Nº 12-001-XIF au catalogue

Techniques d'enquête

Décembre 2006

Statistique Canada
Statistics Canada
Canada

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : 1-800-263-1136).

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

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

Renseignements pour accéder ou commander le produit

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

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

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

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

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par

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

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

Normes de service à la clientèle

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

Statistique Canada

Division des méthodes d'enquêtes auprès des entreprises

Techniques d'enquête

Décembre 2006

Publication autorisée par le ministre responsable de Statistique Canada

Décembre 2006

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

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

Périodicité : semestriel

Ottawa

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

Note de reconnaissance

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

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

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

POLITIQUE DE RÉDACTION

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

Présentation de textes pour la revue

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

Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 $ CA par année. Le prix n'inclus pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada: États-Unis 12 $ CA (6 $ × 2 exemplaires); autres pays, 30 $ CA (15 $ × 2 exemplaires). Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division de la diffusion, Gestion de la circulation, 150 Promenade du Pré Tunney, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au 1 800 700-1033, par télécopieur au 1 800 889-9734 ou par Courriel: order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et statisticiens du Québec. Des versions électroniques sont disponibles sur le site internet de Statistique Canada : www.statcan.ca.

Techniques d'enquête

Une revue éditée par Statistique Canada
Volume 32, numéro 2, décembre 2006

Table des matières

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

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

Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 135-136
Statistique Canada, N° 12-001 au catalogue

Dans ce numéro

Ce numéro de *Techniques d'enquête* débute par le sixième article de la série d'articles annuels sollicités dédiée à Joseph Waksberg. C'est avec tristesse que nous soulignons sa disparition survenue en janvier 2005. Une brève biographie de Joseph Waksberg a été publiée dans notre numéro de juin 2001, qui présentait aussi le premier article de la série. Pour plus d'informations sur la vie et l'oeuvre de Joseph Waksberg, voir l'article de Statistical Science (Vol. 15, No 3) "A Conversation with Joseph Waksberg," par David Morganstein et David Marker disponible à http://projecteuclid.org/Dienst/UI/1.0/Home. J'aimerais remercier les membres du Comité de sélection – David Bellhouse, président, Gordon Brackstone, Sharon Lohr et Wayne Fuller – d'avoir choisi Alastair Scott comme auteur de l'article du prix Waksberg de cette année.

Dans son article intitulé « Études cas-témoins basées sur la population », Scott traite des études cas-témoins pour lesquelles les témoins sont obtenus à partir d'une enquête par sondage complexe. En utilisant l'exemple de la régression logistique, il démontre que les estimations pondérées de l'enquête peuvent être passablement inefficaces en raison du poids assez faible attribué aux cas. Établissant une analogie avec l'estimation du maximum de vraisemblance, il propose ensuite une solution de rechange simple et bien plus efficace, mais biaisée pour l'ordonnée à l'origine. À l'aide d'exemples, il démontre les propriétés d'efficacité et de robustesse. Finalement, il examine brièvement la question des études cas-témoins portant sur des familles.

Pour sa part, Kott examine la pondération par calage pour corriger les erreurs de non-réponse et de couverture. Il donne une description générale de l'estimation par calage et élargit l'approche de la forme fonctionnelle d'Estavo et de Särndal à la fonction générale de calage. Il analyse ensuite les propriétés de cette méthode de calage pour corriger les erreurs dues à la non-réponse des unités et les erreurs de couverture en fonction d'un modèle de quasi-randomisation. Il conclut avec un exemple empirique et une analyse de certaines questions importantes.

À l'aide d'une étude de simulation, Reiter, Raghunathan et Kinney étudient les conséquences de ne pas tenir compte des variables du plan de sondage lors de l'élaboration de modèles d'imputation dans un contexte d'imputation multiple. Ils montrent que les biais potentiels peuvent être réduits en incluant ces variables dans le modèle d'imputation au moyen d'un modèle à effets fixes ou à effets mixtes. Les auteurs concluent que les imputeurs auraient intérêt à inclure comme prédicteurs toutes les variables liées aux variables imputées, en particulier les variables du plan de sondage, de manière à satisfaire aux conditions de l'hypothèse habituelle selon laquelle la non-réponse est ignorable.

Dans leur article, Funaoka, Saigo, Sitter et Toida se penchent sur l'utilisation des estimateurs de variance bootstrap dans l'échantillonnage stratifié à plusieurs degrés lorsque les fractions de sondage sont grandes. Ils proposent une méthode bootstrap de type Bernoulli qui donne des estimations cohérentes de la variance bootstrap lorsque l'échantillonnage aléatoire simple sans remise est utilisé à chaque degré. La méthode proposée est simple à appliquer et peut être étendue à n'importe quel nombre de degrés d'échantillonnage sans trop de complications. La méthode est illustrée à l'aide d'une étude de simulation limitée qui utilise des données de l'Enquête nationale sur les prix menée en 1997 au Japon.

Dans leur article, Kozak et Verma comparent l'approche géométrique de stratification, proposée par Gunning et Horgan (2004), avec deux approches par optimisation : l'algorithme de Lavallé-Hidiroglou (Lavallée et Hidiroglou 1988) et l'algorithme d'optimisation proposé par Kozak (2004). Au moyen de cinq populations fictives de différentes tailles, les trois méthodes sont comparées selon deux scénarios : une comparaison du coefficient de variation (c.v.) résultant selon une taille d'échantillon fixe et une comparaison des tailles d'échantillon résultant selon un niveau de précision fixe.

Deville et Lavallée, pour leur part, présentent les fondements théoriques généraux de la méthode du partage des poids appliquée au sondage indirect. Ils définissent l'important concept d'une matrice de liens dans le sondage indirect, qui précise comment les éléments de la population échantillonnée sont liés à la population cible et qui pondère ces liens pour que l'estimation soit sans biais. Ils examinent d'importantes propriétés de la matrice de liens et définissent les conditions nécessaires et suffisantes à une matrice de liens optimale. Les auteurs démontrent leur théorie avec quelques exemples intéressants.

Deville et Maumy-Bertrand étudient la détermination d'un plan de sondage et d'une méthode d'estimation dans le cas d'une enquête sur la fréquentation touristique. Le problème principal posé par ce type d'enquêtes est l'absence d'une base de sondage permettant d'atteindre directement les touristes. Pour contourner ce problème, les auteurs suggèrent d'échantillonner des services destinés principalement aux touristes. On se trouve donc dans une situation de sondage indirect pour laquelle la méthode généralisée du partage des poids est utilisée pour obtenir les estimations des paramètres d'intérêt. Certaines extensions à la méthode sont nécessaires. Les auteurs portent leur attention plus particulièrement à l'une d'elles et la décrivent en détail.

Félix-Medina et Monjardin examinent une variante de l'échantillonnage par dépistage de liens. Pour ce faire, ils ont recours à une approche Bayesienne pour construire des estimateurs de taille de population. Cependant, pour que les inférences sur la taille de la population soient robustes aux spécifications erronées du modèle hypothétique, les auteurs font ces inférences selon une approche fréquentiste basée sur le plan. L'étude de simulation montre que les estimateurs proposés donnent de meilleurs résultats que les estimateurs du maximum de vraisemblance qui sont utilisés actuellement.

Dans leur article, Dorfman, Lent, Leaver et Wegman comparent la méthodologie de l'indice des prix à la consommation du Royaume-Uni et celle des États-Unis en s'appuyant sur les mêmes données scanographiques. Ils concluent que pour la population étudiée, l'approche du Royaume-Uni, qui suppose une stratification plus rigoureuse et, ce qui importe davantage, un échantillonnage intrastrate par choix raisonné plus restrictif que l'approche par échantillonnage probabiliste des États-Unis, donne un meilleur indice superlatif estimé cible. C'est le cas quel que soit l'estimateur (ratio des moyennes, moyenne géométrique ou moyenne des ratios) utilisé pour produire l'indice de prix de bas niveau.

Dans leur article, Thomas, Raghunathan, Schenker, Katzoff et Johnson ont recours à l'imputation multiple pour analyser des données qui, à cause d'un plan de sondage matriciel, présentent des valeurs manquantes. Dans le plan de sondage matriciel, seul un sous-ensemble de questions sont posées à chaque répondant afin de réduire le fardeau de réponse. Les auteurs ont élaboré une méthode permettant de créer des formulaires de sondage matriciel. Chaque formulaire contient un sous-ensemble de questions à poser à des répondants choisis au hasard. La méthode est conçue pour que chaque formulaire comprenne des questions qui permettent de prédire les réponses aux questions exclues, ce qui permet de déduire une partie de l'information exclue. La méthode proposée et l'imputation multiple sont évaluées à l'aide de données tirées de la National Health and Nutrition Examination Survey.

Harold Mantel, Rédacteur en chef délégué

Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 137-147
Statistique Canada, N° 12-001 au catalogue

Études cas-témoins basées sur la population

Alastair Scott [1]

Résumé

Nous discutons de méthodes d'analyse des études cas-témoins pour lesquelles les témoins sont sélectionnés selon un plan de sondage complexe. La méthode la plus simple est l'approche du sondage standard basée sur des versions pondérées des équations d'estimation pour la population. Nous examinons aussi des méthodes plus efficaces et comparons leur degré de robustesse aux erreurs de spécification du modèle dans des cas simples. Nous discutons également brièvement des études familiales cas-témoins, pour lesquelles la structure intragrappe présente un intérêt en soi.

Mots clés : Études cas-témoins; échantillonnage sélectif; échantillonnage rétrospectif; pondération.

