}_w^2\right] - \frac{2(1-\rho^{\mathrm{I}})S_b^2}{N} \\
&= \frac{2}{C_0}(c_{12}^{\mathrm{I}} + 2mc^{\mathrm{II}}) \\
&\times\left[(1-\rho^{\mathrm{I}})S_b^2 + \frac{2}{m}(1-\rho^{\mathrm{II}})\bar{S}_w^2 - \frac{2}{M}(1-\rho^{\mathrm{II}})\bar{S}_w^2\right] - \frac{2(1-\rho^{\mathrm{I}})S_b^2}{N} \\
&= \frac{2}{C_0}\left\{c_{12}^{\mathrm{I}}(1-\rho^{\mathrm{I}})S_b^2 + 2(1-\rho^{\mathrm{II}})\bar{S}_w^2\left[2c^{\mathrm{II}} - \frac{c_{12}^{\mathrm{I}}}{M}\right]\right. \\
&\quad \left. + \frac{2}{m}c_{12}^{\mathrm{I}}(1-\rho^{\mathrm{II}})\bar{S}_w^2 + 2mc^{\mathrm{II}}\left[(1-\rho^{\mathrm{I}})S_b^2 - \frac{2}{M}(1-\rho^{\mathrm{II}})\bar{S}_w^2\right]\right\} \\
&\quad - \frac{2(1-\rho^{\mathrm{I}})S_b^2}{N} \\
&= \frac{2}{C_0}\left\{c_{12}^{\mathrm{I}}(1-\rho^{\mathrm{I}})S_b^2 + 2(1-\rho^{\mathrm{II}})\bar{S}_w^2\left[2c^{\mathrm{II}} - \frac{c_{12}^{\mathrm{I}}}{M}\right]\right. \\
&\quad + 2(1-\rho^{\mathrm{II}})\sqrt{2[(1-\rho^{\mathrm{I}})S_b^2 - \bar{S}_w^2/M]\bar{S}_w^2 c^{\mathrm{II}} c_{12}^{\mathrm{I}}} \\
&\quad \left. + \sqrt{\frac{2\bar{S}_w^2 c_{12}^{\mathrm{I}} c^{\mathrm{II}}}{(1-\rho^{\mathrm{I}})S_b^2 - \bar{S}_w^2/M}}\left[(1-\rho^{\mathrm{I}})S_b^2 - \frac{2}{M}(1-\rho^{\mathrm{II}})\bar{S}_w^2\right]\right\} \\
&\quad - \frac{2(1-\rho^{\mathrm{I}})S_b^2}{N}.
\end{aligned}$$

Preuve de la proposition 6. La fonction lagrangienne de minimisation de (2.15) sous la contrainte (3.6) est

$$\begin{aligned}
L(n, m, \lambda) &= 2\left(1-\frac{n}{N}\right)\frac{(1-\rho^{\mathrm{I}})S_b^2}{n} + 2\left(1-\frac{m}{M}\right)\frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2}{nm} \\
&\quad - \lambda(c_{12}^{\mathrm{I}} n + c_{12}^{\mathrm{II}} nm - C_0).
\end{aligned}$$

Les conditions de premier ordre sont :

$$\begin{aligned}
0 &= \frac{\partial L}{\partial n} = -2\frac{(1-\rho^{\mathrm{I}})S_b^2}{n^2} \\
&\quad - 2\left(1-\frac{m}{M}\right)\frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2}{n^2 m} - \lambda(c_{12}^{\mathrm{I}} + c_{12}^{\mathrm{II}} m), \\
0 &= \frac{\partial L}{\partial m} = -2\frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2}{nm^2} - \lambda c_{12}^{\mathrm{II}} n, \\
0 &= \frac{\partial L}{\partial \lambda} = c_{12}^{\mathrm{I}} n + c_{12}^{\mathrm{II}} nm - C_0.
\end{aligned}$$

En exprimant $-\lambda n^2$ d'après ces conditions, nous obtenons :

$$\begin{aligned}
-\lambda n^2/2 &= \frac{(1-\rho^{\mathrm{I}})S_b^2}{c_{12}^{\mathrm{I}} + c_{12}^{\mathrm{II}} m} + \left(1-\frac{m}{M}\right)\frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2}{m(c_{12}^{\mathrm{I}} + c_{12}^{\mathrm{II}} m)} \\
&= \frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2}{m^2 c_{12}^{\mathrm{II}}}.
\end{aligned}$$

D'où,

$$\begin{aligned}
&(1-\rho^{\mathrm{I}})S_b^2 Mm^2 c_{12}^{\mathrm{II}} + (M-m)(1-\rho^{\mathrm{II}})\bar{S}_w^2 mc_{12}^{\mathrm{II}} \\
&\qquad - (1-\rho^{\mathrm{II}})M\bar{S}_w^2(c_{12}^{\mathrm{I}} + c_{12}^{\mathrm{II}} m) = 0, \\
&[(1-\rho^{\mathrm{I}})S_b^2 Mc_{12}^{\mathrm{II}} - (1-\rho^{\mathrm{II}})\bar{S}_w^2 c_{12}^{\mathrm{II}}]m^2 \\
&\qquad - [(1-\rho^{\mathrm{II}})M\bar{S}_w^2 c_{12}^{\mathrm{I}}] = 0, \\
m &= \sqrt{\frac{(1-\rho^{\mathrm{II}})M\bar{S}_w^2 c_{12}^{\mathrm{I}}}{[(1-\rho^{\mathrm{I}})S_b^2 M - (1-\rho^{\mathrm{II}})\bar{S}_w^2]c_{12}^{\mathrm{II}}}} \\
&= \sqrt{\frac{c_{12}^{\mathrm{I}}}{c_{12}^{\mathrm{II}}}\frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2}{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2/M}}.
\end{aligned}$$

D'après le budget de l'enquête (3.6),

$$n = \frac{C_0}{c_{12}^{\mathrm{I}} + c_{12}^{\mathrm{II}} m} = \frac{C_0}{c_{12}^{\mathrm{I}} + \sqrt{\dfrac{(1-\rho^{\mathrm{II}})\bar{S}_w^2 c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}}}{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2/M}}}.$$

Enfin, la variance de l'estimateur de la différence est

$$
\begin{aligned}
&\mathrm{V}_{e,o}[d] \\
&= \frac{2}{C_0}\left(c_{12}^{\mathrm{I}} + \sqrt{\frac{(1-\rho^{\mathrm{II}})\,\bar{S}_w^2 c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}}}{(1-\rho^{\mathrm{I}})\,S_b^2 - (1-\rho^{\mathrm{II}})\,\bar{S}_w^2 \,/\, M}} \right) \\
&\quad \times \left[(1-\rho^{\mathrm{I}})\, S_b^2 \right. \\
&\qquad \left. + \left(\sqrt{\frac{c_{12}^{\mathrm{II}}}{c_{12}^{\mathrm{I}}} \frac{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\,\bar{S}_w^2 \,/\, M}{(1-\rho^{\mathrm{II}})\,\bar{S}_w^2}} - \frac{1}{M} \right) (1-\rho^{\mathrm{II}})\,\bar{S}_w^2 \right] \\
&\quad - \frac{2(1-\rho^{\mathrm{I}})S_b^2}{N} \\
&= \frac{2}{C_0}\left\{ (1-\rho^{\mathrm{I}}) S_b^2 c_{12}^{\mathrm{I}} \right. \\
&\qquad + (1-\rho^{\mathrm{II}})\,\bar{S}_w^2 \sqrt{c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}} \frac{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\,\bar{S}_w^2 \,/\, M}{(1-\rho^{\mathrm{II}})\,\bar{S}_w^2}} \\
&\qquad + \left[(1-\rho^{\mathrm{I}})\, S_b^2 - \frac{1}{M}(1-\rho^{\mathrm{II}})\,\bar{S}_w^2 \right] \\
&\qquad \times \sqrt{\frac{(1-\rho^{\mathrm{II}})\,\bar{S}_w^2 c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}}}{(1-\rho^{\mathrm{I}})\,S_b^2 - (1-\rho^{\mathrm{II}})\,\bar{S}_w^2 \,/\, M}} \\
&\qquad \left. + (1-\rho^{\mathrm{II}})\,\bar{S}_w^2 \left(c_{12}^{\mathrm{II}} - \frac{c_{12}^{\mathrm{I}}}{M} \right) \right\} \\
&\quad - \frac{2(1-\rho^{\mathrm{I}})S_b^2}{N}.
\end{aligned}
$$

Preuve de la proposition 7. En ignorant les termes de correction pour population finie d'ordres $O(N^{-1})$ et $O(M^{-1})$, l'équation (3.3) peut s'écrire :

$$
\begin{aligned}
\mathrm{V}_{e,i}[d] &\approx \frac{4\left(c^{\mathrm{I}} + \sqrt{c^{\mathrm{I}} c^{\mathrm{II}} \bar{S}_w^2 / S_b^2} \right)}{C_0} \left[S_b^2 + \left(\sqrt{\frac{c^{\mathrm{II}}}{c^{\mathrm{I}}}} \bar{S}_w^2 S_b^2 \right) \right] \\
&= \frac{4}{C_0}\left(c^{\mathrm{I}} S_b^2 + c^{\mathrm{II}} \bar{S}_w^2 + 2\sqrt{c^{\mathrm{I}} c^{\mathrm{II}} S_b^2 \bar{S}_w^2} \right) \\
&= \frac{4}{C_0}\left(\sqrt{c^{\mathrm{I}} S_b^2} + \sqrt{c^{\mathrm{II}} \bar{S}_w^2} \right)^2.
\end{aligned}
$$

De même, l'équation (3.8) peut s'écrire

$$
\begin{aligned}
\mathrm{V}_{e,o}[d] &\approx \frac{2}{C_0}\left[(1-\rho^{\mathrm{I}})\, S_b^2 c_{12}^{\mathrm{I}} \right. \\
&\qquad \left. + 2\sqrt{c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}} (1-\rho^{\mathrm{I}})\, S_b^2 (1-\rho^{\mathrm{II}})\,\bar{S}_w^2} + (1-\rho^{\mathrm{II}})\,\bar{S}_w^2 c_{12}^{\mathrm{II}} \right] \\
&= \frac{2}{C_0}\left[\sqrt{(1-\rho^{\mathrm{I}})\, S_b^2 c_{12}^{\mathrm{I}}} + \sqrt{(1-\rho^{\mathrm{II}})\,\bar{S}_w^2 c_{12}^{\mathrm{II}}} \right]^2.
\end{aligned}
$$

L'énoncé de la proposition 7 découle directement de ces deux expressions.

Bibliographie

Binder, D.A., et Hidiroglou, M.A. (1988). Sampling in time. Dans *Handbook of Statistics*, (Éds., P.R. Krishnaiah et C.R. Rao), North Holland, Amsterdam, 6, 187-211.

Cochran, W.G. (1977). *Sampling Techniques*, 3e Éd., New York : John Wiley & Sons, Inc.

Eckler, A.R. (1955). Rotation sampling. *Annals of Mathematical Statistics*, 26(4), 664-685.

Ernst, L.R. (1999). The maximization and minimization of sample overlap problems: A half century of results. Rapport technique, U.S. Bureau of Labor Statistics.

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

Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York : John Wiley & Sons, Inc.

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

Kish, L. (1995). *Survey Sampling*, 3e Éd., New York : John Wiley & Sons, Inc.

Lehtonen, R., et Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys, Statistics in Practice*, 2e Éd., New York : John Wiley & Sons, Inc.

Mas-Colell, A., Whinston, M.D. et Green, J.R. (1995). *Microeconomic Theory*, Oxford University Press, Oxford, UK.

McDonald, T.L. (2003). Review of environmental monitoring methods: Survey designs. *Environmental Monitoring and Assessment*, 85, 277-292.

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

Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, Séries B, 12(2), 241-255.

Rao, J.N.K., et Graham, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59(306), 492-509.

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

Scott, C.T. (1998). Sampling methods for estimating change in forest resources. *Ecological Applications*, 8(2), 228-233.

Thompson, M.E. (1997). Theory of Sample Surveys. *Monographs on Statistics and Applied Probability*, New York : Chapman & Hall/CRC, 74.

Thompson, S.K. (1992). *Sampling*, New York : John Wiley & Sons, Inc.

Wolter, K.M. (2007). *Introduction to Variance Estimation*, 2e Éd., New York : Springer.

Techniques d'enquête, juin 2011
Vol. 37, N° 1, pp. 103-110
Statistique Canada, N° 12-001-X au catalogue

À propos de l'efficacité de l'échantillonnage à probabilité proportionnelle à la taille aléatoire

Paul Knottnerus [1]

Résumé

Dans le présent document, on examine l'efficacité de l'estimateur Horvitz-Thompson au moyen d'un échantillon systématique de probabilité proportionnelle à la taille (PPT) tiré d'une liste en ordre aléatoire. Plus précisément, l'efficacité est comparée avec celle d'un estimateur par quotient ordinaire. Les résultats théoriques sont confirmés d'une manière empirique à l'aide d'une étude de simulation basée sur des données hollandaises de l'Indice des prix à la production.

Mots clés : Estimateur de Horvitz-Thompson ; Indice des prix à la production ; estimateur par quotient ; coefficient d'autocorrélation d'échantillonnage.

1. Introduction

Lorsque la variable y à l'étude dans une population de N unités est plus ou moins proportionnelle à une variable de taille x, on peut s'appuyer sur l'estimateur par quotient d'un échantillon aléatoire simple (EAS) de taille n sans remplacement. Un estimateur de rechange en pareil cas est l'estimateur Horvitz-Thompson (HT), associé à un échantillon systématique à probabilité proportionnelle à la taille tiré d'une liste en ordre aléatoire, ci-après appelé l'échantillon PPT aléatoire.

Ces dernières années, plusieurs auteurs ont étudié les procédures d'estimation de la variance pour l'estimateur HT à partir d'un échantillon PPT aléatoire. Voir, entre autres, Brewer et Donadio (2003), Cumberland et Royall (1981), Deville (1999), Knottnerus (2003), Kott (1988 et 2005), Rosén (1997) et Stehman et Overton (1994). Pour une comparaison entre l'efficacité de l'estimateur par quotient et celle de l'estimateur PPT aléatoire, le lecteur est invité à consulter le rapport de Foreman et Brewer (1971), Cochran (1977) et les références fournies dans le présent document. L'un des inconvénients de ces comparaisons est que la correction pour population finie n'est pas prise en compte. Hartley et Rao (1962) tiennent compte de la correction pour population finie, mais sans formule explicite pour l'efficacité. En examinant de plus près les résultats de Gabler (1984), Qualité (2008) démontre que l'estimateur HT connexe d'un échantillonnage réjectif de Poisson de taille n est plus efficace que l'estimateur de Hansen-Hurwitz pour un plan d'échantillonnage avec remplacement. Aucune formule pour l'efficacité accrue n'est toutefois donnée.

Le principal objectif du présent document est d'établir des formules pour l'efficacité de l'estimateur PPT aléatoire par rapport à l'estimateur par quotient. À cette fin, nous présentons une formule simple pour la variation de la taille de l'échantillon requise pour maintenir la même variance lorsqu'un estimateur PPT aléatoire est remplacé par un estimateur par quotient. De la perspective du plan de sondage, ces formules sont valides lorsque $n = o(N)$, pour $N \to \infty$. Cette condition suggère que la correction pour population finie peut être omise pour ce genre de plan d'échantillonnage. Fait étonnant, comme nous le verrons dans un exemple à la section 4, l'échantillonnage PPT aléatoire peut réduire la variance de plus de 30 % comparativement à l'échantillonnage PPT *avec* remplacement, même lorsque la fraction d'échantillonnage n/N est beaucoup plus petite que 30 % ; voir aussi Kott (2005, page 436). De plus, les formules demeurent appropriées d'une perspective basée sur un modèle lorsque n et N sont du même ordre, à condition que N soit grand et que le modèle hypothétique pour les observations Y_i $(i = 1, ..., N)$ satisfasse à des conditions peu contraignantes.

Les grandes lignes du document sont décrites ci-après. La section 2 décrit une expression de rechange pour la variance de l'estimateur HT en fonction du coefficient d'autocorrélation de l'échantillonnage. L'estimateur de variance correspondant pour l'échantillonnage PPT aléatoire est prouvé comme non négatif avec une probabilité 1. La section 3 présente les formules pour l'efficacité de l'estimateur PPT aléatoire par rapport à l'estimateur par quotient pour divers modèles de données souvent rencontrés dans la pratique. La section 4 montre un exemple de données sur l'indice des prix à la production aux Pays-Bas, qui illustre les gains d'efficacité considérables qui peuvent être obtenus en pratique. Un contre-exemple démontre que l'échantillonnage PPT aléatoire n'est pas *toujours* avantageux. Le document se termine par un sommaire.

2. Une autre expression de la variance pour l'échantillonnage PPT aléatoire

Si l'on suppose une population $U = \{1, ..., N\}$, et que s est un échantillon de taille fixe n tiré de U sans

1. Paul Knottnerus, Statistics Netherlands, PO Box 24500, 2490 HA The Hague, Pays-Bas. Courriel : pkts@cbs.nl.

remplacement en fonction d'un plan d'échantillonnage donné, avec des probabilités d'inclusion du premier ordre de π_i et des probabilités d'inclusion du deuxième ordre de $\pi_{ij} (i, \quad j = 1, ..., N)$. L'estimateur HT de la population totale, $Y = \Sigma_{i \in U} Y_i$, est défini par $\hat{Y}_{\mathrm{HT}} = \Sigma_{i \in s} Y_i / \pi_i$. Supposons qu'il y a une mesure de la taille relative X_i (c'est-à-dire $X = \Sigma_{i \in U} X_i = 1$), de manière à ce que $X_i \leq 1/n$. En fait, on présume ici que les unités où $X_i > 1/n$ sont regroupées dans une strate de certitude distincte. Lorsque les π_i sont proportionnels à ces mesures de la taille, $\pi_i = nX_i$. En définissant que $Z_i = Y_i / X_i$, nous pouvons écrire Y comme moyenne pondérée de Z_i, c'est-à-dire que $Y = \mu_z = \Sigma_{i \in U} X_i Z_i$. De même, nous pouvons écrire l'estimateur HT de Y dans l'échantillonnage PPT aléatoire comme $\hat{Y}_{\mathrm{HT}} = \hat{Y}_{\mathrm{PPT}} = \bar{z}_s$, où $\bar{z}_s$ est la moyenne de l'échantillon de Z_i.

La variance de l'estimateur PPT aléatoire $\hat{Y}_{\mathrm{PPT}}$ est

$$\mathrm{var}(\hat{Y}_{\mathrm{PPT}}) = \frac{1}{n^2} \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) Z_i Z_j \qquad (1)$$

$$= -\frac{1}{2n^2} \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j)(Z_i - Z_j)^2 \qquad (2)$$

où $\pi_{ii} = \pi_i$. La première équation est attribuée à Horvitz et Thompson (1952), et la deuxième, à Sen (1953) et à Yates et Grundy (1953). L'expression de rechange suivante pour la variance est mieux adaptée à nos besoins :

$$\mathrm{var}(\hat{Y}_{\mathrm{PPT}}) = \mathrm{var}(\bar{z}_s) = \{1 + (n-1)\rho_z\} \frac{\sigma_z^2}{n}, \qquad (3)$$

où $\sigma_z^2 = \sum_{i \in U} X_i (Z_i - \mu_z)^2$, et

$$\rho_z = \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \frac{\pi_{ij}}{n(n-1)} \left(\frac{Z_i - \mu_z}{\sigma_z} \right) \left(\frac{Z_j - \mu_z}{\sigma_z} \right). \qquad (4)$$

Pour une preuve de (3), voir Knottnerus (2003, page 103). Soulignons que σ_z^2 / n aurait été la variance si l'échantillon avait été tiré avec remplacement, avec des probabilités de tirage de X_i.

Le coefficient d'autocorrélation de l'échantillonnage ρ_z dans (4) est une généralisation du coefficient de corrélation interclasse plus courant ρ dans l'échantillonnage systématique avec probabilités égales ; voir, par exemple, Cochran (1977, pages 209 et 240) et Särndal, Swensson et Wretman (1992, page 79). Soulignons que ρ_z est un paramètre de population fixe. L'expression *autocorrélation d'échantillonnage* est employée parce que ρ_z désigne l'autocorrélation entre deux observations sélectionnées au hasard, disons z_{s1} et z_{s2}, à partir de s. Par conséquent, la valeur de ρ_z dépend du plan d'échantillonnage. Plus précisément, dans le cas d'un échantillonnage avec remplacement, $\rho_z = 0$, tandis que selon l'EAS, $\rho_z = -1/(N-1)$.

Bien que les expressions exactes pour π_{ij} selon l'échantillonnage PPT aléatoire soient disponibles, elles peuvent être lourdes lorsque N est grand. Pour une expression exacte, voir Connor (1966) et pour une modification, voir Hidiroglou et Gray (1980). Nous utilisons ici une approximation proposée par Knottnerus (2003, page 197) :

$$\pi_{ijK} = n(n-1) \frac{X_i X_j (1 - X_i - X_j)}{\gamma (1 - 2X_i)(1 - 2X_j)} \qquad (5)$$

$$\gamma = \frac{1}{2} + \frac{1}{2} \sum_{i \in U} \frac{X_i}{1 - 2X_i}.$$

Il a été démontré que ces π_{ijK} satisfont aux restrictions de deuxième ordre pour π_{ij} :

$$\sum_{i,\, j \in U (j \neq i)} \pi_{ij} = n(n-1),$$

et

$$\sum_{j \in U (j \neq i)} \pi_{ij} = (n-1)\pi_i.$$

En outre, (5) est correct pour l'EAS lorsque $n \leq N$, tandis que π_{ijK} coïncide avec π_{ijBD} en fonction des plans spéciaux proposés par Brewer (1963a) et Durbin (1967) pour les échantillons PPT, où $n = 2$. De plus, l'élément π_{ijK} dans (5) peut être écrit sous forme factorisée, comme proposé par Brewer et Donadio (2003). Autrement dit,

$$\pi_{ijK} = \pi_i \pi_j (c_i + c_j) / 2, \qquad (6)$$

et

$$c_i = (n-1)/n\gamma(1 - 2X_i).$$

Une implication de l'approximation (5) est que $\pi_{ijK}/n(n-1)$ ne dépend pas de n. Par conséquent, l'approximation correspondante de ρ_z ne dépend pas de n (rappelons que nous avons présumé que $X_i < 1/n$).