1. Introduction

L'étude cas-témoins, dans laquelle des échantillons distincts sont tirés parmi les « cas » (disons, les personnes présentant une maladie d'intérêt) et parmi les « témoins » (personnes n'ayant pas la maladie), est l'une des méthodes d'étude les plus fréquentes en recherche sur la santé. En fait, Breslow (1996) a qualifié ce genre d'étude de « pivot de l'épidémiologie ». Nous nous concentrerons ici sur les applications biostatistiques, mais le plan de base représente une stratégie d'échantillonnage efficace dans les situations où les cas sont rares et est aussi utilisé fréquemment dans de nombreux autres domaines (commerce, sciences sociales, écologie, études de marché, par exemple). On a notamment assisté dans la littérature économétrique au développement parallèle de la plupart de la théorie de l'échantillonnage basé sur les choix (voir, par exemple, Manski et McFadden 1981; Cosslett 1981).

Il existe deux types fondamentalement différents d'études cas-témoins, à savoir les études avec cas-témoins appariés, dans lesquelles chaque cas est apparié à un ou à plusieurs témoins, et les études avec cas-témoins non appariés, dans lesquelles les échantillons de cas et de témoins sont tirés indépendamment, quoiqu'il puisse exister un vague « appariement fréquentiste », l'échantillon de témoins étant réparti entre des strates définies par des variables démographiques de base de façon telle que la distribution de ces variables dans l'échantillon de témoins soit la même que celle prévue dans l'échantillon de cas. Nous ne nous intéressons ici qu'aux études sans appariement et, plus précisément, uniquement à la catégorie restreinte d'études sur la population dans lesquelles les témoins (et, à l'occasion, les cas également) sont sélectionnés en utilisant des méthodes standard d'échantillonnage.

Une excellente introduction à l'échantillonnage cas-témoins et à ses points forts et ses inconvénients éventuels est donnée dans Breslow (1996, 2004). L'un des plus grands défis que doit relever toute personne qui conçoit une étude de ce genre consiste à s'assurer que les témoins soient réellement sélectionnés à partir de la même population, selon les mêmes protocoles, que les cas. Comme l'a dit Miettinen (1985), les cas et les témoins « devraient être représentatifs de la même expérience fondamentale » [traduction]. Le fait qu'au début, des mesures adéquates n'aient pas été prises à cet égard lors de certaines études a suscité des doutes au sujet de l'échantillonnage cas-témoins chez de nombreux chercheurs. Une discussion approfondie des principes qui devraient régir la sélection des témoins figure dans Wacholder, McLaughlin, Silverman et Mandel (1991). Puisque l'essence de l'échantillonnage tient aux méthodes suivies pour tirer des échantillons représentatifs à partir d'une population cible, il est devenu naturel de penser aux méthodes de sondage pour obtenir les témoins. De plus en plus fréquemment, au cours des quelque 25 dernières années, les témoins (et parfois les cas également) ont été sélectionnés en utilisant des plans de sondage complexes stratifiés à plusieurs degrés. Le lecteur trouvera au chapitre 9 de Korn et Graubard (1999) un bon historique de cette évolution.

L'analyse de ce genre d'études est un sujet tout indiqué pour le présent article, car Joe Waksberg lui-même a été l'un des principaux artisans de l'adoption des méthodes de sondage (et de la composition aléatoire, en particulier) pour l'obtention des témoins (voir, par exemple, Waksberg 1998, ainsi que DiGaetano et Waksberg 2002).

2. Exemples

Nous commençons par deux exemples en vue d'illustrer le genre de problèmes que nous voulons résoudre. Le premier est typique des études à grande échelle réalisées par le National Cancer Institute, aux employés duquel nous devons la plupart des progrès réalisés dans le domaine. Joe Waksberg, et ses collègues de Westat ont eu une

1. Alastair Scott, Department of Statistics, University of Auckland, Auckland 1, Nouvelle-Zélande. Courriel : a.scott@auckland.ac.nz.

influence considérable sur le choix des méthodes d'échantillonnage adoptées pour réaliser ces études (voir Hartge, Brinton, Rosenthal, Cahill, Hoover et Waksberg 1984 qui donnent aussi une description de plusieurs études similaires), de sorte qu'il s'agit d'un point de départ naturel.

Exemple 1

En 1977 – 1978, le National Cancer Institute et l'Environmental Protection Agency des États-Unis ont réalisé une étude cas-témoins sur la population afin d'examiner les effets des rayonnements ultraviolets sur le cancer de la peau de type non-mélanome au cours d'une période d'un an (Hartge, Brinton, Rosenthal, Cahill, Hoover et Waksberg 1984; Fears et Gail 2000). L'étude a été réalisée dans huit régions géographiques où l'intensité des rayonnements solaires ultraviolets différait. Dans chaque région, un échantillon de personnes atteintes d'un cancer de la peau de type non-mélanome âgées de 20 à 74 ans et un échantillon de témoins sélectionnés dans la population générale ont été interviewés par téléphone afin d'obtenir des renseignements sur les facteurs de risque. Pour chaque région, un échantillon aléatoire simple de 450 cas et un échantillon supplémentaire de 50 cas du groupe des 20 à 49 ans ont été sélectionnés en vue d'une prise de contact. Pour les témoins, 500 ménages ont été échantillonnés dans chaque région selon la méthode de composition aléatoire de Mitofsky-Waksberg (Waksberg 1978). On s'est efforcé, dans la mesure du possible, d'interviewer tous les adultes de 65 à 74 ans, ainsi qu'une personne de chaque sexe de 20 à 64 ans choisie au hasard. En outre, un deuxième échantillon de Mitofsky-Waksberg contenant de 500 à 2 100 ménages a été sélectionné et des renseignements ont été recueillis au sujet de tous les adultes de 65 à 74 ans. Cela a donné un échantillon d'environ 3 000 cas et un échantillon d'environ 8 000 témoins, le taux d'échantillonnage des cas étant environ 300 fois celui des témoins, selon l'âge.

Le deuxième exemple revêt une importance particulière pour moi, car il correspond à la première incursion que Chris Wild et moi-même avons faite dans ce domaine.

Exemple 2

L'étude de la méningite à Auckland a été exécutée à la demande du ministère de la Santé et du Conseil de recherche sur la santé de la Nouvelle-Zélande en vue d'examiner les facteurs de risque de méningite chez les jeunes enfants parmi lesquels la progression de la maladie prenait des proportions épidémiques (voir Baker, McNicholas, Garrett, Jones, Stewart, Koberstein et Lennon 2000). L'étude avait pour population cible l'ensemble des enfants de moins de neuf ans dans la région d'Auckland entre 1997 et 2000.

Tous les cas de méningite relevés dans le groupe d'âge cible au cours des trois années qu'a duré l'étude ont été inclus, ce qui a donné environ 250 cas. Un nombre comparable de témoins ont été sélectionné parmi les autres enfants faisant partie de la population étudiée selon un plan d'échantillonnage complexe à plusieurs degrés. Au premier degré d'échantillonnage, 300 îlots de recensement (contenant chacun environ 70 ménages) ont été sélectionnés avec probabilité proportionnelle au nombre de maisons dans l'îlot. Au deuxième degré, un échantillon systématique de 20 ménages a été tiré dans chaque îlot sélectionné et les enfants provenant de ces ménages ont été choisis pour l'étude avec une probabilité variant selon l'âge et l'ethnicité, déterminée de façon qu'elle corresponde à la fréquence prévue parmi les cas. Les probabilités de sélection sont présentées plus loin au tableau 1 (IP signifie originaire des îles du Pacifique). La taille des échantillons de grappes varie de 1 à 6, et environ 250 témoins ont été sélectionnés en tout. Cela correspond à une fraction d'échantillonnage d'environ 1 pour 400, en moyenne, de sorte que les cas ont été échantillonnés à un taux d'environ 400 fois celui des témoins.

Ces deux études sont assez représentatives de celles dont nous voulons discuter. Elles illustrent aussi les deux principales méthodes d'échantillonnage utilisées, à savoir la composition aléatoire et l'échantillonnage aréolaire. Une vive discussion des mérites relatifs de ces deux stratégies figure dans Brogan, Denniston, Liff, Flagg, Coates et Brinton (2001), ainsi que dans DiGaetano et Waksberg (2002).

Tableau 1
Probabilités de sélection

ÂGE	MAORI	ÎLES DU PACIFIQUE	AUTRE
≤ 1 an	0,29	0,70	0,10
≤ 3 ans	0,15	0,50	0,07
≤ 5 ans	0,15	0,31	0,04
≤ 8 ans	0,15	0,17	0,04

3. Conditions générales

Supposons que nous ayons une variable de réponse binaire, Y, avec $Y = 1$ dénotant un cas et $Y = 0$ dénotant un témoin, et un vecteur de variables explicatives éventuelles, $\mathbf{x}$. Nous supposons que la valeur de Y est connue pour chacune des N unités d'une population cible donnée, mais qu'au moins certaines composantes de $\mathbf{x}$ sont inconnues. Nous stratifions la population en cas et en témoins, tirons un échantillon dans chaque strate d'après les variables que nous connaissons pour toutes les unités, et mesurons les valeurs des covariables manquantes pour les unités échantillonnées (en pratique, l'échantillon de témoins est souvent tiré à partir de l'ensemble de la population, plutôt que parmi les unités pour lesquelles $Y = 0$. Si la

proportion de cas est faible, la différence est négligeable. Sinon, il est simple d'adapter les résultats qui suivent à cette variante – pour un développement rigoureux, voir Lee, Scott et Wild 2006). Habituellement, nous voulons ensuite utiliser les données d'échantillon pour ajuster un modèle de régression binaire de la probabilité marginale qu'une unité soit un cas ayant la forme d'une fonction des covariables. Le modèle utilisé est presque toujours logistique avec

$$\text{logit}\,\{P(Y = 1 \mid \mathbf{x})\} = \log\left(\frac{P(Y = 1 \mid \mathbf{x})}{P(Y = 0 \mid \mathbf{x})}\right) = \beta_0 + \mathbf{x}^T\boldsymbol{\beta}_1 \qquad (1)$$

disons, où β_0 et $\boldsymbol{\beta}_1$ sont des paramètres inconnus et, tout au long de l'article, nous supposerons que nous avons affaire au modèle (1). Les extensions à des modèles de régression plus généraux sont simples en principe (voir Scott et Wild 2001b), mais les expressions résultantes sont un peu moins élégantes que le modèle logistique.

Comment devrions-nous nous y prendre pour ajuster le modèle (1) sachant les données d'échantillon? Les méthodes efficaces sont faciles à appliquer en cas d'échantillonnage aléatoire simple ou stratifié, mais nous nous intéressons ici à des méthodes d'échantillonnage plus complexes. Très souvent, on omet tout bonnement de tenir compte de l'échantillonnage complexe, ce qui risque d'entraîner tous les problèmes qui se posent habituellement lorsqu'on ne prend pas en compte la structure du plan d'échantillonnage. Des probabilités de sélection variables pourraient fausser la structure de la moyenne de sorte que les estimations produites par les programmes standard risquent d'être non cohérentes. La corrélation intragrappe pourrait réduire la taille effective d'échantillon, de sorte que les erreurs-types produites couramment seraient trop faibles, les intervalles de confiance, trop courts, les valeurs p, trop faibles, et ainsi de suite. Une stratégie simple qu'adoptent certains chercheurs pour réduire au minimum les effets consiste à garder petit le nombre de sujets dans chaque grappe (voir Graubard, Fears et Gail 1989, par exemple). Cela réduit l'effet de plan, donc l'incidence sur la mise en grappes, mais le remède peut être coûteux. Dans les sections qui suivent, nous examinons certains moyens éventuels d'utiliser les plans d'échantillonnage standard, plus rentables.

4. Approche de la pondération de sondage

Une option évidente consiste à suivre l'approche standard des équations d'estimations pondérées qui est intégrée dans la plupart des progiciels contemporains d'analyse de données d'enquête (voir Binder 1983). Supposons d'abord que nous ayons des données provenant de la population finie complète. Si nous émettons l'hypothèse que cette population finie est tirée à partir d'une superpopulation dans laquelle le modèle logistique conditionnel (1) est vérifié, alors nous pourrions estimer $\boldsymbol{\beta}$ en résolvant les équations d'estimation relatives à l'ensemble de la population, c'est-à-dire sous recensement, suivantes

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_1^N \mathbf{x}_i(y_i - p_1(\mathbf{x}_i;\boldsymbol{\beta})) = 0, \qquad (2)$$

où $p_1(\mathbf{x};\boldsymbol{\beta}) = e^{\beta_0 + \mathbf{x}^T\boldsymbol{\beta}_1}/(1 + e^{\beta_0 + \mathbf{x}^T\boldsymbol{\beta}_1})$. (Il s'agit des équations de vraisemblance si l'on suppose que les unités de population sont échantillonnées indépendamment à partir d'une superpopulation, mais que les estimateurs résultants sont convergents sous des structures de population nettement plus réalistes, à condition que le modèle (1) soit vérifié marginalement – voir Rao, Scott et Skinner 1998 pour une discussion plus approfondie.)

Maintenant, pour toute valeur fixée de $\boldsymbol{\beta}$, $\mathbf{S}(\boldsymbol{\beta})$ dans l'équation (2) est simplement un vecteur de totaux de population. Nous pouvons donc l'estimer d'après l'échantillon, disons par

$$\hat{\mathbf{S}}(\boldsymbol{\beta}) = \sum_{\text{échantillon}} w_i\mathbf{x}_i(y_i - p_1(\mathbf{x}_i;\boldsymbol{\beta})), \qquad (3)$$

où w_i est l'inverse de la probabilité de sélection, peut-être corrigée de la non-réponse et de la poststratification. Fixer $\hat{\mathbf{S}}(\boldsymbol{\beta})$ égal à 0 nous donne notre estimateur, $\hat{\boldsymbol{\beta}}$. Nous pourrions appliquer la linéarisation ou le jacknife directement à $\hat{\boldsymbol{\beta}}$ pour obtenir les erreurs-types. Nous pouvons aussi développer $\hat{\mathbf{S}}(\hat{\boldsymbol{\beta}})$ autour de la valeur réelle, $\boldsymbol{\beta}$, et obtenir comme matrice de covariance estimée l'estimateur « sandwich »

$$\hat{\text{Cov}}\{\hat{\boldsymbol{\beta}}\} \approx \mathbf{J}(\hat{\boldsymbol{\beta}})^{-1}\,\hat{\text{Cov}}\{\hat{S}(\hat{\boldsymbol{\beta}})\}\,\mathbf{J}(\hat{\boldsymbol{\beta}})^{-1}, \qquad (4)$$

où $\mathbf{J}(\boldsymbol{\beta}) = -\partial\mathbf{S}/\partial\boldsymbol{\beta}^T = \sum_{\text{échantillon}} w_i\,p_1(\mathbf{x}_i;\boldsymbol{\beta})\,p_0(\mathbf{x}_i;\boldsymbol{\beta})\,\mathbf{x}_i\mathbf{x}_i^T$ avec $p_0 = 1 - p_1$. Puisque $\hat{\mathbf{S}}(\boldsymbol{\beta})$ est un vecteur de totaux, $\hat{\text{Cov}}\{\hat{\mathbf{S}}(\boldsymbol{\beta})\}$ devrait être disponible d'office pour tout plan de sondage standard. De nos jours, la plupart des grands progiciels statistiques (par exemple, SAS (PROC SURVEYLOGISTIC), SPSS (CSLOGISTIC), STATA (SVY:LOGIT), SUDAAN (LOGISTIC)) peuvent traiter de façon courante la régression logistique en cas d'échantillonnage complexe et de pondération. Par conséquent, il est assez simple de produire des estimations pondérées et de faire les inférences connexes.

Strictement parlant, dans notre cadre fondé sur un modèle, les probabilités de sélection sont souvent elles-mêmes des variables aléatoires basées sur une population finie hypothétiquement générée à partir du modèle. Nous pouvons tenir compte de ce fait en utilisant les résultats de Rao (1973), mais la correction est d'ordre $1/N$ et peut être ignorée dans la plupart des études de grande taille.

L'inconvénient de la pondération est, en général, qu'elle a tendance à être inefficace si les poids sont très variables (une règle empirique parfois proposée est que $w_{\max} / w_{\min}$ ne devrait pas être supérieur à 10). Dans les études cas-témoins, la variation des poids est à peu près aussi extrême qu'elle peut l'être. Par exemple, le ratio de $w_{\max}$ par rapport à $w_{\min}$ est approximativement de 300 pour 1 dans l'exemple 1, et de 1 000 pour 1 dans l'exemple 2. Et des ratios encore plus extrêmes ne sont pas inhabituels. Aucun spécialiste chevronné de l'échantillonnage ne s'étonnerait de constater que la pondération n'est guère efficace dans ces circonstances.

Pouvons-nous trouver une solution plus efficace? La réponse est assurément affirmative dans certains cas. Des méthodes de vraisemblance entièrement efficaces ont été élaborées dans des situations où les cas et les témoins sont sélectionnés par échantillonnage aléatoire simple ou stratifié, et ces méthodes peuvent être nettement plus efficaces que les méthodes de pondération. Nous passons ces résultats en revue à la section suivante.

5. Révision : Cas simple

Examinons d'abord la situation la plus simple où les cas et les témoins sont sélectionnés par échantillonnage aléatoire simple et où nous ne disposons d'information au niveau de la population sur aucune des covariables à l'étape de l'élaboration du plan de sondage. Pour ce cas, il existe des méthodes semi-paramétriques entièrement efficaces d'estimation du maximum de vraisemblance bien établies. En outre, ces méthodes sont très faciles à appliquer en utilisant les logiciels standard (Prentice et Pyke 1979) (les méthodes sont semi-paramétriques, parce que la vraisemblance complète dépend de la distribution inconnue des covariables, ce que nous voulons en général modéliser).

En fait, il nous suffit d'ajuster le modèle (1) en utilisant un programme standard de régression logistique sans aucune pondération. Plus précisément, la résolution de l'équation non pondérée

$$\sum_{\text{échantillon}} \mathbf{x}_i (y_i - p_1(\mathbf{x}_i; \boldsymbol{\beta})) = \mathbf{0}, \qquad (5)$$

produit des estimations efficaces de tous les coefficients, sauf l'ordonnée à l'origine. Fait peut-être plus important, toutes les erreurs-types et les inférences résultantes que nous obtenons au moyen du programme standard sont également valides, de nouveau à l'exception de tout résultat où intervient l'ordonnée à l'origine. Il est assez simple de corriger les inférences dans lesquelles intervient l'ordonnée à l'origine, à condition que nous connaissions les fractions d'échantillonnage, mais nous ne nous intéressons souvent qu'aux autres coefficients de toute façon.

Les résultats s'étendent directement à l'échantillonnage aléatoire stratifié, à condition d'inclure dans le modèle une ordonnée à l'origine distincte pour chaque strate. De nouveau, il est possible d'obtenir des estimateurs semi-paramétriques efficaces de tous les coefficients, sauf les ordonnées à l'origine de strate en traitant simplement les données à l'aide d'un programme de régression logistique (non pondérée) ordinaire. De nouveau, les erreurs-types estimées et les inférences connexes sont également valides. Comme dans le cas de l'échantillonnage aléatoire simple, nous pouvons corriger les résultats pour les ordonnées à l'origine de strate à condition que nous connaissions les fractions d'échantillonnage dans les strates, mais, encore une fois, celles-ci ne présentent habituellement que peu d'intérêt.