Cette non-dépendance à n serait également survenue si nous avions utilisé l'approximation proposée par Hartley et Rao (1962) pour l'échantillonnage PPT aléatoire :

$$\pi_{ijHR} = n(n-1) X_i X_j \{1 + X_i + X_j - \mu_x + 2(X_i^2 + X_j^2 + X_i X_j) - 3\mu_x (X_i + X_j - \mu_x - 2\Sigma_{i \in U} X_i^3)\}, \qquad (7)$$

où $\mu_x = \Sigma_{i \in U} X_i^2$ (rappelons que $\mu_z = \Sigma_{i \in U} X_i Z_i$). Manifestement, $\pi_{ijHR} / n(n-1)$ ne dépend pas de n. À l'époque, Hartley et Rao ont supposé que $n = O(1)$ lorsque $N \to \infty$. De plus, en se reportant à une conversation privée avec J.N.K. Rao, Thompson et Wu (2008) déclarent que l'approximation (7) est valide lorsque $n = o(N)$ lorsque $N \to \infty$. Pour un exemple où (5) et (7) ne peuvent pas être utilisés pour toutes valeurs de n et N, voir l'annexe A.

Étant donné que (5) et (7) entraînent des approximations pour ρ_z dans l'échantillonnage PPT aléatoire qui sont $\rho_z\{1+o(1)\}$ lorsque $N \to \infty$, avec $n = o(N)$, (5) peut être utilisé pour calculer ρ_z dans la pratique lorsque $n << N$ et que N est grand. Pour faciliter l'exposé, on présume ici qu'il y a une constante positive c, de sorte que $\rho_z < -c/N$. Voir aussi Kott (2005, page 436), qui explique l'estimation de la variance sous l'échantillonnage PPT lorsque $n = O(N^{2/3})$.

Supposons que $\gamma = 1 + \mu_x + O(1/N^2)$ et que $\mu_x = O(1/N)$ (qui suit selon les conditions du théorème 1 ci-après). Il n'est pas difficile de voir que, après élimination des termes $O(1/nN)$, c_i dans (6) est identique à $c_{iHR} = (n-1)/\{n(1+\mu_x - 2X_i)\}$. Cette dernière expression est l'équation (11) de Brewer et Donadio, qui est basée sur π_{ijHR} en (7).

L'approche proposée ici est un peu différente de Knottnerus (2003). D'abord, réécrivez (5) comme suit

$$\pi_{ijK} = n(n-1)\frac{X_i X_j}{\gamma}\left(\frac{1/2}{1-2X_i} + \frac{1/2}{1-2X_j}\right). \quad (8)$$

En remplaçant (8) dans (4), nous obtenons une nouvelle approximation simple pour ρ_z :

$$\rho_z = \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} \frac{X_i X_j}{\gamma}\left(\frac{1/2}{1-2X_i}+\frac{1/2}{1-2X_j}\right)\left(\frac{Z_i-Y}{\sigma_z}\right)\left(\frac{Z_j-Y}{\sigma_z}\right)$$

$$= \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} \frac{X_i X_j}{\gamma}\left(\frac{1}{1-2X_i}\right)\left(\frac{Z_i-Y}{\sigma_z}\right)\left(\frac{Z_j-Y}{\sigma_z}\right)$$

$$= 0 - \sum_{i\in U}\frac{X_i^2}{\gamma(1-2X_i)}\left(\frac{Z_i-Y}{\sigma_z}\right)^2. \quad (9)$$

À la deuxième ligne, nous avons utilisé l'égalité $\sum_{i,j} m_{ij} v_i = \sum_{i,j} m_{ij} v_j$ lorsque $m_{ij} = m_{ji}$. À la dernière ligne, nous avons utilisé $\sum_{j\in U} X_j (Z_j - Y) = 0$.

Ensuite, supposons que $\bar{X}$ indique la moyenne de la population de $X_1, ..., X_N$ et définissons σ_x^2 et V_x^2 par

$$\sigma_x^2 = \sum_{i\in U} X_i (X_i - \mu_x)^2,$$

et

$$V_x^2 = \sum_{i\in U} (X_i - \bar{X})^2 / N,$$

respectivement. Dans le théorème suivant, (9) est encore plus simplifié.

Théorème 1. On suppose que $(Z_i - Y)/\sigma_z = O(1)$, lorsque $N \to \infty$ et qu'il y a des constantes positives c et C, de sorte que $V_x/\bar{X} < c$, $\sigma_x/\mu_x < c$ et $0 < X_i < C < 1/2$. Ainsi, pour de grandes valeurs de N et $n << N$,

$$\rho_z = -\frac{\sum_{i\in U} X_i^2 (Z_i - Y)^2}{\sum_{i\in U} X_i (Z_i - Y)^2}\left\{1 + O\left(\frac{1}{N}\right)\right\} + O\left(\frac{1}{N^2}\right). \quad (10)$$

Preuve. Comme $\bar{X} = 1/N$, on conclut des suppositions qui précèdent que la moyenne pondérée μ_x [$=\Sigma X_i^2 = N(V_x^2 + \bar{X}^2)$] est de l'ordre de $1/N$ et donc, $\sigma_x = O(1/N)$. Comme $(1-2X_i)^{-1} = 1 + 2X_i + O(X_i^2)$ pour $0 < X_i < C < 1/2$, ρ_z, de (9) peut être écrit lorsque $N \to \infty$, comme

$$\rho_z = -\sum_{i\in U}\frac{X_i^2}{\gamma}\left(\frac{Z_i - Y}{\sigma_z}\right)^2 + \frac{1}{\gamma} O\left(\sum_{i\in U} X_i^3\right),$$

où $\sum_{i\in U} X_i^3 = \sigma_x^2 + \mu_x^2 = O(N^{-2})$, et

$$\gamma = \frac{1}{2} + \frac{1}{2}\sum_{i\in U} X_i\{1 + 2X_i + O(X_i^2)\}$$

$$= 1 + \mu_x + O\left(\frac{1}{N^2}\right) = 1 + O\left(\frac{1}{N}\right),$$

dont suit (10). Voilà qui conclut la preuve.

En substituant (10) dans (3), nous obtenons

$$\text{var}(\hat{Y}_{\text{PPT}}) = \frac{\sigma_z^2}{n} - \frac{n-1}{n}\sum_{i\in U} X_i^2 (Z_i - Y)^2$$

$$= \frac{1}{n}\sum_{i\in U} X_i\{1 - (n-1)X_i\}(Z_i - Y)^2, \quad (11)$$

une équation qui est également proposée par Hartley et Rao (1962). Il convient de souligner que l'approximation (10) découle également directement de la substitution de l'approximation simple $\pi_{ijAP} = n(n-1)X_i X_j$ dans (4). De même, l'utilisation de π_{ijHR} donne lieu à une expression semblable à (9), et donc, à (10). De plus, l'utilisation directe de π_{ijAP} à (1) ou (2) pour le cas d'EAS, où $X_i = X_j = 1/N$ pourrait entraîner des erreurs de plus de 100 % pour les populations où $\bar{Y} = V_y^2$; voir Knottnerus (2003, pages 274-6). Donc, (1) et (2) sont plus sensibles aux petites erreurs dans π_{ij} que (3) et (4). Qui plus est, soulignons que lorsque n est petit à un point tel que $|n\rho_z| << 1$, nous pouvons établir que $\rho_z = 0$, ce qui donne la formule de variance avec remplacement de Hansen et Hurwitz (1943).

Pour estimer (3) au moyen de ρ_z, comme précédemment, dénotons par z_{s1} une observation choisie au hasard de *s*. Nous obtenons alors

$$\sigma_z^2 = \text{var}(z_{s1}) = \text{var}\{E(z_{s1}|s)\} + E\{\text{var}(z_{s1}|s)\}$$

$$= \text{var}(\bar{z}_s) + E\left(\frac{n-1}{n} s_z^2\right),$$

où

$$s_z^2 = \frac{1}{n-1}\sum_{i\in s}(Z_i - \bar{z}_s)^2.$$

Maintenant, à partir de (3), on constate que $s_z^2/(1-\rho_z)$ est un estimateur sans biais pour σ_z^2. Lorsque ρ_z est très petit, le terme $(1-\rho_z)$ peut être omis. Lorsque n est assez grand, le ratio ρ_z de (9) peut être estimé comme suit

$$\hat{\rho}_{z9} = -\frac{\sum_{i\in s} X_i (Z_i - \bar{z}_s)^2 / \hat{\gamma}(1-2X_i)}{\sum_{i\in s}(Z_i - \bar{z}_s)^2},$$

où

$$\hat{\gamma} = \frac{1}{2} + \frac{1}{2n}\sum_{i\in s}\frac{1}{1-2X_i}.$$

Comme $\hat{\gamma} \geq 1$ et $X_i \leq 1/n$, nous obtenons $\hat{\rho}_{z9} \geq -1/(n-2)$. Pour le biais d'un ratio estimatif lorsque n est petit, voir Cochran (1977, page 160).

De même, l'élément ρ_z dans (10) peut être estimé comme suit

$$\hat{\rho}_{z10} = -\frac{\sum_{i\in s} X_i (Z_i - \bar{z}_s)^2}{\sum_{i\in s}(Z_i - \bar{z}_s)^2} \geq \frac{-1}{n} > \frac{-1}{n-1}.$$

Donc, en remplaçant σ_z^2 et ρ_z dans (3) par $s_z^2/(1-\hat{\rho}_{z10})$ et $\hat{\rho}_{z10}$, respectivement, on obtient un estimateur de variance non négatif avec une probabilité de 1. On peut en dire autant pour $\hat{\rho}_{z9}$, lorsque $X_i \leq 1/(n+1)$. L'estimateur pour $\text{var}(\hat{Y}_{\text{PPT}})$ ainsi obtenu devient

$$\hat{\text{var}}_{\rho}(\hat{Y}_{\text{PPT}}) = \frac{\{1+(n-1)\hat{\rho}_{z9}\}s_z^2}{n(1-\hat{\rho}_{z9})}.$$

En outre, pour les valeurs modérées de N, l'estimateur $\hat{\rho}_{z9}$ a probablement de meilleures propriétés que $\hat{\rho}_{z10}$ parce que les éléments π_{ijK} sous-jacents à (9) satisfont exactement aux restrictions de deuxième ordre sans égard aux valeurs de n et de N.

3. Efficacité de $\hat{Y}_{\text{PPT}}$ pour grands n et N

3.1 Formules d'efficacité

Comme $X = 1$, l'estimateur par quotient pour Y devient

$$\hat{Y}_R = \frac{\bar{y}_s}{\bar{x}_s} = \frac{\sum_{i\in s} X_i Z_i}{\sum_{i\in s} X_i}.$$

Lorsque n est assez grand, l'approximation souvent utilisée pour sa variance est la suivante

$$\text{var}(\hat{Y}_R) = \frac{N(N-n)}{n(N-1)}\sum_{i\in U} X_i^2 (Z_i - Y)^2. \qquad (12)$$

À partir de (3) et (12), on peut constater que l'efficacité de $\hat{Y}_{\text{PPT}}$ par rapport à $\hat{Y}_R$ peut être écrite comme suit

$$Eff_{P/R} = \frac{\text{var}(\hat{Y}_R)}{\text{var}(\hat{Y}_{\text{PPT}})} = \frac{(N-n)\sum_{i\in U} X_i^2 (Z_i - Y)^2}{\{1+(n-1)\rho_z\}\sigma_z^2}, \qquad (13)$$

en supposant que $N/(N-1) \approx 1$. Lorsque l'on combine (10) et (13), on obtient

$$Eff_{P/R} = \frac{-(N-n)\rho_z}{1+(n-1)\rho_z}. \qquad (14)$$

Supposons maintenant que les observations Y_i satisfont au modèle :

$$Y_i = \mu X_i + \varepsilon_i, \qquad (15)$$

où $E(\varepsilon_i) = 0$, $E(\varepsilon_i^2) = \sigma^2 X_i^{\delta}$, et $E(\varepsilon_i \varepsilon_j) = 0$ $(i \neq j)$. Par conséquent, pour Z_i, nous obtenons $Z_i = \mu + u_i$, où $E(u_i) = 0$, $E(u_i^2) = \sigma^2 X_i^{\delta-2}$, et $E(u_i u_j) = 0 (i \neq j)$. D'après Kott (1988), δ se trouve souvent entre 1 et 2. Voir aussi Brewer (1963b). Brewer et Donadio (2003) ont démontré qu'en supposant un modèle comme (15), (7) et donc (10) et (14) se maintiennent lorsque n et N sont du même ordre lorsque $N \to \infty$. De plus, lorsque N est suffisamment grand, nous pouvons remplacer Y, ainsi que le numérateur et le dénominateur dans (10) par les valeurs attendues du modèle, ce qui donne

$$\rho_z = -\frac{\Sigma_{i\in U} X_i^{\delta}}{\Sigma_{i\in U} X_i^{\delta-1}}. \qquad (16)$$

Dans les prochaines sous-sections, nous examinons de plus près la relation entre δ et l'efficacité de $\hat{Y}_{\text{PPT}}$.

3.2 Efficacité de $\hat{Y}_{\text{PPT}}$ si $\delta = 2$

Comme $\delta = 2$, (16) donne $\rho_z = -\Sigma_{i\in U} X_i^2 = -\mu_x$, ce qui peut également être écrit comme suit

$$\rho_z = -\frac{1}{N}(1 + CV_x^2), \qquad (17)$$

parce que

$$\frac{1}{N}\sum_{i\in U} X_i^2 = V_x^2 + \bar{X}^2 = \bar{X}^2(1 + CV_x^2),$$

où $\bar{X} = 1/N$ et $CV_x = V_x/\bar{X}$ est le coefficient de variation de X_i. Lorsque l'on remplace (17) dans (14), on obtient

$$Eff_{P/R} = \frac{(N-n)(1+CV_x^2)}{N-(n-1)(1+CV_x^2)}.$$

Par conséquent, si $\delta = 2$, l'efficacité de l'échantillon PPT aléatoire est élevée lorsque la variabilité de X_i est élevée. Lorsque $CV_x = 0$, l'échantillonnage PPT aléatoire équivaut à l'échantillonnage EAS et manifestement, $Eff_{P/R} = 1$, en supposant que $(N-n+1) \approx (N-n)$; soulignons que cette hypothèse se maintient lorsque N est suffisamment grand et que $n/N < f_0 < 1$.

Soulignons qu'en remplaçant $n = n_{\text{PPT}}(1+CV_x^2)$ dans (12), on obtient à peu près le même résultat qu'en (3) et (10) en remplaçant n par n_{PPT}. Ainsi, lorsque $CV_x = 1{,}5$, l'échantillonnage PPT aléatoire dont la taille de l'échantillon $n_{\text{PPT}} = 100$ est aussi efficace que l'estimateur par quotient d'un EAS de taille $n_{\text{EAS}} = 325$. D'une manière plus

générale, en supposant que $(n-1)/n \approx 1$, on constate à partir de (3), (10) et (12) qu'un estimateur par quotient d'un EAS de taille n_{EAS} est aussi efficace qu'un échantillon PPT de taille n_{PPT} lorsque

$$n_{\text{EAS}} = -n_{\text{PPT}}\,\rho_z N. \qquad (18)$$

3.3 Efficacité de $\hat{Y}_{\text{PPT}}$ pour $\delta < 1$ par rapport à $\delta \geq 1$

Un autre cas spécial est $\delta = 1$. À partir de (16), $\rho_z = -1/N$ lorsque $\delta = 1$. Ensuite, on constate d'après (14) qu'en vertu du modèle (15), $Eff_{P/R} = 1 + O(N^{-1})$, à condition que $n/N < f_0 < 1$, lorsque $N \to \infty$ sans égard à la valeur de CV_x. De plus, on peut démontrer que $Eff_{P/R}$ est une fonction croissante de δ. Cette théorie est prouvée ci-après au Lemme 1. Par conséquent, pour $\delta < 1$, l'estimateur PPT aléatoire est moins efficace que l'estimateur par quotient, alors que pour $\delta > 1$, l'estimateur PPT aléatoire est plus efficace que l'estimateur par quotient.

Lemme 1. Supposons que $Eff_{P/R}$ et que ρ_z sont définis par (14) et (16), respectivement. Si $V_x^2 > 0$, alors $Eff_{P/R}$ est une fonction croissante monotone de δ.

Preuve. Écrivez ρ_z à partir de (16) comme une moyenne pondérée de X_i (négatif).

$$\rho_z = -u(\delta) = -\sum_{i \in U} w_i X_i,$$

où

$$w_i = \frac{X_i^{\delta-1}}{\sum_{i \in U} X_i^{\delta-1}} \quad [\text{Notons que } \mu_x = u(2)].$$

Supposons que $X_i > X_j (i \neq j)$, et définissons $h(\delta)$ comme $w_i/w_j = (X_i/X_j)^{\delta-1}$. Étant donné que $h(\delta)$ augmente en fonction de δ, le poids du grand X_i augmente comparativement à celui de X_j lorsque δ augmente. Par conséquent, $u(\delta)$ augmente et ρ_z décroît en fonction de δ. Il suffit donc de démontrer que $Eff_{P/R}$ diminue en fonction de ρ_z. En écrivant (14) comme

$$Eff_{P/R} = \frac{-(N-n)}{\rho_z^{-1} + (n-1)},$$

on constate que $Eff_{P/R}$ diminue effectivement en fonction de ρ_z. Voilà qui conclut la preuve.

3.4 Une structure de rechange parmi les perturbations

Enfin, supposons que la variance des perturbations à (15) est de la forme suivante :

$$\text{var}(\varepsilon_i) = c_1 X_i + c_2 X_i^2 \qquad (0 < c_1,\, c_2 \leq 1).$$

Voir Kott (1988). Pour ce cas, nous obtenons par analogie avec (16)

$$\rho_z = -\sum_{i \in U} \omega_i X_i,$$

où

$$\omega_i = \frac{1 + \phi X_i}{\sum_{i \in U}(1 + \phi X_i)}, \quad \text{et} \quad \phi = c_2/c_1$$

lorsque $\varphi = 0$, $\rho_z = -1/N$. Ainsi, lorsque $c_2 = 0$, l'efficacité de l'échantillonnage PPT est seulement équivalente à l'estimateur par quotient ordinaire de l'EAS. Dans la même veine que la preuve du Lemme 1, on peut démontrer que ρ_z diminue en fonction de φ, tandis que $Eff_{P/R}$ augmente en fonction de φ. Par conséquent, dans ce cas-ci, l'estimateur PPT aléatoire est toujours plus efficace que l'estimateur par quotient lorsque c_2 est positif.

4. Une application de l'Indice des prix à la production

L'Indice des prix à la production (IPP) des Pays-Bas est basé sur environ 2 500 indices de prix de marchandises, organisés par type de produit. L'indice de prix pour un produit précis peut être écrit comme suit

$$Y = \sum_{i \in U} X_i Z_i,$$

où Z_i est la variation des prix pour ce produit de l'établissement i par rapport à la période de base, tandis que X_i représente les ventes relatives de ce produit par l'établissement i pendant la période de base (rappelons que $\Sigma X_i = 1$).

Dans l'exemple donné ici, nous examinons les variations de prix de 70 établissements pour le produit *Métal de base* en décembre 2005 par rapport à décembre 2004 ; voir le tableau 1. Nous comparons la variance de l'estimateur par quotient d'un échantillon EAS à la variance de l'estimateur HT d'un échantillon PPT aléatoire lorsque $n = 9$. En appliquant (12) à ces données, on obtient $\text{var}(\hat{Y}_R) = 101$. Si l'échantillon avait été tiré avec remplacement, la variance aurait été de 116. En appliquant (3) et (9) pour un échantillon PPT aléatoire, on obtient $\text{var}(\hat{Y}_{\text{PPT},\gamma}) = 29{,}9$. Ce résultat tient compte de γ et est proche du résultat $V_{\text{PPT}}^{(sim)} = 29{,}2$ d'une expérience de simulation comprenant 80 000 échantillons PPT aléatoires de taille $n = 9$ à partir de l'ensemble de 70 établissements. Ainsi, $Eff_{P/R} = 3{,}5$. Étant donné que la formule (12) pour $\text{var}(\hat{Y}_R)$ est seulement asymptotiquement sans biais, nous avons également effectué des simulations pour évaluer l'erreur quadratique

moyenne (EQM) et le biais de $\hat{Y}_R$, pour obtenir $\mathrm{EQM}_R^{(sim)} = 108$ et un biais relativement petit de 0,7. Voilà qui confirme la conjecture que (12) donne une sous-estimation importante de la variance réelle ; voir Cochran (1977). Par conséquent, pour des échantillons modérés, la véritable valeur de $Eff_{P/R}$ pourrait être quelque peu plus élevée que (14) le suggère.

De plus, il convient de souligner que la formule simplifiée (10) pour ρ_z combinée à (3) donne presque le même résultat $\mathrm{var}(\hat{Y}_{\mathrm{PPT}}) = 30{,}7$, même si $N = 70$ n'est pas très grand. La variance PPT avec remplacement aurait été de 43,8. Par conséquent, la réduction de la variance pour l'échantillonnage PPT aléatoire est supérieure à 30 %, même si la fraction d'échantillonnage n/N est bien plus petite. Selon (18), la formule (12) où $n_{\mathrm{EAS}} = 26$ donne à peu près le même résultat que (3) où $n_{\mathrm{PPT}} = 9$; nota : $\rho_z = -0{,}042$. Par conséquent, la taille des échantillons diverge d'un facteur 2,9, ce qui correspond plus ou moins au facteur $(1 + CV_x^2) = 3{,}1$ de la sous-section 3.2. Il ne faut pas s'en étonner, parce que les fluctuations de prix et leur variabilité ne dépendent guère de la taille des entreprises. L'ajustement d'une régression bilogarithmique

$$\ln(Z_i - Y)^2 = \alpha + \beta \ln X_i + v_i \qquad (19)$$

donne lieu à l'estimation $\hat{\beta} = 0{,}07$ pour les données dans le tableau 1 ; les unités où $Z_i = Y$ devraient être omises dans la régression. L'estimation $\hat{\beta} = 0{,}07$ correspond à $\hat{\delta} = 2{,}07$ pour les perturbations dans (15), ce qui explique la supériorité de l'échantillonnage PPT aléatoire pour ce type de données. De plus, pour d'autres marchandises, $\hat{\delta}$ se situait souvent autour de 2 ; voir Enthoven (2007).