Donc, dans ces situations simples, les estimations du maximum de vraisemblance sont plus faciles à calculer que les estimations pondérées, et elles sont aussi plus efficaces. Dans quelle mesure le sont-elles? Cela dépend du nombre de covariables, de la grandeur de leur coefficient et du ratio des fractions d'échantillonnage, mais la différence est souvent importante (ainsi, les estimations pondérées ont une efficacité d'environ 50 % dans l'exemple 2 de l'introduction, et de moins de 20 % dans l'exemple du cancer du cerveau que nous examinons à la section 8. Lawless, Kalbfleisch et Wild (1999) discutent de situations où l'efficacité est encore plus faible).

Enfin, nous notons que les estimations du maximum de vraisemblance présentent encore un autre avantage par rapport aux estimations pondérées : elles ont tendance à avoir de nettement meilleures propriétés sur petit échantillon, surtout quand l'efficacité des estimations pondérées est faible. Essentiellement, la pondération réduit la taille effective d'échantillon et cette dernière est l'élément qui détermine le moment où la théorie asymptotique commence à donner une bonne approximation (voir Scott et Wild 2001a pour plus de précisions). Manifestement, le prix de l'adhésion stricte aux poids de population peut être très lourd.

6. Échantillonnage plus complexe

Dans les deux exemples de la section 2, les témoins ont été obtenus par échantillonnage complexe à plusieurs degrés plutôt que par échantillonnage aléatoire simple. Comme nous l'avons mentionné dans l'introduction, cette situation est de plus en plus fréquente dans les études cas-témoins à grande échelle (à l'occasion, comme dans l'exemple 1, les cas sont également sélectionnés selon un plan d'échantillonnage complexe). Il est possible de dériver des estimateurs semi-paramétriques efficaces pour l'échantillonnage stratifié à plusieurs degrés, en supposant que les unité primaire

d'échantillonnage sont sélectionnées indépendamment dans les strates (hypothèse qui est de toute façon celle faite dans tous les progiciels offrant l'approche de pondération par les poids de sondage), mais cela nous oblige à construire des modèles multivariés pour le vecteur des réponses dans une unité primaire d'échantillonnage. Le lecteur trouvera des renseignements détaillés dans Neuhaus, Scott et Wild (2002, 2006). À moins que nous ne nous intéressions à la structure intragrappe en soi (comme dans les études familiales cas-témoins considérées à la section 9, par exemple), l'exercice demande beaucoup trop d'efforts pour être faisable, du moins dans les analyses de routine.

Pouvons-nous faire quelque chose de plus simple sans perdre trop d'efficacité? Naturellement, nous pouvons toujours nous rabattre sur les estimations pondérées. Cependant, elles sont tout aussi inefficaces pour les plans de sondage complexes que pour le cas simple examiné à la section précédente. En fait, nous pouvons obtenir de sensiblement meilleurs résultats sans trop de complications supplémentaires.

Revenons un instant à la situation de la section précédente où nous avons un échantillon aléatoire simple de taille n_1 provenant de la strate de cas et un échantillon aléatoire simple indépendant de taille n_0 provenant de la strate de témoins. Ici, toutes les unités de la strate ℓ ont un poids $w_i \propto W_\ell / n_\ell$, où W_ℓ dénote la proportion de la population dans la strate, pour $\ell = 0,1$. Si nous divisons tout au long par N et posons que $p_0(\mathbf{x};\boldsymbol{\beta}) = 1 - p_1(\mathbf{x};\boldsymbol{\beta})$, alors nous pouvons réécrire l'équation (3) pour l'estimateur pondéré sous la forme

$$W_1 \frac{\sum_{\text{cas}} \mathbf{x}_i \, p_0(\mathbf{x}_i;\boldsymbol{\beta})}{n_1} - W_0 \frac{\sum_{\text{témoins}} \mathbf{x}_i \, p_1(\mathbf{x}_i;\boldsymbol{\beta})}{n_0} = \mathbf{0}. \quad (6)$$

De même, nous pouvons écrire l'équation (5) pour l'estimateur efficace du maximum de vraisemblance sous la forme

$$\omega_1 \frac{\sum_{\text{cas}} \mathbf{x}_i \, p_0(\mathbf{x}_i;\boldsymbol{\beta})}{n_1} - \omega_0 \frac{\sum_{\text{témoins}} \mathbf{x}_i \, p_1(\mathbf{x}_i;\boldsymbol{\beta})}{n_0} = \mathbf{0}, \quad (7)$$

où $\omega_\ell = n_\ell /(n_0 + n_1)$, pour $\ell = 0,1$. Ces deux expressions sont des cas particuliers de l'ensemble général d'équations d'estimation

$$\lambda_1 \frac{\sum_{\text{cas}} \mathbf{x}_i \, p_0(\mathbf{x}_i;\boldsymbol{\beta})}{n_1} - \lambda_0 \frac{\sum_{\text{témoins}} \mathbf{x}_i \, p_1(\mathbf{x}_i;\boldsymbol{\beta})}{n_0} = \mathbf{0}. \quad (8)$$

À mesure que $n_0, n_1 \to \infty$, sous des contraintes faibles quant à la façon dont la population finie est générée à partir de la superpopulation, la solution de (8) converge presque certainement vers la solution $\boldsymbol{\beta}^*$ de

$$\lambda_1 E_1\{\mathbf{X}\, p_0(\mathbf{X};\boldsymbol{\beta}^*)\} - \lambda_0 E_0\{\mathbf{X}\, p_1(\mathbf{X};\boldsymbol{\beta}^*)\} = \mathbf{0}, \quad (9)$$

où $E_\ell\{\cdot\}$ dénote l'espérance conditionnelle sachant que $Y = \ell$ pour $\ell = 0,1$. Si le modèle (1) est vérifié, alors l'équation (8) a pour solution $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_1$ et $\beta_0^* = \beta_0 + b_\lambda$ avec $b_\lambda = \log(\lambda_1 W_0 / \lambda_0 W_1)$ pour toute valeur positive de λ_0, λ_1 [voir Scott et Wild (1986) pour des détails de la preuve]. Donc, la solution de l'équation (8) produit des estimateurs convergents pour tous les coefficients de régression, sauf le terme constant pour tout $\lambda_\ell > 0$ ($\ell = 0,1$). Comme dans le cas simple, il est facile de corriger les inférences au sujet du terme constant, à condition de connaître la proportion de cas dans la population.

Maintenant, passons à des plans d'échantillonnage plus complexes. Puisque le premier membre de l'équation (9) ne fait intervenir que deux moyennes de sous-population, nous pouvons encore estimer ces moyennes pour tout plan de sondage standard. Cela suggère un estimateur, disons $\hat{\boldsymbol{\beta}}_\lambda$, pour les plans d'échantillonnage généraux qui satisfait

$$\hat{\mathbf{S}}_\lambda(\boldsymbol{\beta}) = \lambda_1 \hat{\boldsymbol{\mu}}_1(\boldsymbol{\beta}) - \lambda_0 \hat{\boldsymbol{\mu}}_0(\boldsymbol{\beta}) = \mathbf{0}, \quad (10)$$

où $\hat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta})$ est l'estimateur sur échantillon de la moyenne de sous-population $E_\ell\{\mathbf{X}(1 - p_\ell(\mathbf{X};\boldsymbol{\beta}))\}$ ($\ell = 0,1$). La matrice de covariance de $\hat{\boldsymbol{\beta}}_\lambda$ peut alors être obtenue à l'aide d'arguments standard de linéarisation, ce qui nous mène à une matrice de covariance estimée (« sandwich »)

$$\hat{\mathrm{Cov}}\{\hat{\boldsymbol{\beta}}_\lambda\} \approx \mathbf{J}_\lambda(\hat{\boldsymbol{\beta}}_\lambda)^{-1} \hat{\mathrm{Cov}}\{\hat{\mathbf{S}}_\lambda(\hat{\boldsymbol{\beta}}_\lambda)\} \mathbf{J}_\lambda(\hat{\boldsymbol{\beta}}_\lambda)^{-1}, \quad (11)$$

avec $\mathbf{J}_\lambda(\boldsymbol{\beta}) = (-\partial \hat{\mathbf{S}}_\lambda(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}^T)$ et $\hat{\mathrm{Cov}}\{\hat{\mathbf{S}}_\lambda(\boldsymbol{\beta})\} = \lambda_1^2 \hat{\mathrm{Cov}}\{\hat{\boldsymbol{\mu}}_1(\boldsymbol{\beta})\} + \lambda_0^2 \hat{\mathrm{Cov}}\{\hat{\boldsymbol{\mu}}_0(\boldsymbol{\beta})\}$. Ici, $\hat{\mathrm{Cov}}\{\hat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta})\}$ dénote l'estimation par sondage habituelle qui devrait être disponible systématiquement pour tout plan de sondage standard, puisque $\hat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta})$ est simplement une moyenne estimée.