Tableau 1
Variations des prix (Z_i) et tailles (X_i) de 70 établissements

i	variation des prix	taille	*i*	variation des prix	taille
1	-18,4 %	0,0608	36	34,8 %	0,0427
2	-16,0 %	0,0784	37	13,1 %	0,0121
3	3,3 %	0,0762	38	31,7 %	0,0351
4	12,5 %	0,0100	39	-24,8 %	0,0074
5	0,0 %	0,0029	40	55,3 %	0,0009
6	8,3 %	0,0006	41	40,5 %	0,0066
7	-39,0 %	0,0182	42	34,6 %	0,0022
8	-25,1 %	0,0020	43	1,7 %	0,0001
9	1,1 %	0,0040	44	0,0 %	0,0039
10	4,4 %	0,0066	45	3,9 %	0,0304
11	-4,9 %	0,0039	46	25,4 %	0,0209
12	-8,9 %	0,0070	47	25,6 %	0,0062
13	-7,0 %	0,0148	48	0,0 %	0,0033
14	-15,0 %	0,0108	49	-0,3 %	0,0019
15	-10,7 %	0,0087	50	66,6 %	0,0346
16	-9,0 %	0,1079	51	0,0 %	0,0039
17	-11,3 %	0,0247	52	-2,9 %	0,0007
18	10,6 %	0,0024	53	15,8 %	0,0011
19	-23,2 %	0,0001	54	0,0 %	0,0026
20	-25,4 %	0,0001	55	0,0 %	0,0018
21	-80,7 %	0,0002	56	11,6 %	0,0057
22	13,4 %	0,0005	57	0,0 %	0,0042
23	-42,5 %	0,0010	58	0,0 %	0,0236
24	-34,8 %	0,0014	59	-1,5 %	0,0015
25	-30,0 %	0,0126	60	0,0 %	0,0003
26	8,0 %	0,0530	61	11,7 %	0,0067
27	0,0 %	0,0208	62	0,0 %	0,0012
28	2,1 %	0,0119	63	0,8 %	0,0040
29	11,3 %	0,0208	64	2,0 %	0,0009
30	0,7 %	0,0322	65	2,3 %	0,0018
31	9,5 %	0,0447	66	4,7 %	0,0026
32	11,5 %	0,0018	67	0,9 %	0,0064
33	5,8 %	0,0174	68	-1,0 %	0,0309
34	-6,9 %	0,0197	69	-0,5 %	0,0005
35	0,0 %	0,0124	70	0,0 %	0,0006

Nous concluons cette section par un exemple succinct démontrant que la PPT n'est pas *toujours* meilleure que l'estimateur par quotient. Bien que les données dans le tableau 2 pour une population de cinq unités soient artificielles, un modèle de données de ce genre pourrait survenir dans les succursales financières où de très petites entreprises financières peuvent grandir très vite à l'égard de certaines variables financières. Cette forte variabilité des taux de croissance des petites entreprises entraîne une faible valeur pour δ. Pour un échantillon EAS où $n = 2$ pour les cinq unités au tableau 2, la variance de l'estimateur par quotient est de 211 selon (12) ; les simulations donnent $\mathrm{EQM}_R^{(sim)} =$ 323. C'est beaucoup moins que la variance de 557 déterminée dans une simulation comportant 80 000 échantillons PPT aléatoires de taille $n = 2$. La formule (3) agencée à (9) donne le même résultat : 557. Ce serait également la bonne variance si l'échantillon avait été tiré selon Brewer (1963a) ou Durbin (1967). La formule (11), basée sur (10), donne une valeur légèrement différente : 556.

La régression (19) avec les données du tableau 2 donne $\hat{\beta} = -3,0$, et donc $\hat{\delta} = -1,0$. Conformément aux résultats de la sous-section 3.3, cette faible valeur $\hat{\delta} = -1,0$ explique pourquoi $\hat{Y}_{\mathrm{PPT}}$ est moins efficace que $\hat{Y}_R$ dans cet exemple. De plus, l'estimateur direct ordinaire $N\bar{y}_s$ d'un échantillon EAS a une variance de 356, qui est encore plus petite ici que la variance dans l'échantillonnage PPT aléatoire, sachant que $\bar{y}_s$ est la moyenne de l'échantillon de Y_i. Par conséquent, pour ce type de données, l'estimateur par quotient est la meilleure solution. Rappelons que l'estimateur par quotient a une plus petite variance que $N\bar{y}_s$ lorsque $b > Y/2X$, où b est la pente de la droite de régression de Y_i sur X_i et une constante $(i = 1, ..., N)$; voir Knottnerus (2003, page 117). Ainsi, les données Y_i $(= X_i Z_i)$ dans le tableau 2 ne présentent certainement pas une tendance stable.

Tableau 2
Taux de croissance des biens (Z_i) et taille (X_i) de cinq établissements

i	Taux de croissance	Taille
1	200 %	0,0455
2	33 %	0,1364
3	75 %	0,1818
4	33 %	0,2727
5	62 %	0,3636

5. Sommaire

Le présent document compare la variance de l'estimateur HT $\hat{Y}_{\mathrm{PPT}}$ à partir d'un échantillon PPT aléatoire avec la variance de l'estimateur classique par quotient $\hat{Y}_R$ à partir d'un échantillon EAS de la même taille. Dans cette comparaison, le coefficient d'autocorrélation de l'échantillonnage ρ_z joue un rôle capital.

Lorsque les tendances des données des variables x et z $(= y/x)$ sont telles que $\rho_z < -1/(N-1)$, on peut démontrer dans certaines conditions peu contraignantes que $\hat{Y}_{\mathrm{PPT}}$ est plus efficace que $\hat{Y}_R$ lorsque n et N sont assez grands, à condition que X_i et Z_i soient sans corrélation. En vertu du modèle (15) où $E(\varepsilon_i^2) = \sigma^2 X_i^{\delta}$, on constate que $\rho_z < -1/(N-1)$ lorsque $\delta > 1$. Par conséquent, pour ce type de données, il faut préconiser $\hat{Y}_{\mathrm{PPT}}$. De plus, il ressort de (14) et (16) que pour $\delta = 2$, l'efficacité relative de l'échantillonnage PPT comparativement à celle de l'estimateur par quotient augmente parallèlement au CV_x. En outre, il faut privilégier $\hat{Y}_R$ lorsque les données correspondent à un modèle où $\delta < 1$. Ces résultats sont confirmés de façon empirique avec une étude de simulation axée sur deux différents ensembles de données. Lorsque le modèle (15) ne s'applique pas, l'efficacité relative de $\hat{Y}_{\mathrm{PPT}}$ est donnée par (14), à condition que n soit grand et que N soit relativement plus grand. En pratique, le ρ_z inconnu dans (14) est remplacé par $\hat{\rho}_{z9}$. Le fait que $n << N$ ne signifie pas nécessairement que le facteur $(n-1)\rho_z$ dans (3) est toujours négligeable.

Remerciements

Les opinions exprimées dans l'article sont celles de l'auteur et ne correspondent pas nécessairement à la politique du bureau central de la statistique des Pays-Bas. L'auteur tient à remercier Peter-Paul de Wolf, Sander Scholtus, le rédacteur en chef adjoint et deux réviseurs anonymes pour leurs suggestions et leurs corrections utiles.

Annexe A

Contre-exemple

Les équations (5) et (7) ne peuvent pas toujours être utilisées pour l'échantillonnage PPT aléatoire lorsque n et N sont du même ordre, tandis que X_i et Z_i sont corrélés. Pour le constater, supposons une population U composée de deux groupes, U_1 et U_2, avec des moyennes de $\bar{Y}_1$ et $\bar{Y}_2$, respectivement. Les deux tailles de strates sont $N/2$. Supposons que s est un échantillon PPT aléatoire de taille $n = 3N/4$ tiré de l'ensemble de la population U. Supposons que X_i est tel que

$$\pi_i = nX_i = \begin{cases} 1 & \text{si} \quad i \in U_1 \\ 0,5 & \text{si} \quad i \in U_2. \end{cases}$$

De toute évidence, le groupe 1 ne contribue pas à la variance. Les unités sélectionnées en s de U_2 constituent un

échantillon EAS ordinaire de taille $N/4$. Par conséquent, pour l'échantillonnage PPT aléatoire, la bonne formule de variance dans cet exemple est la suivante

$$\text{var}(\hat{Y}_{\text{PPT}}) = \left(\frac{N}{2}\right)^2 \left(1 - \frac{1}{2}\right) \frac{S_{y2}^2}{N/4} = \frac{NS_{y2}^2}{2},$$

et

$$S_{y2}^2 = \frac{2}{N-2} \sum_{i \in U_2} (Y_{2i} - \bar{Y}_2)^2.$$

Cependant, l'approximation (11) donne un résultat complètement différent et plus grand, à moins que $\bar{Y}_1 = 2\bar{Y}_2$.

Bibliographie

Brewer, K.R.W. (1963a). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5, 5-13.

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

Brewer, K.R.W., et Donadio, M.E. (2003). La variance sous grande entropie de l'estimateur de Horvitz-Thompson. *Techniques d'enquête*, 29, 213-220.

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

Connor, W.S. (1966). An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement. *Journal of the American Statistical Association*, 61, 384-390.

Cumberland, W.G., et Royall, R.M. (1981). Prediction models and unequal probability sampling. *Journal of the Royal Statistical Society*, B, 43, 353-367.

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 219-230.

Durbin, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Applied Statistics*, 16, 152-164.

Enthoven, L. (2007). *Cohort calculations* (en néerlandais). Rapport MIC-2007-21, Statistics Netherlands, Voorburg.

Foreman, E.K., et Brewer, K.R.W. (1971). The efficient use of supplementary information in standard sampling procedures. *Journal of the Royal Statistical Society*, B, 33, 391-400.

Gabler, S. (1984). On unequal probability sampling: Sufficient conditions for the superiority of sampling without replacement. *Biometrika*, 71, 171-175.

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

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

Hidiroglou, M.A., et Gray, G.B. (1980). Construction of joint probability of selection for systematic P.P.S. sampling. *Applied Statistics*, 29, 107-112.

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.

Knottnerus, P. (2003). *Sample Survey Theory*: *Some Pythagorean Perspectives.* New York : Springer-Verlag.

Kott, P.S. (1988). Model-based finite population correction for the Horvitz-Thompson estimator. *Biometrika*, 75, 797-799.

Kott, P.S. (2005). A note on the Hartley-Rao variance estimator. *Journal of Official Statistics*, 21, 433-439.

Qualité, L. (2008). A comparison of conditional Poisson sampling versus unequal probability sampling with replacement. *Journal of Statistical Planning and Inference*, 138, 1428-1432.

Rosén, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62, 159-191.

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

Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5, 119-127.

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

Thompson, M.E., et Wu, C. (2008). Échantillonnage PPT systématique randomisé basé sur la simulation en cas de substitution d'unités. *Techniques d'enquête*, 34, 3-11.

Yates, F., et Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, B, 15, 253-261.

Techniques d'enquête, juin 2011
Vol. 37, N° 1, pp. 111-116
Statistique Canada, N° 12-001-X au catalogue

Utilisation des équations estimantes pour réaliser un calage sur des paramètres complexes

Éric Lesage [1]

Résumé

Dans la méthode du calage de Deville et Särndal (1992), les équations de calage ne prennent en compte que les estimations exactes de totaux des variables auxiliaires. L'objectif de cet article est de s'intéresser à d'autres paramètres que le total pour caler. Ces paramètres que l'on qualifie de complexes sont par exemple le ratio, la médiane ou la variance de variables auxiliaires.

Mots clés : Calage ; paramètre complexe ; équation estimante ; poids de calage.

1. Introduction

En statistique d'enquête, à la phase d'estimation, deux grandes approches d'estimateurs sont utilisées : d'une part les estimateurs dits « assistés par un modèle » (comme l'estimateur par la régression ou l'estimateur par le ratio) et d'autre part les estimateurs par calage (comme le raking ratio) de Deville et Särndal (1992). Ces deux approches ont d'ailleurs une certaine proximité comme le montre l'estimateur par la régression qui correspond à l'estimateur par calage avec la distance du χ^2 (méthode de calage dite « linéaire »).

L'objet de cet article est d'étendre la famille des estimateurs par calage. Actuellement la méthode du calage permet de caler sur des totaux. L'idée est de pouvoir prendre en compte dans les contraintes de calage des statistiques ou paramètres complexes, comme un ratio, une médiane ou une moyenne géométrique. Ce qui motive cette démarche c'est qu'il est possible qu'on dispose d'une information auxiliaire sous la forme d'une statistique complexe et non sous la forme de totaux. Par exemple, on connaîtrait un ratio sur la population totale mais pas le total du numérateur ou du dénominateur.

D'autres auteurs ont déjà abordé la question des paramètres complexes dans le cadre du calage. Särndal (2007) passe en revue un certain nombre d'entre eux. Il cite notamment les travaux de Harms et Duchesne (2006) sur l'estimation de quantiles par calage, et ceux de Plikusas (2006) et Krapavickaite et Plikusas (2005) sur les estimateurs par calage de certaines fonctions de totaux.

L'originalité de l'approche de cet article est de ramener le calage sur un paramètre complexe à un calage sur un total pour une nouvelle variable auxiliaire *ad hoc*. L'avantage de cette approche est qu'elle permet d'utiliser les outils actuels de calage et de ne pas avoir à résoudre un programme d'optimisation complexe.

Dans la deuxième partie de cet article, nous rappellerons le fonctionnement de la méthode du calage, définirons le calage sur paramètres complexes et présenterons des cas simples où le calage sur un paramètre complexe peut se ramener à un calage sur un total. Dans la troisième partie, nous nous intéresserons aux paramètres qui peuvent se définir comme solution d'une équation estimante (Godambe et Thompson 1986). Nous introduirons la notion de calage sur paramètre complexe défini par une équation estimante et montrerons que l'équation de calage résultante peut être remplacée par une équation de calage sur un total.

2. Paramètre complexe défini comme une fonction de totaux

2.1 Rappels sur le calage sur des totaux

Soit U, une population finie de taille N. Les unités statistiques de cette population sont indexées par une étiquette k où $k \in \{1, \ldots, N\}$. Un échantillon s est tiré suivant un plan de sondage $p(s)$. Sa taille est notée n, elle peut être aléatoire. On note π_k la probabilité d'inclusion de k dans l'échantillon s et $d_k = 1/\pi_k$ son poids d'échantillonnage.

Pour n'importe quelle variable z qui prend les valeurs z_k pour les unités de U indexées par k, on appelle total de z sur U la somme $t_z = \sum_{k \in U} z_k$.

Soient $y^{(1)}, \ldots, y^{(Q)}, Q$ variables d'intérêt, dont les valeurs sont connues uniquement sur l'échantillon s et $\theta_{\mathbf{y}}$, le paramètre d'intérêt qui est une fonction des totaux $t_{y^{(1)}}, \ldots, t_{y^{(Q)}}$:

$$\theta_{\mathbf{y}} = f(t_{y^{(1)}}, \ldots, t_{y^{(Q)}}).$$

L'estimateur de $\theta_{\mathbf{y}}$ est :

$$\hat{\theta}_{\mathbf{y},\pi} = f(\hat{t}_{y^{(1)},\pi}, \ldots, \hat{t}_{y^{(Q)},\pi}).$$

1. Éric Lesage, CREST(ENSAI) et IRMAR(UEB), Campus de Ker Lann, F-35172 BRUZ, France. Courriel : eric.lesage@ensai.fr.

Il s'agit simplement de la fonction $f(\cdot, ..., \cdot)$ dans laquelle les totaux $t_{y^{(q)}}$ ont été remplacés par leur estimateur Horvitz-Thompson $\hat{t}_{y^{(q)},\pi} = \sum_{k \in s} d_k y_k^{(q)}$ (Särndal, Swensson and Wretman 1992). On peut qualifier cet estimateur d'estimateur par substitution.

Soient $x^{(1)}, ..., x^{(P)}$, P variables auxiliaires, connues sur s et $t_{x^{(1)}}, ..., t_{x^{(P)}}$ les totaux sur U de ces variables auxiliaires qui sont également connus. Pour un individu k, on note $\mathbf{x}_k' = (x_k^{(1)}, ..., x_k^{(P)})$ le vecteur des valeurs prises par les variables auxiliaires sur k.

L'estimateur par calage de $\theta_{\mathbf{y}}$ est :

$$\hat{\theta}_{\mathbf{y},\mathrm{CAL}} = f(\hat{t}_{y^{(1)},\mathrm{CAL}}, \ldots, \hat{t}_{y^{(Q)},\mathrm{CAL}})$$

avec $\hat{t}_{y^{(q)},\mathrm{CAL}} = \sum_{k \in s} w_k y_k^{(q)}$, et une série de poids $\{w_k\}_{(k \in s)}$, dits poids de calage (que l'on devrait noter $w_k(s)$, puisque ces poids dépendent de l'échantillonnage), obtenus par résolution du programme d'optimisation suivant :

$$\min_{\{w_k\}_{(k \in s)}} \sum_{k \in s} d(w_k, d_k)$$

sous contraintes :

$$\begin{cases} \hat{t}_{x^{(1)},\mathrm{CAL}} = t_{x^{(1)}} \\ \ldots \\ \hat{t}_{x^{(P)},\mathrm{CAL}} = t_{x^{(P)}} \end{cases}$$

$d(\cdot,\cdot)$ est une pseudo-distance, *i.e.*, une fonction qui mesure l'écart du poids de calage au poids d'échantillonnage (à la différence d'une distance, la pseudo-distance n'est pas nécessairement symétrique sur ses deux arguments). Le programme se résout à l'aide d'un Lagrangien. Dans le cas où la distance utilisée est la distance du χ^2 (*i.e.*, $d(w_k, d_k) = (1/2)(w_k - d_k)^2 / d_k$), on trouve comme solution : $w_k = d_k(1 + \mathbf{x}_k' \lambda)$ (où λ est le vecteur de taille P des multiplicateurs de Lagrange).

2.2 Calage sur un paramètre complexe $\eta_{\mathbf{x}}$

Définition 1 : *Soient $x^{(1)}, \ldots, x^{(P)}$, P variables auxiliaires, connues sur s et $\eta_{\mathbf{x}} = g(t_{x^{(1)}}, ..., t_{x^{(P)}})$ un paramètre complexe, fonction des totaux de ces variables auxiliaires, également connu.*

Dans le cas du calage sur le paramètre complexe $\eta_{\mathbf{x}}$, les poids de calage sont obtenus par résolution du programme d'optimisation suivant :

$$\min_{\{w_k\}_{(k \in s)}} \sum_{k \in s} d(w_k, d_k)$$

sous contraintes :

$$\hat{\eta}_{\mathbf{x},\mathrm{CAL}} = g(\hat{t}_{x^{(1)},\mathrm{CAL}}, \ldots, \hat{t}_{x^{(P)},\mathrm{CAL}}) = \eta_{\mathbf{x}}.$$

Les totaux $t_{x^{(q)}}$ n'ont pas besoin d'être connus, seul le paramètre complexe $\eta_{\mathbf{x}}$ doit l'être.

Prenons l'exemple du ratio

$$R_{\mathbf{x}} = \frac{t_{x^{(1)}}}{t_{x^{(2)}}} = \frac{\sum_{k \in U} x_k^{(1)}}{\sum_{k \in U} x_k^{(2)}}.$$

L'estimateur par calage de $R_{\mathbf{x}}$ est de la forme :

$$\hat{R}_{\mathbf{x},\mathrm{CAL}} = \frac{\sum_{k \in s} w_k x_k^{(1)}}{\sum_{k \in s} w_k x_k^{(2)}}.$$

L'équation de calage dans le cas d'un calage sur le ratio s'écrit :

$$\hat{R}_{\mathbf{x},\mathrm{CAL}} = \frac{\sum_{k \in s} w_k x_k^{(1)}}{\sum_{k \in s} w_k x_k^{(2)}} = R_{\mathbf{x}}$$

$R_{\mathbf{x}}$ est une information auxiliaire connue, comme l'est habituellement le total des variables auxiliaires. Ce cas de figure peut se produire lorsqu'on a des proportions bien connues et qui par exemple sont stables dans le temps, mais qu'on ne connaît pas bien les totaux du dénominateur et du numérateur.

On a présenté le cas du calage sur un seul paramètre complexe, mais on voit qu'il est facile de caler sur plusieurs paramètres complexes. Il y a alors autant de contraintes que de paramètres de calage.

2.3 Cas simples où le calage sur paramètre complexe peut se ramener au calage sur un total

Il n'est pas évident de savoir d'emblée si l'équation de calage sur un paramètre complexe va pouvoir s'écrire sous la forme d'une équation de calage sur un total. En d'autres termes, il n'est pas toujours trivial de trouver une « nouvelle » variable auxiliaire z, associée au paramètre complexe, sur le total de laquelle on va caler.

Par exemple, cela est assez immédiat pour tous les moments d'une variable auxiliaire x (on suppose ici que le plan de sondage permet d'estimer parfaitement la taille de la population N). Si $\mu_{x^m} = N^{-1} \sum_{k \in U} x_k^m$ est une information auxiliaire, alors il suffit de prendre $z_k = x_k^m / N$ et de caler sur $\mu_{x^m} : \sum_{k \in s} w_k \, x_k^m / N = \mu_{x^m}$.

Si on veut caler sur la variance et la moyenne de la variable x en ayant μ_x et σ_x^2 comme information auxiliaire, alors on peut utiliser les deux nouvelles variables auxiliaires :

$$z_k^{(1)} = \frac{x_k}{N}$$

et

$$z_k^{(2)} = \frac{(x_k - \mu_x)^2}{N}.$$

En revanche, si on ne connaît pas μ_x, mais qu'on a σ_x^2 dans l'information auxiliaire et qu'on veut caler sur cette variance, alors les choses se compliquent. Pour s'en convaincre, il faut écrire l'estimateur par substitution de σ_x^2

(dans le cas où le plan de sondage permet d'estimer parfaitement la taille de la population N) :

$$\hat{\sigma}^2_{x,\mathrm{CAL}} = \frac{1}{N}\sum_{k\in s} w_k \left(x_k - \left(\frac{\sum_{l\in s} w_l x_l}{N} \right) \right)^2 .$$

Trouver une nouvelle variable auxiliaire z n'est pas immédiat, dans la mesure où l'équation de calage initiale n'est pas linéaire par rapport au vecteur des poids. Nous reviendrons sur le cas de la variance à la section 3.3 de cet article.

Exemple du Ratio

Proposition 1 : *Le calage sur un ratio est équivalent au calage sur le total de la nouvelle variable auxiliaire* : $z_k = x_k^{(1)} - R_{\mathbf{x}} x_k^{(2)}$.

L'équation de calage s'écrit :

$$\hat{t}_{z,\mathrm{CAL}} = t_z = 0.$$

Démonstration :

$$\hat{t}_{z,\mathrm{CAL}} = t_z$$

$$\Leftrightarrow \sum_{k\in s} w_k (x_k^{(1)} - R_{\mathbf{x}} x_k^{(2)}) = \sum_{k\in U} (x_k^{(1)} - R_{\mathbf{x}} x_k^{(2)})$$

$$\Leftrightarrow \hat{t}_{x^{(1)},\mathrm{CAL}} - R_{\mathbf{x}}\, \hat{t}_{x^{(2)},\mathrm{CAL}} = t_{x^{(1)},\mathrm{CAL}} - R_{\mathbf{x}}\, t_{x^{(2)},\mathrm{CAL}} = 0$$

$$\Leftrightarrow \frac{\hat{t}_{x^{(1)},\mathrm{CAL}}}{\hat{t}_{x^{(2)},\mathrm{CAL}}} = R_{\mathbf{x}}$$

i.e., $\hat{R}_{\mathbf{x},\mathrm{CAL}} = R_{\mathbf{x}}$.