Tout cela peut aussi être exécuté facilement au moyen de n'importe quel progiciel capable de traiter la régression logistique sous plan de sondage complexe, simplement en spécifiant le vecteur approprié de poids. Plus précisément, supposons que

$$\hat{\boldsymbol{\mu}}_\ell(\boldsymbol{\beta}) = \frac{\sum_{i \in S_\ell} w_i \, \mathbf{x}_i \, (y_i - p_1(\mathbf{x}_i;\boldsymbol{\beta}))}{\sum_{i \in S_\ell} w_i}, \quad (12)$$

où S_1 dénote la sous-population de cas (c'est-à-dire l'ensemble des unités pour lesquelles $Y = 1$) et S_0 dénote la sous-population de témoins (l'ensemble des unités pour lesquelles $Y = 0$). Alors, l'équation d'estimation (9) peut s'écrire sous la forme

$$\hat{\mathbf{S}}_\lambda(\boldsymbol{\beta}) = \sum_{\text{échantillon}} w_i^* \, \mathbf{x}_i (y_i - p_1(\mathbf{x}_i;\boldsymbol{\beta})) = \mathbf{0}, \quad (13)$$

avec $w_i^* \propto \lambda_\ell w_i / \sum_{i \in S_\ell} w_i$ pour les unités comprises dans $S_\ell (\ell = 0,1)$. Autrement dit, nous devons simplement rééchelonner les poids des cas et des témoins séparément, de sorte que la somme de ceux des cas soit proportionnelle à λ_1 et que la somme de ceux des témoins soit proportionnelle à λ_0, puis les introduire, avec la spécification habituelle de la structure du plan de sondage (strates, unités primaires d'échantillonnage), dans le programme de notre choix. Soulignons que le choix de la constante de proportionnalité n'a pas d'incidence sur le résultat.

Il nous reste à décider des bonnes valeurs de λ_1 et λ_0. Nous pouvons souvent réaliser des gains importants en utilisant les poids d'échantillon ($\lambda_\ell = n_\ell / n$) plutôt que les poids de population ($\lambda_\ell = W_\ell$). Scott et Wild (2002) ont fait état de gains d'efficacité de 50 % ou plus dans l'exemple 2 et dans des simulations basées sur cette population. Les gains devenaient plus importants à mesure que s'intensifiait la force de la relation et qu'augmentait l'effet de la mise en grappe. De surcroît, la couverture des intervalles de confiance était plus proche de la valeur nominale en cas de pondération par les poids d'échantillon dans les simulations.

L'utilisation des poids d'échantillon est la stratégie la plus efficace disponible lorsqu'on a affaire à des échantillons aléatoires simples de cas et de témoins, mais pour des plans de sondage plus complexes, elle n'est plus entièrement efficace. Nous pourrions nous attendre à ce que des poids basés sur une forme de tailles d'échantillon équivalentes donnent de meilleurs résultats. Cette approche produit effectivement certains gains d'efficacité dans des simulations limitées décrites dans Scott et Wild (2001a). Toutefois, les gains sont relativement faibles, du moins lorsque l'effet de plan de l'échantillon de témoins est inférieur à 2, puisque $\mathrm{Cov}\{\hat{\boldsymbol{\beta}}_\lambda\}$ est une fonction de $\boldsymbol{\lambda}$ très plate près de son minimum. Les considérations relatives à la robustesse dont nous discutons à la section 8 pourraient jouer un rôle plus important dans le choix de $\boldsymbol{\lambda}$.

Les avantages qu'offre la pondération d'échantillon peuvent dépendre en grande partie du problème à l'étude. Korn et Graubard (1999, page 327) font remarquer que, dans leur expérience, la stratégie de pondération par les poids d'échantillon produit rarement de gros gains d'efficacité. De toute évidence, la poursuite des travaux, tant empiriques que théoriques, est nécessaire ici. Quoi qu'il en soit, il semble prudent d'ajuster le modèle en utilisant systématiquement les poids d'échantillon ainsi que les poids de population. Si les estimations des coefficients sont semblables, alors nous pouvons porter un jugement en nous basant sur les erreurs-types estimées. Par contre, des écarts significatifs entre les estimations des coefficients indiquent que le modèle a été mal spécifié. Si nous sommes incapables de corriger les déficiences du modèle, alors nous devons bien réfléchir à ce que nous essayons vraiment d'estimer. Nous examinons cette question à la section 8.

7. Échantillonnage stratifié

Le compromis proposé à la section précédente (c'est-à-dire utiliser la pondération standard par les poids de sondage dans les sous-populations définies d'après la situation de cas ou de témoin, mais combiner les sous-populations en utilisant les proportions d'échantillon) semble donner d'assez bons résultats en pratique, mais elle est entièrement *ad hoc*. Pourrions-nous obtenir de meilleurs résultats en adoptant une approche plus systématique?

Dans le cas particulier de l'échantillonnage aléatoire stratifié, où des échantillons indépendants de cas et de témoins sont tirés dans chaque strate, il existe des méthodes entièrement efficaces bien établies et faciles à appliquer. En particulier, si notre modèle comprend une ordonnée à l'origine distincte pour chaque strate, alors la régression logistique non pondérée ordinaire (avec un simple ajustement pour les ordonnées à l'origine de strate si l'on veut les obtenir) est la méthode semi-paramétrique efficace du maximum de vraisemblance (Prentice et Pyke 1979). Il est assez facile de l'étendre à des plans stratifiés plus généraux. Notre modèle est maintenant

$$\mathrm{logit}\{P(Y = 1 \mid \mathbf{x}, \text{ Strate h})\} = \beta_{0h} + \mathbf{x}^T \beta_1, \qquad (14)$$

et l'équivalent stratifié de l'équation d'estimation (7) est

$$\sum_h \left(\lambda_{1h} \frac{\sum_{\text{cas}} \mathbf{x}_i p_{0h}(\mathbf{x}_i; \boldsymbol{\beta})}{n_{1h}} - \lambda_{0h} \frac{\sum_{\text{témoins}} \mathbf{x}_i p_{1h}(\mathbf{x}_i; \boldsymbol{\beta})}{n_{0h}} \right) = \mathbf{0}. \qquad (15)$$

À mesure que $n_{0h}, n_{1h} \to \infty$, la solution de (7) converges presque certainement vers la solution de

$$\sum_h (\lambda_{1h} E_{1h}\{\mathbf{X}\, p_{0h}(\mathbf{X}; \boldsymbol{\beta})\} - \lambda_{0h} E_0\{\mathbf{X}\, p_{1h}(\mathbf{X}; \boldsymbol{\beta})\}) = \mathbf{0}, \qquad (16)$$

avec l'extension évidente de la notation utilisée pour le cas non stratifié. Si le modèle (13) est vérifié, alors l'équation (8) a pour solution $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_1$ et $\beta_{0h}^* = \beta_{0h} + b_{\lambda h}$ avec $b_{\lambda h} = \log(\lambda_{1h} W_{0h} / \lambda_{0h} W_{1h})$. Puisque l'équation (14) ne fait intervenir que des moyennes de strate, nous pouvons estimer facilement ces dernières en utilisant les données provenant de tout plan de sondage raisonnable, par exemple par

$$\hat{\boldsymbol{\mu}}_{\ell h}(\boldsymbol{\beta}) = \frac{\sum_{i \in S_{\ell h}} w_{ih} \mathbf{x}_{ih} (y_{ih} - p_1(\mathbf{x}_{ih}; \boldsymbol{\beta}))}{\sum_{i \in S_{\ell h}} w_{ih}}.$$

La substitution de ces estimateurs aux moyennes d'échantillon dans l'équation (14) donne l'équation d'estimation

$$\hat{\mathbf{S}}_\lambda(\boldsymbol{\beta}) = \sum_h \sum_{i \in S_h} w_{ih}^* \mathbf{x}_i (y_i - p_{1h}(\mathbf{x}_i; \boldsymbol{\beta})) = \mathbf{0}, \qquad (17)$$

avec $w_{ih}^* \propto \lambda_{\ell h} w_{ih} / \sum_{i \in S_{\ell h}} w_{ih}$ pour les unités comprises dans $S_{\ell h} (\ell = 0,1; h = 1, ..., H)$. Ce modèle peut être ajusté dans tout programme standard d'analyse de données d'enquête en introduisant ces poids et l'information appropriée sur le plan de sondage. Notons que nous devons faire attention à la façon dont nous incluons ce que nous appelons des « strates » dans la spécification du plan. Si les unités primaires d'échantillonnage sont emboîtées dans les « strates », comme c'est le cas des régions géographiques dans l'exemple 1, il n'y a pas de problème et les strates doivent être incluses de la façon standard. Toutefois, si les unités primaires d'échantillonnage recoupent les « strates », comme c'est le cas de l'âge dans l'exemple 1, et de l'âge et de l'ethnicité dans l'exemple 2, il ne s'agit plus de strates au sens habituel du terme en échantillonnage. Elles ne devraient pas être incluses dans les spécifications du plan, mais simplement être traitées par la pondération.

Parfois, nous voulons modéliser la contribution des variables de strate en utilisant une courbe paramétrique lisse au lieu de les inclure à l'aide de variables muettes. Par exemple, nous pourrions fort bien vouloir inclure une fonction linéaire de l'âge dans notre modèle, tant dans l'exemple 1 que dans l'exemple 2. La méthode de pondération par les poids de sondage et la pondération de compromis proposée à la section 6 s'appliquent l'une et l'autre, et aucun nouveau développement théorique n'est nécessaire. Par contre, les méthodes plus efficaces ne sont guères aussi simples. Des méthodes entièrement efficace ont été élaborées dans la situation où des échantillons aléatoire simple de cas et de témoins sont tirés dans chaque strate (voir Scott et Wild 1997, ainsi que Breslow et Holubkov 1997), mais les équations d'estimation résultantes ne sont pas des combinaisons linéaires des moyennes de strates, et il n'existe aucune manière évidente de les généraliser à des plans d'échantillonnage plus complexes. Néanmoins, il existe un moyen un peu moins efficace, mais facile à étendre. Si nous modifions le modèle (14) en incluant $b_{\lambda h} = \log(\lambda_{1h} W_{0h} / \lambda_{0h} W_{1h})$ comme correction, c'est-à-dire si nous supposons que

$$\text{logit}\{P^*(Y = 1 \mid \mathbf{x}, \text{ Strate h})\} = b_{\lambda h} + \beta_{0h} + x^T \beta_1, \qquad (18)$$

alors l'équation (15) produit des estimations convergentes, entièrement efficaces, de tous les coefficients, y compris $\beta_{0h} (h = 1, ..., H)$. L'introduction des mêmes corrections dans les modèles ne contenant pas de terme β_{0h} et où le vecteur $\mathbf{x}$ comprend des fonctions de la variable de stratification produit des estimateurs convergents pour tous les coefficients avec une efficacité habituellement élevée (quoi que non totale) (voir Fears et Brown 1986, ainsi que Breslow et Cain 1988). Cela se généralise à des plans de sondage arbitraires immédiatement. Il nous suffit d'utiliser l'équation (16) en remplaçant p_{1h} par p_{1h}^* défini en fixant $\text{logit}(p_{1h}^*) = b_{\lambda h} + \mathbf{x}^T \beta$. Alors, tout programme d'analyse de données d'enquête qui permet d'appliquer des corrections peut être utilisé pour ajuster le modèle et fournir des estimations des erreurs-types, *etc*.

Quel est notre gain d'efficacité dans ce cas-ci? Nous avons exécuté plusieurs simulations, dont certaines sont décrites dans Scott et Wild (2002). La plupart des scénarios sont fondés sur l'étude de la méningite de l'exemple 2 et nous fixons le ratio de la fraction d'échantillonnage de strate la plus grande à la plus faible dans l'échantillon de témoins à environ 10 pour 1. Sans aucune mise en grappes, le gain d'efficacité dû à l'utilisation de la méthode de correction (qui est le maximum de vraisemblance complète dans ce cas-ci) comparativement à la méthode *ad hoc* n'a jamais été supérieur à 10 %. Les efficacités relatives sont demeurées à peu près les mêmes lors de l'introduction d'une mise en grappes sur l'ensemble des strates. Quand nous sommes passés à la mise en grappes emboîtée dans les strates, les gains ont disparus progressivement à mesure que l'effet de plan augmentait et la méthode *ad hoc* est, en fait, devenue plus efficace que la méthode de correction, lorsque la valeur de l'effet de plan a atteint environ 1,5.

Comme nous l'avons mentionné plus haut, il est possible de produire des estimateurs semi-paramétriques entièrement efficaces si nous sommes prêts à modéliser la structure de dépendance à l'intérieur des unités primaires d'échantillonnage. Nous avons commencé à exécuter certaines simulations. Les premiers résultats donnent à penser que le travail supplémentaire que demande la modélisation ne vaudra presque jamais la peine si nous nous intéressons uniquement aux paramètres du modèle marginal (1). Notre conclusion provisoire est que les méthodes *ad hoc* partiellement pondérées (avec les poids d'échantillon) sont faciles à utiliser et donnent de suffisamment bons résultats pour la plupart des objectifs pratiques couverts par notre expérience, mais il s'agit toutefois d'un autre domaine où il conviendrait de poursuivre les travaux empiriques. Nous soulignons cependant que, pour certains problèmes, comme l'étude familiale cas-témoins dont il est question à la section 9, le comportement intragrappe est intéressant en soi. Il faut alors recourir à des méthodes plus perfectionnées.

8. Robustesse

Il doit y avoir un piège quelque part. Que se passe-t-il si le modèle est incorrect? Quel est alors le prix du gain d'efficacité?

Par construction, l'estimateur pondéré en fonction de la population estime toujours l'approximation logistique linéaire que nous obtiendrions si nous disposions de données pour l'ensemble de la population. Par contre, ce que l'estimateur plus efficace pondéré en fonction de l'échantillon estime dépend des tailles d'échantillons particulières utilisées. Certaines personnes considéreraient cet élément à lui seul comme une raison suffisamment valable d'utiliser l'estimateur pondéré d'après la population et je soupçonne que fort peu d'entre elles jugeraient entièrement satisfaisant que la cible de leur inférence dépende du choix arbitraire de la taille d'échantillon.

Notre estimateur général $\hat{\boldsymbol{\beta}}_\lambda$ satisfaisant (10) converge vers la solution de l'équation (9), disons $\mathbf{B}_\gamma$, avec $\gamma = \lambda_0/(\lambda_0 + \lambda_1)$, qui dépend du modèle réel et de la distribution des covariables, ainsi que de γ. Dans Scott et Wild (2002), nous avons examiné ce qui arrive à $\mathbf{B}_\gamma$ lors d'écarts faibles par rapport au modèle hypothétique (nous nous intéressons aux petits écarts, car en principe, les grands sont décelés par les procédures courantes de vérification des modèles qui devraient alors être améliorés en conséquence). Pour simplifier, supposons que nous ajustions un modèle linéaire ne contenant qu'une seule variable explicative pour le logarithme du rapport de cotes, mais que le modèle réel soit quadratique, disons

$$\text{logit}\{P(Y = 1 \mid x)\} = \beta_0 + \beta_1 x + \delta x^2 \qquad (19)$$

où δ est petit.

De toute évidence, la pente réelle de l'échelle logit, $\beta_1 + 2\delta x$, varie lorsque nous nous déplaçons le long de la courbe. Pour tout $0 < \gamma < 1$, $\mathbf{B}_{\gamma 1}$ est égal à la pente réelle à un point donné sur la courbe. Dénotons cette valeur par $x = x_\gamma$. Soit x_0 la valeur attendue de x dans la population de témoins et soit x_1 la valeur attendue de x dans la population de cas. Nous supposons que $\beta_1 > 0$, de sorte que $x_0 < x_1$. Il s'avère que x_γ est toujours compris entre x_0 et x_1, et que x_γ augmente quand la valeur de γ passe de 0 à 1. Rappelons que la pondération par les poids de sondage correspond à $\gamma = W_0$ et que la pondération par les poids d'échantillon correspond à $\gamma = \omega_0 = n_0/n$. Habituellement, W_0 est sensiblement plus grand que ω_0, de sorte que l'utilisation des poids de sondage donnent une estimation de la pente pour des valeurs plus grandes de x, où la probabilité d'un cas est plus élevée, tandis que la pente estimée d'après la pondération d'échantillon s'approche davantage de la valeur moyenne de x dans la population. La figure 1, adaptée de Scott et Wild (2002), illustre la position dans deux scénarios, l'un avec une courbure positive et l'autre, une courbure négative, basés approximativement sur l'exemple 2. Nous choisissons une valeur de δ telle que celui-ci serait décelé à l'aide d'un test standard du rapport de vraisemblance dans environ 50 % des cas si nous sélectionnions des échantillons aléatoires simples de taille $n_0 = n_1 = 200$ à partir de la population.

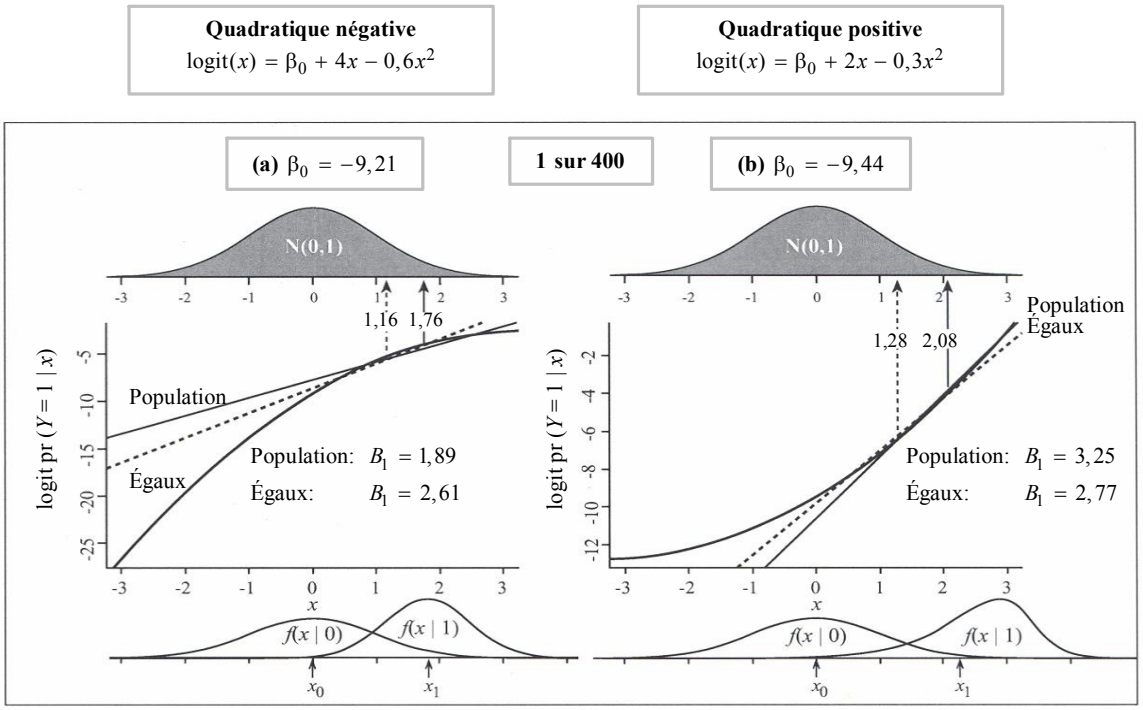

Figure 1. Comparaison entre les poids de population et les poids égaux.