Fonction d'un ratio de combinaisons linéaires de totaux

Soit $\eta_{\mathbf{x}}$ un paramètre complexe qui est une fonction bijective d'un ratio de combinaisons linéaires de totaux :

$$\eta_{\mathbf{x}} = h\left(\frac{\alpha' \cdot \mathbf{t}_{\mathbf{x}}}{\beta' \cdot \mathbf{t}_{\mathbf{x}}} \right) \qquad (1)$$

avec $\alpha' = (\alpha_1, ..., \alpha_P)$ et $\beta' = (\beta_1, ..., \beta_P)$ des vecteurs de coefficients réels de taille P, et $\mathbf{t}'_{\mathbf{x}} = (t_{x^{(1)}}, \ldots, t_{x^{(P)}})$.

Proposition 2 : *Réaliser un calage sur le paramètre complexe* η_x *défini par la fonction* (1) *est équivalent à caler sur le total de la nouvelle variable auxiliaire* :

$$z_k = (\alpha' - h^{-1}(\eta_{\mathbf{x}})\beta') \cdot \mathbf{x_k}$$

avec l'équation de calage :

$$\hat{t}_{z,\mathrm{CAL}} = \sum_{k\in s} w_k z_k = t_z = 0.$$

Démonstration :

$$\hat{\eta}_{\mathbf{x},\mathrm{CAL}} = \eta_{\mathbf{x}} \Leftrightarrow h\left(\frac{\alpha' \cdot \hat{\mathbf{t}}_{\mathbf{x,CAL}}}{\beta' \cdot \hat{\mathbf{t}}_{\mathbf{x,CAL}}} \right) = \eta_{\mathbf{x}}$$

$$\Leftrightarrow \frac{\alpha' \cdot \hat{\mathbf{t}}_{\mathbf{x,CAL}}}{\beta' \cdot \hat{\mathbf{t}}_{\mathbf{x,CAL}}} = h^{-1}(\eta_{\mathbf{x}})$$

$$\Leftrightarrow (\alpha' - h^{-1}(\eta_{\mathbf{x}})\, \beta') \cdot \hat{\mathbf{t}}_{\mathbf{x,CAL}} = 0$$

$$\Leftrightarrow \sum_{k\in s} w_k (\alpha' - h^{-1}(\eta_{\mathbf{x}})\, \beta') \cdot \mathbf{x_k} = 0.$$

On peut prendre la moyenne géométrique comme exemple :

$$\mu_{\mathrm{Geo},\mathbf{x}} = \left(\prod_{k\in U} x_k \right)^{1/N} .$$

On peut réécrire cette expression :

$$\mu_{\mathrm{Geo},\mathbf{x}} = \exp\left(\frac{\sum_{k\in U} \ln(x_k)}{\sum_{k\in U} 1} \right).$$

On note : $\mathbf{x}'_k = (x_k^{(1)}, x_k^{(2)}) = (\ln(x_k), 1)$, $\alpha' = (1, 0)$, $\beta' = (0, 1)$ et $h^{-1}(u) = \exp^{-1}(u) = \ln(u)$.

On a alors comme nouvelle variable auxiliaire :

$$z_k = \ln(x_k) - \ln(\mu_{\mathrm{Geo},\mathbf{x}}) \cdot 1.$$

Nous allons voir dans la suite de cet article que la méthode des équations estimantes offre une autre approche pour exhiber la (les) nouvelle(s) variable(s) auxiliaire(s) **z**.

3. Paramètre défini par une équation estimante

3.1 Principe d'estimation par équation estimante

Certains paramètres $\theta_{\mathbf{y}}$ se définissent, ou peuvent se définir, comme solution d'une fonction implicite appelée *équation estimante sur U* (Godambe et Thompson 1986), *i.e.* :

$$\sum_{k\in U} \Phi(\theta_{\mathbf{y}}, \mathbf{y}_k) = 0$$

avec $\mathbf{y}'_k = (y_k^{(1)}, ..., y_k^{(Q)})$ le vecteur des valeurs prises par les variables d'intérêt pour l'individu k.

Dans ce contexte, on définit un estimateur de $\theta_{\mathbf{y}}$ à partir de l'échantillon s, noté $\hat{\theta}_{\mathbf{y},ee,\pi}$, qui est la solution de *l'équation estimante sur s* (voir, entre autres, Hidiroglou, Rao et Yung 2002) :

$$\sum_{k\in s} d_k \Phi(\hat{\theta}_{\mathbf{y},ee,\pi}, \mathbf{y}_k) = 0.$$

Tableau 1
Exemples de paramètres définis par des équations estimantes sur *U*

Paramètres	$\Phi(\theta_{\mathbf{y}}, \mathbf{y}_k)$	Équations estimantes sur *U*
moyenne μ	$(\mathbf{y}_k - \mu)$	$\sum_{k\in U} (y_k - \mu) = 0$
ratio $R = \mu_1 / \mu_2$	$(y_k^{(1)} - R\, y_k^{(2)})$	$\sum_{k\in U} (y_k^{(1)} - R y_k^{(2)}) = 0$
médiane m	$(1_{y_k \le m} - 1/2)$	$\sum_{k\in U} (1_{y_k \le m} - 1/2) = 0$

On peut donner également l'exemple du coefficient d'une régression logistique. Soient $y^{(1)}$ une variable dichotomique qui prend les valeurs 0 et 1 sur *U*, et $y^{(2)}$ une variable quantitative. On suppose que la valeur $y_k^{(1)}$ prises par $y^{(1)}$ pour l'unité *k* est une réalisation de la variable aléatoire $Y_k^{(1)}$ qui suit une loi de Bernoulli

$$\mathfrak{B}\left(1,\ p_k = \frac{1}{1+\exp(-\beta_0 y_k^{(2)})}\right).$$

Nous avons limité le nombre de paramètres à un, mais il serait aussi simple de considérer le cas multidimensionnel. Toutefois, il faudrait donner une définition des équations estimantes qui prennent en compte le cas des paramètres vectoriels.

Le paramètre qui nous intéresse est l'estimateur de β_0, noté β, calculé sur la population finie par la méthode du maximum de vraisemblance. L'équation estimante de β sur *U* sera l'équation du maximum de vraisemblance. La log vraisemblance dans le cas de variables de Bernoulli vaut :

$$\mathbf{L}(\beta) = \sum_{k\in U} y_k^{(1)} \ln(p_k) + \sum_{k\in U} (1 - y_k^{(1)}) \ln(1 - p_k).$$

On obtient facilement l'équation estimante de β sur *U* :

$$\sum_{k\in U} y_k^{(2)} \left(y_k^{(1)} - \frac{1}{1+\exp(-\beta y_k^{(2)})} \right) = 0.$$

L'équation estimante sur *s* qui définit l'estimateur $\hat{\beta}_{ee,\pi}$ à partir des poids d'échantillonnage est :

$$\sum_{k\in s} d_k y_k^{(2)} \left(y_k^{(1)} - \frac{1}{1+\exp(-\hat{\beta}_{ee,\pi} y_k^{(2)})} \right) = 0.$$

L'équation estimante n'est pas linéaire dans le paramètre ; $\hat{\beta}_{ee,\pi}$ ne peut pas être exprimé comme une fonction simple des observations.

L'exemple de la régression logistique est très intéressant car il met en évidence qu'on n'a pas besoin de connaître l'expression de $\hat{\beta}_{ee,\pi}$ pour faire le calage. On verra dans la sous-section suivante qu'il suffit d'avoir le terme générique de l'équation estimante sur

$$U,\ \Phi(\beta, \mathbf{y}_k) = y_k^{(2)} \left(y_k^{(1)} - \frac{1}{1+\exp(-\beta y_k^{(2)})} \right),$$

pour tout les $k \in s$.

3.2 Calage dans le cas de paramètres définis par des équations estimantes

Soient $\mathbf{x}_k' = (x^{(1)}, ..., x^{(P)})$, le vecteur des *P* variables auxiliaires connues sur *s* et $\eta_{\mathbf{x}}$ un paramètre complexe, également connu, défini par l'équation estimante :

$$\sum_{k\in U} \Psi(\eta_{\mathbf{x}}, \mathbf{x}_k) = 0.$$

Définition 2 : *Dans le cas du calage sur le paramètre complexe* $\eta_{\mathbf{x}}$, *les poids de calage sont obtenus par résolution du programme d'optimisation suivant* :

$$\min_{\{w_k\}_{(k\in s)}} \sum_{k\in s} d(w_k, d_k)$$

sous contraintes :

$$\sum_{k\in s} w_k \Psi(\eta_{\mathbf{x}}, \mathbf{x}_k) = 0.$$

Proposition 3 : *Le calage sur un paramètre complexe* $\eta_{\mathbf{x}}$, *défini par une équation estimante, est équivalent à un calage sur le total de la nouvelle variable auxiliaire* : $z_k = \Psi(\eta_{\mathbf{x}}, \mathbf{x}_k)$, *avec la contrainte de calage* $\sum_{k\in s} w_k z_k = 0$.

Définition 3 : *Un estimateur par calage du paramètre d'intérêt* $\theta_{\mathbf{y}}$, *noté* $\hat{\theta}_{\mathbf{y},ee,\mathrm{CAL}}$, *est solution de l'équation estimante sur s pondérée par les poids de calage* $\{w_k\}_{(k\in s)}$:

$$\sum_{k\in s} w_k \Phi(\hat{\theta}_{\mathbf{y},ee,\mathrm{CAL}}, \mathbf{y}_k) = 0.$$

Dans la plupart des cas la solution de l'équation estimante est unique. La médiane est un exemple de paramètre qui peut offrir plusieurs solutions. On prend alors souvent l'infimum comme estimateur.

Proposition 4 : *S'il y a unicité de la solution de l'équation* : $\sum_{k\in s} w_k \Psi(\hat{\eta}_{\mathbf{x},ee,\mathrm{CAL}}, \mathbf{x}_k) = 0$, *alors*

$$\hat{\eta}_{\mathbf{x},ee,\mathrm{CAL}} = \eta_{\mathbf{x}}.$$

Démonstration : $\eta_{\mathbf{x}}$ *est solution de l'équation estimante qui définit* $\hat{\eta}_{\mathbf{x},ee,\mathrm{CAL}}$. *Puisqu'il y a une solution unique, on a* $\hat{\eta}_{\mathbf{x},ee,\mathrm{CAL}} = \eta_{\mathbf{x}}$.

3.3 Calage sur une variance

L'objet de cette section est d'étudier le calage sur la variance σ_x^2 qui est un paramètre complexe plus compliqué que ceux vus précédemment. On montrera que lorsqu'on a uniquement ce paramètre comme information auxiliaire, on peut procéder à un calage approché qui fournit des poids de calage qui ont de meilleures propriétés que les poids d'échantillonnage.

Revenons donc sur le cas de la variance. La moyenne μ_x et la variance σ_x^2 sur *U* de la variable auxiliaire *x* peuvent être définies par deux équations estimantes sur *U* :

$$\begin{cases} \sum_{k\in U}(x_k - \mu_x) = 0 & (2) \\ \sum_{k\in U}((x_k - \mu_x)^2 - \sigma_x^2) = 0. & (3) \end{cases}$$

Si on connaît ces deux paramètres, le calage sur ceux-ci est facile, puisqu'il suffit de caler sur les totaux des deux nouvelles variables auxiliaires $z^{(1)} = x - \mu_x$ et $z^{(2)} = (x - \mu_x)^2 - \sigma_x^2$.

Par contre, si on considère le cas d'école où la moyenne μ_x n'est pas connue, alors le paramètre σ_x^2 ne peut pas être défini par une équation estimante unique. En effet, si on remplace μ_x par sa définition explicite

$$\mu_x = \frac{\sum_{l\in U} x_l}{\sum_{j\in U} 1}$$

dans l'équation (3), on obtient l'équation

$$\sum_{k\in U}\left(\left(x_k - \frac{\sum_{l\in U} x_l}{\sum_{j\in U} 1}\right)^2 - \sigma_x^2\right) = 0,$$

qui ne peut pas s'écrire sous la forme d'une équation estimante : $\sum_{k\in U}\Psi(\sigma_x^2, x_k) = 0$.