Dans les deux scénarios, la valeur de β_0 est fixée de sorte que la proportion de cas dans la population soit de 1 sur 400, c'est-à-dire $W_0 = 0,9975$. La densité globale de x est représentée à la partie supérieure du graphique et les densités conditionnelles pour les cas et les témoins le sont au bas du graphique. Les valeurs de x_γ et $\mathbf{B}_{\gamma 1}$ sont données pour $\gamma = W_0$ (étiquetées « Population ») et $\gamma = 0,5$ (étiquetées « Égaux »). La seconde valeur correspond à la pondération d'échantillon si nous tirons des nombres égaux de cas et de témoins. Manifestement, dans les deux scénarios, la pondération d'échantillon produit une estimation de la pente appropriée pour des valeurs de x située plus à l'extrémité de la queue supérieure de la distribution (c'est-à-dire pour les personnes à haut risque) que dans le cas de la pondération égale.

Notons que, si nous sélectionnions des échantillons aléatoires simples de taille $n_0 = n_1 = 200$ à partir de la population de la figure 1 (a), l'efficacité relative de la pondération d'échantillon ne serait que d'environ 16 %, et le biais de petit échantillon serait de 0,24. Dans ce cas, même si nous prenions la valeur de population comme cible, la pondération par les poids de sondage produirait une erreur quadratique moyenne plus grande que la pondération d'échantillon.

Un plus grand nombre de résultats sont présentés dans Scott et Wild (2002), où nous examinons aussi l'effet des covariables omises. Celles-ci s'avèrent avoir un effet semblable, mais un peu plus faible, que l'omission d'un terme quadratique.

Quelle est la valeur de γ qu'il convient d'utiliser? Cela dépend clairement de l'utilisation que nous voulons faire du modèle résultant. Si notre principal intérêt est d'utiliser le modèle pour estimer des rapports de cotes à des valeurs de x où la probabilité d'un cas est élevée, et que l'échantillon est suffisamment grand pour que la variance et le biais de petit échantillon soient moins importants, nous pourrions utiliser les poids de population. Pour les tailles d'échantillon plus petites, ou si nous nous intéressons à des valeurs de x plus proche de la moyenne de population, les poids d'échantillon conviendraient mieux. Parfois, une valeur intermédiaire entre les poids de population et les poids d'échantillon pourrait représenter un compromis raisonnable. Par exemple, l'élagage des poids à 10 pour 1 (c'est-à-dire fixer $\gamma \approx 0,91$) dans l'exemple, au lieu de 1 pour 1 (pondération d'échantillon) ou 400 pour 1 (pondération de population) donne une efficacité de 70 % et un biais de petit échantillon de 0,04. Les valeurs correspondantes pour la pondération de population sont 16 % et 0,24. La valeur de $x_{0,91}$ est située presque exactement à mi-chemin entre $x_{0,5}$ et $x_{0,9975}$.

9. Études familiales cas-témoins

Si nous nous intéressons principalement aux paramètres du modèle marginal (1), alors les méthodes dont nous avons discuté aux sections précédentes sont faciles à appliquer et raisonnablement efficaces. L'utilisation de méthodes entièrement efficaces requiert la construction de modèles paramétriques de la dépendance intragrappe et l'effort supplémentaire que cela demande en vaut rarement la peine. Cependant, il existe des situations où la structure de dépendance présente un intérêt intrinsèque. En particulier, il est de plus en plus fréquent que les épidémiologistes généticiens étoffent les données d'une étude cas-témoins standard au moyen d'information sur les réponses et les covariables fournie par des membres de la famille, afin d'essayer d'obtenir des renseignements sur le rôle de la génétique et de l'environnement. Cette approche peut être considérée comme un échantillonnage en grappes stratifié, où les familles sont les grappes et, dans ce cas, la structure intragrappe est de toute première importance. L'exemple qui suit est assez typique.

Exemple 3

Wrensch, Lee, Miike, Newman, Barger, Davis, Wiencke et Neuhaus (1997) ont réalisé une étude cas-témoins sur la population du gliome, forme la plus fréquente de tumeur maligne du cerveau, dans la région de la baie de San Francisco. Ils ont recueilli des renseignements sur tous les cas de gliome diagnostiqués durant un intervalle de temps particulier et sur un échantillon comparable de témoins sélectionnés par la méthode de composition aléatoire. Ils ont également recueilli des renseignements sur la situation de tumeur du cerveau et sur les covariables auprès des membres de la famille des sujets sélectionnés dans l'échantillon cas-témoins original. L'étude portait sur 476 familles comptant un cas de tumeur du cerveau et 462 familles comptant un témoin.

Nous pourrions utiliser les méthodes dont nous venons de discuter pour ajuster un modèle marginal de la probabilité de devenir une victime du gliome, mais les chercheurs s'intéressaient avant tout à l'estimation des caractéristiques intrafamiliales. Une approche consisterait à ajuster un modèle logistique mixte comprenant un ou plusieurs effets familiaux aléatoires.

Notons que, strictement parlant, le plan d'échantillonnage de l'exemple 3 n'est pas compris dans ce plan d'étude cas-témoins. Ici, la stratification est reliée à la variable réponse, mais n'est pas entièrement déterminée par cette dernière. La strate 1 contient les 476 familles dans lesquelles un cas a été diagnostiqué durant un petit intervalle de temps déterminé, tandis que la strate 2 contient les 1 942 490 autres familles, dont certaines comprennent des victimes du cancer du cerveau.

Dans Neuhaus et coll. (2006), nous élaborons des méthodes semi-paramétriques efficaces pour l'échantillonnage stratifié à plusieurs degrés dans des situations où la stratification dépend de la réponse, éventuellement d'une façon non spécifiée qui doit être modélisée, et les observations dans une unité primaire d'échantillonnage sont reliées au moyen d'un modèle paramétrique. Le calcul des estimations requiert la résolution des $p+1$ équations d'estimation, où p est la dimension du vecteur de paramètres. La matrice de covariance peut aussi être estimée facilement en utilisant une analogue de l'inverse de la matrice d'information observée. La procédure complète peut être exécutée à l'aide d'une routine de maximisation raisonnablement générale, mais demande néanmoins une certaine expertise en calcul.

Nous pourrions aussi ajuster les mêmes modèles en utilisant des estimateurs pondérés par les poids de sondage, ce qui offre l'énorme avantage de ne nécessiter aucun logiciel spécialisé. Dans notre exemple, les familles comprenant un cas auraient un poids de 1 et les familles comprenant un témoin auraient un poids de $1\,942\,490/462 \approx 4\,200$. Étant donné cette grande différence, nous pourrions nous attendre à ce que les estimations pondérées soient très inefficaces. Malheureusement, il s'avère presque impossible d'ajuster un modèle intéressant pour lequel les estimations pondérées convergent. L'un des problèmes est que les estimations pondérées sont fondées presque entièrement sur l'échantillon de témoins et que l'on possède fort peu d'information au sujet des effets familiaux dans les familles de témoins (un autre problème est que nous ne possédons pas d'information sur l'âge des membres de la famille et que toute spécification du modèle sans la variable d'âge était exagérément incorrecte). Donc, nous avons dû recourir à une simulation, qui est loin d'être achevée à ce stade. Il semble cependant qu'ici l'efficacité des estimations pondérées soit inférieure à 10 % des estimations par la méthode semi-paramétrique du maximum de vraisemblance. Plus de détails sont donnés dans Neuhaus et coll. (2002, 2006).

Bien que nos simulations en soient à un stade très précoce, il est possible de tirer quelques conclusions provisoires. La principale est que les grandeurs intrafamiliales sont fort mal estimées, même en utilisant des méthodes entièrement efficaces. Les plans d'étude familiale cas-témoins, où l'information sur les membres de la famille est obtenue à titre de supplément à un plan cas-témoins standard, ne fournissent tout simplement pas suffisamment d'information pour estimer les paramètres qui intéressent les épidémiologistes généticiens, à moins que les associations soient extrêmement (voire déraisonnablement) fortes (il convient de souligner que tous les épidémiologistes généticiens ne sont pas d'accord sur ce point). L'utilisation de variantes plus efficaces est néanmoins possible. Ainsi, si nous pouvions identifier les familles contenant plus d'un cas, il serait alors possible d'atteindre une efficacité sensiblement plus grande en suréchantillonnant fortement ces familles. Essentiellement, nous considérerions la famille comme l'unité d'échantillonnage, définirions une « famille-cas » comme contenant plusieurs cas individuels, puis sélectionnerions un échantillon cas-témoins de familles. Il s'agit d'un domaine important où de nombreux travaux restent à accomplir.

10. Conclusion

L'étude cas-témoins sur la population est l'un des domaines où la pratique a devancé la théorie. Autant que je sache, le seul ouvrage où le sujet est abordé en profondeur est celui de Korn et Graubard (1999, chapitre 9). Un aspect auquel a été accordée une attention théorique raisonnablement importante dans la littérature est la stratification. Des méthodes efficaces en vue d'intégrer des variables de stratification dans l'analyse ont été élaborées, entre autres, par Scott et Wild (1997), Breslow et Holubkov (1997), ainsi que Lawless et coll. (1999), dans des circonstances où les variables peuvent prendre uniquement un ensemble fini de valeurs. Breslow et Chatterjee (1999) ont examiné le meilleur moyen d'utiliser ce genre d'information à l'étape de la conception de l'étude. L'extension de tous ces travaux (analyses ainsi que conception) à des situations où nous possédons de l'information sur des variables continues, comme l'âge, pour tous les membres de la population est un domaine où les travaux doivent se poursuivre. Bien que l'échantillonnage à plusieurs degrés soit d'usage répandu, l'effet de la mise en grappes a fait couler nettement moins d'encre. Font exception Graubard et coll. (1989), Fears et Gail (2000), ainsi que Scott et Wild (2001a). Le présent article suscitera peut-être d'autres travaux portant sur ce sujet important. En particulier, puisque le problème se résume essentiellement à l'estimation de deux moyennes de population (voir l'équation (8)), il devrait être possible d'appliquer une grande partie des connaissances sur les plans de sondage efficaces à la résolution de ce problème.