μ_x apparaît donc comme un paramètre de nuisance (Binder 1991). Pour contourner cette difficulté, on peut le remplacer dans l'équation (3) par son estimateur par substitution : $\hat{\mu}_{x,\pi} = \hat{t}_{x,\pi}/\hat{N}_\pi$, avec $\hat{N}_\pi = \sum_{k\in s} d_k 1$ l'estimateur Horvitz-Thompson de la taille de la population U. On obtient alors comme équation de calage « approchée » :

$$\sum_{k\in s} w_k\left(\left(x_k - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}\right)^2 - \sigma_x^2\right) = 0. \quad (4)$$

Proposition 5 : *Avec l'équation estimante* (4), *le calage sur la variance n'est pas parfait et nous avons* :

$$\hat{\sigma}^2_{x,ee,\mathrm{CAL}} = \sigma_x^2 - \left(\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} - \frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}}\right)^2. \quad (5)$$

Démonstration :

- L'équation de calage « approchée » est l'équation (4).
- La définition des estimateurs par calage des paramètres :

$$\begin{cases} \sum_{k\in s} w_k(x_k - \hat{\mu}_{x,ee,\mathrm{CAL}}) = 0 \\ \sum_{k\in s} w_k((x_k - \hat{\mu}_{x,ee,\mathrm{CAL}})^2 - \hat{\sigma}^2_{x,ee,\mathrm{CAL}}) = 0. \end{cases}$$

Ce qui peut se réécrire :

$$\begin{cases} \hat{\mu}_{x,ee,\mathrm{CAL}} = \dfrac{\sum_{k\in s} w_k x_k}{\sum_{k\in s} w_k} = \dfrac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}} \\ \sum_{k\in s} w_k\left(\left(x_k - \dfrac{\hat{t}_{x,CAL}}{\hat{N}_{\mathrm{CAL}}}\right)^2 - \hat{\sigma}^2_{x,ee,\mathrm{CAL}}\right) = 0. \end{cases}$$

- Si on fait la différence terme à terme de l'équation de calage approchée et de la deuxième équation estimante, on obtient :

$$\sum_{k\in s} w_k\left(\left(x_k - \frac{\hat{t}_{x,\pi}}{\hat{N}}\right)^2 - \left(x_k - \frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}}\right)^2 - \sigma_x^2 + \hat{\sigma}^2_{x,ee,\mathrm{CAL}}\right) = 0.$$

En utilisant l'identité : $a^2 - b^2 = (a-b)(a+b)$, on a :

$$\sum_{k\in s} w_k\left(\left(\frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}} - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}\right)\left(2x_k - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} - \frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}}\right)\right) - \hat{N}_{\mathrm{CAL}}(\sigma_x^2 - \hat{\sigma}^2_{x,ee,\mathrm{CAL}}) = 0$$

$$\left(\frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}} - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}\right)\sum_{k\in s} w_k\left(2x_k - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} - \frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}}\right) - \hat{N}_{\mathrm{CAL}}(\sigma_x^2 - \hat{\sigma}^2_{x,ee,\mathrm{CAL}}) = 0$$

$$\left(\frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}} - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}\right)\left(2\hat{t}_{x,\mathrm{CAL}} - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}\hat{N}_{\mathrm{CAL}} - \hat{t}_{x,\mathrm{CAL}}\right) - \hat{N}_{\mathrm{CAL}}(\sigma_x^2 - \hat{\sigma}^2_{x,ee,\mathrm{CAL}}) = 0$$

$$\hat{N}_{\mathrm{CAL}}\left(\frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}} - \frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}\right)^2 - \hat{N}_{\mathrm{CAL}}(\sigma_x^2 - \hat{\sigma}^2_{x,ee,\mathrm{CAL}}) = 0.$$

Ce qui correspond bien à l'expression de $\hat{\sigma}^2_{x,ee,\mathrm{CAL}}$ donné par l'équation (5).

Ce résultat est intéressant car, à défaut d'un calage exact, on a un estimateur par calage de σ_x^2 qui est asymptotiquement plus précis que l'estimateur par substitution $\hat{\sigma}^2_{x,\pi}$. En effet, si on se place dans le cadre asymptotique habituellement utilisé en sondage et qu'on utilise la linéarisation des estimateurs complexes (Deville 1999), on a :

$$\hat{\sigma}^2_{x,\pi} - \sigma_x^2 = O_p\left(\frac{1}{\sqrt{n}}\right)$$

et

$$(\hat{\sigma}^2_{x,ee,\mathrm{CAL}} - \sigma_x^2)^{1/2} = \left(\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi} - \frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}}\right) = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Ce qui donne :

$$\hat{\sigma}^2_{x,ee,\mathrm{CAL}} - \sigma^2_x = O_p\left(\frac{1}{n}\right).$$

4. Conclusion

Cet article a présenté une méthode simple pour réaliser un calage dans le cas où l'information auxiliaire prend la forme d'un paramètre complexe. Cette méthode s'appuie sur la notion d'équation estimante. Son gros avantage est qu'elle peut être mise en oeuvre avec les logiciels actuels de calage.

Une piste de recherche intéressante serait d'étudier dans quels cas pratiques l'utilisation de paramètres complexes dans le calage améliore la précision des paramètres d'intérêt.

Remerciements

L'auteur remercie le rédacteur associé de la revue, les examinateurs, Guillaume Chauvet, François Coquet et Jean-Claude Deville pour leurs commentaires constructifs sur les versions provisoires de cette note.

Bibliographie

Binder, D.A. (1991). Use of estimating functions for interval estimation from complex surveys. *Proceedings of the survey research methods section*, American Statistical Association, 34-42.

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 219-230.

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

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

Harms, T., et Duchesne, P. (2006). De l'estimation des quantiles par calage. *Techniques d'enquête*, 32, 41-57.

Hidiroglou, M., Rao, J.N.K. et Yung, W. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā*, 64, 2, 364-378.

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

Plikusas, A. (2006). Non-linear calibration. Recueil du Colloque sur les méthodes de sondage, Venspils, Latvia. Riga : Central Statistical Bureau of Latvia.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New-York : Springer-Verlag, 162-163.

Särndal, C.-E. (2007). La méthode de calage dans la théorie et la pratique des enquêtes. *Techniques d'enquête*, 33, 2, 113-135.

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 26, No. 4, 2010

Editorial
Ingegerd Jansson and Boris Lorenc i

Degrees of Freedom Approximations and Rules-of-Thumb
Richard Valliant and Keith F. Rust 585

Combining Link-Tracing Sampling and Cluster Sampling to Estimate Totals and Means of Hidden Human Populations
Martín H. Félix-Medina and Pedro E. Monjardin 603

Increasing Respondents' Use of Definitions in Web Surveys
Andy Peytchev, Frederick G. Conrad, Mick P. Couper and Roger Tourangeau 633

A Framework for Cut-off Sampling in Business Survey Design
Roberto Benedetti, Marco Bee and Giuseppe Espa 651

Statistical Model of the 2001 Czech Census for Interactive Presentation
Jiří Grim, Jan Hora, Pavel Boček, Petr Somol and Pavel Pudil 673

An Optimal Multivariate Stratified Sampling Design Using Auxiliary Information: An Integer Solution Using Goal Programming Approach
M.G.M. Khan, T. Maiti and M.J. Ahsan 695

Letter to the Editor
Rainer Lenz 709

Book Reviews 711
Editorial Collaborators 717

All inquires about submissions and subscriptions should be directed to jos@scb.se

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 27, No. 1, 2011

The 2010 Morris Hansen Lecture Dealing with Survey Nonresponse in Data Collection, in Estimation
Carl-Erik Särndal 1

Discussion
J. Michael Brick 23
Roger Tourangeau 29

Breakoff and Unit Nonresponse Across Web Surveys
Andy Peytchev 33

Assessing Mode Effects in a National Crime Victimization Survey using Structural Equation Models: Social Desirability Bias and Acquiescence
Dirk Heerwegh and Geert Loosveldt 49

Designing Input Fields for Non-Narrative Open-Ended Responses in Web Surveys
Mick P. Couper, Courtney Kennedy, Frederick G. Conrad and Roger Tourangeau 65

Using Register Data to Evaluate the Effects of Proxy Interviews in the Norwegian Labour Force Survey
Ib Thomsen and Ole Villund 87

Linear Regression Influence Diagnostics for Unclustered Survey Data
Jianzhu Li and Richard Valliant 99

Evaluating the Small-Sample Bias of the Delete-a-Group Jackknife for Model Analyses
Phillip S. Kott and Steven T. Garren 121

Letter to the Editor
James R. Knaub 135

Book Review 139
In Other Journals 149

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 38, No. 4, December/décembre 2010

Abbas Khalili
New estimation and feature selection methods in mixture-of-experts models 519

Iván A. Carrillo, Jiahua Chen and Changbao Wu
The pseudo-GEE approach to the analysis of longitudinal surveys 540

Irène Gijbels, Marek Omelka and Dominik Sznajder
Positive quadrant dependence tests for copulas 555

Hanfeng Chen, Jiahua Chen and Shun-Yi Chen
Confidence intervals for the mean of a population containing many zero values under unequal-probability sampling 582

Mahmoud Torabi and Jon N.K. Rao
Mean squared error estimators of small area means using survey weights 598

Zhiqiang Tan
Nonparametric likelihood and doubly robust estimating equations for marginal and nested structural models 609

Xiaogang Duan, Jing Qin and Qihua Wang
Optimal estimation in surrogate outcome regression problems 633

Jesse Frey
Testing for equivalence of variances using Hartley's ratio 647

Yanyuan Ma and Guosheng Yin
Semiparametric median residual life model and inference 665

Samiran Sinha
An estimated-score approach for dealing with missing covariate data in matched case-control studies 680

Mahmoud Torabi and Rhonda J. Rosychuk
Spatio-temporal modelling of disease mapping of rates 698

Guohua Yan, William J. Welch and Ruben H. Zamar
Model-based linear clustering 716

Tingting Gou and Duncan Murdoch
Simulation of extremes of diffusions 738

CONTENTS TABLE DES MATIÈRES

Volume 39, No. 1, March/mars 2011

Baoying Yang, Gengsheng Qin and Jing Qin
Empirical likelihood-based inferences for a low income proportion 1

Gengsheng Qin, Xiaoping Jin and Xiao-Hua Zhou
Non-parametric interval estimation for the partial area under the ROC curve..... 17

Grace Y. Yi, Leilei Zeng and Richard J. Cook
A robust pairwise likelihood method for incomplete longitudinal binary data arising in clusters 34

Konstantinos Kalogeropoulos, Petros Dellaportas and Gareth O. Roberts
Likelihood-based inference for correlated diffusions..... 52

Chi-Chung Wen, Steve Y.H. Huang and Yau-Hung Chen
Cox regression for current status data with mismeasured covariates 73

Jianqiang C. Wang and Mary C. Meyer
Testing the monotonicity or convexity of a function using regression splines..... 89

Weixing Song and Juan Du
A note on testing the regression functions via nonparametric smoothing..... 108

Hirokazu Yanagihara, Ken-Ichi Kamo and Tetsuji Tonda
Second-order bias-corrected AIC in multivariate normal linear models under non-normality 126

Li-Chun Zhang and Nina Hagesæther
A domain outlier robust design and smooth estimation approach..... 147

Steven N. MacEachern and Subharup Guha
Parametric and semiparametric hypotheses in the linear model..... 165