Remerciements

Je tiens à remercier les examinateurs, ainsi que Barry Graubard et Graham Kalton, dont la discussion réfléchie d'une version antérieure du présent article a fait progresser considérablement ma compréhension du sujet. Enfin, j'aimerais remercier tout spécialement mes collaborateurs de longue date Chris Wild, avec lequel ont été réalisés presque tous les travaux qui sous-tendent le présent article, et Jon Rao, auquel je dois essentiellement tout mon savoir sur l'analyse des données d'enquête.

Bibliographie

Baker, M., McNicholas, A., Garrett, N., Jones, N., Stewart, J., Koberstein, V. et Lennon, D. (2000). Household crowding: A major risk factor for epidemic meningococcal disease in Auckland children. *Pediatric Infectious Disease Journal*, 19, 983-990

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

Breslow, N.E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, 91, 14-28.

Breslow, N.E. (2004). Case-control studies. Dans *Handbook of Epidemiology*. (Éds. W. Aherns et I. Pigeot). New York : Springer. 287-319.

Breslow, N.E., et Cain, K.C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75, 11-20.

Breslow, N.E., et Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. *Applied Statistics*, 48, 457-468.

Breslow, N.E., et Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase outcome-dependent sampling. *Journal of the Royal Statistical Society*, B, 59, 447-461.

Brogan, D.J., Denniston, M.M., Liff, J.M., Flagg, E.W., Coates, R.J. et Brinton, L.A. (2001). Comparison of telephone sampling and area sampling: Response rates and within-household coverage. *American Journal of Epidemiology*, 153, 1119-1127.

Cosslett, S.R. (1981). Maximum likelihood estimation for choice-based samples. *Econometrica*, 49, 1289-1316.

DiGaetano, R., et Waksberg, J. (2002). Trade-offs in the development of a sample design for case-control studies. *American Journal of Epidemiology*, 155, 771-775.

Fears, T.R., et Brown, C.C. (1986). Logistic regression models for retrospective case-control studies using complex sampling procedures. *Biometrics*, 42, 955-960.

Fears, T.R., et Gail, M.H. (2000). Analysis of a two-stage case-control study with cluster sampling of controls: Application to nonmelanoma skin cancer. *Biometrics*, 56, 190-198.

Graubard, B.I., Fears, T.R. et Gail, M.H. (1989). Effects of cluster sampling on epidemiologic analysis in population-based case-control sampling. *Biometrics*, 45, 1053-1071.

Hartge, P., Brinton, L.A., Rosenthal, J.F., Cahill, J.I., Hoover, R.N. et Waksberg, J. (1984). Random digit dialing in selecting a population-based control group. *American Journal of Epidemiology*, 120, 825-833.

Hartge, P., Brinton, L.A., Cahill, J.I., West, D., Hauk, M., Austin, D., Silverman, D. et Hoover, R.N. (1984). Design and methods in a multi-center case-control interview study. *American Journal of Public Health*, 74, 52-56.

Korn, E.L., et Graubard, B.I. (1999). *Analysis of Health Surveys*. New York : John Wiley & Sons, Inc.

Lawless, J.F., Kalbfleisch, J.D. et Wild, C.J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society*, B, 61, 413-38.

Lee, A.J., Scott, A.J. et Wild, C.J. (2006). Fitting binary regression models with case-augmented samples. *Biometrika*, 95 (A paraître).

Manski, C.F., et McFadden, D. (Éds) (1981). *Structural Analysis of Discrete Data with Econometric Applications*. New York : John Wiley & Sons, Inc.

Miettinen, O.S. (1985). The case-control study: Valid selection of subjects. *American Journal of Epidemiology*, 135, 1042-1050.

Prentice, R.L., et Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403-411.

Neuhaus, J., Scott, A.J. et Wild, C.J. (2002). The analysis of retrospective family studies. *Biometrika*, 89, 23-37.

Neuhaus, J., Scott, A.J. et Wild, C.J. (2006). Family-specific approaches to the analysis of retrospective family data. *Biometrics*, 62, sous presse.

Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.

Rao, J.N.K., Scott, A.J. et Skinner, C.J. (1998). Quasi-score tests with survey data. *Statistica Sinica*, 8, 1059-1070.

Scott, A.J., et Wild, C.J. (1986). Fitting logistic models under case-control or choice-based sampling. *Journal of the Royal Statistical Society*, B, 48, 170-182.

Scott, A.J., et Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 83, 57-72.

Scott, A.J., et Wild, C.J. (2001a). The analysis of clustered case-control studies. *Applied Statistics*, 50, 57-71.

Scott, A.J., et Wild, C.J. (2001b). Fitting regression models to case-control data by maximum likelihood. *Journal of Statistical Planning and Inference*, 96, 3-27.

Scott, A.J., et Wild, C.J. (2002). On the robustness of weighted methods for fitting model to case-control data by maximum likelihood. *Journal of the Royal Statistical Society*, B, 64, 207-220.

Wacholder, S., McLaughlin, J.K., Silverman, D.T. et Mandel, J.S. (1991). Selection of controls in case-control studies. I. Principles. *American Journal of Epidemiology*, 135, 1019-1028.

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

Waksberg, J. (1998). Random digit dialing sampling for case-control studies. Dans *Encyclopedia of Biostatistics*. (Éds. P.Armitage et T. Colton). New York : John Wiley & Sons, Inc., 3678-3682.

Wrensch, M., Lee, M., Miike, R., Newman, B., Barger, G., Davis, R., Wiencke, J. et Neuhaus, J. (1997). Familial and personal medical history of cancer and nervous system conditions among adults with glioma and controls. *American Journal of Epidemiology*, 145, 581-93.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 149-160
Statistique Canada, N° 12-001 au catalogue

Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture

Phillip S. Kott [1]

Résumé

La pondération par calage peut être utilisée pour corriger la non-réponse totale et (ou) les erreurs de couverture sous des modèles appropriés de quasi-randomisation. Divers ajustements par calage qui sont asymptotiquement identiques dans un contexte d'échantillonnage pur peuvent diverger lorsqu'ils sont utilisés de cette manière. L'introduction de variables instrumentales dans la pondération par calage permet que la non-réponse (disons) soit une fonction d'un ensemble de caractéristiques différentes de celles comprises dans le vecteur de calage. Si l'ajustement par calage a une forme non linéaire, une variante du jackknife permet d'éliminer le besoin d'itération dans l'estimation de la variance.

Mots clés : Modèle prédictif; modèle de quasi-randomisation; convergent sous quasi-randomisation; variable instrumentale; ajustement proportionnel itératif (raking) généralisé.

1. Introduction

La méthode de pondération par calage a été mise au point au départ en vue de réduire les erreurs d'échantillonnage en maintenant la convergence sous randomisation. Deville et Särndal (1992) ont démontré que de nombreuses formes de pondération par calage sont asymptotiquement identiques dans le contexte de l'échantillonnage, ce qui a fait progresser grandement notre compréhension des méthodes courantes de repondération, telles que la méthode itérative du quotient aussi appelée l'ajustement proportionnel itératif (API, "raking" en anglais), qui ne se trouve pas sous le format de l'estimateur par la régression généralisée (GREG).

Folsom et Singh (2000) ont montré que la pondération par calage peut aussi être utilisée pour corriger les erreurs connues de couverture et (ou) la non-réponse totale sous des modèles appropriés de quasi-randomisation. Ces travaux n'ont été publiés dans aucune revue avec comité de lecture. Le cœur du présent article est une répétition des principaux résultats publiés dans Folsom et Singh, y compris une modification nécessaire de l'approche de Deville-Särndal en vue de modéliser l'estimation de la variance ou de l'erreur quadratique moyenne sous randomisation dans ce contexte élargi. Une version antérieure, strictement linéaire, de la pondération par calage pour l'ajustement pour la non-réponse totale peut être consultée dans Fuller, Loughin et Baker (1994). Voir aussi Lundström et Särndal (1999).

Nous faisons une distinction entre le modèle prédictif qui sous-tend habituellement le calage et le modèle de quasi-randomisation de Folsom et Singh. Toutefois, contrairement à ces deux auteurs, nous examinons ici les propriétés dans les deux cas. En outre, les variables explicatives du modèle de quasi-randomisation peuvent différer des variables de calage, ce qui est également permis dans Lundström et Särndal.

Nous proposons un nouveau jackknife qui est analogue à l'estimateur de la variance par linéarisation de Deville-Särndal. Il repose sur l'utilisation de poids de rééchantillonnage calculés en une étape, quoique les poids de calage proprement dits puissent être déterminés itérativement.

Après la présentation de la notion bien connue de pondération par calage, à la section 2, nous passons en revue le cas particulier de l'estimateur GREG dans un contexte d'échantillonnage pur. À la section 3, nous décrivons l'extension de la pondération par calage d'Estevao et Särndal (2000) dans sa forme linéaire, afin d'inclure des variables instrumentales. À la section 4, nous étendons le traitement de la pondération par calage de Deville et Särndal, afin d'inclure la possibilité de variables instrumentales. À la section 5, nous passons en revue l'estimation de la variance ou de l'erreur quadratique moyenne, et proposons un nouveau jackknife pour certains plans d'échantillonnage. À la section 6, nous décrivons comment la pondération par calage peut être utilisée pour la correction de la non-réponse. Dans ce contexte, les diverses formes fonctionnelles de la pondération par calage ne doivent plus nécessairement être asymptotiquement identiques. À la section 7, nous discutons des modèles de quasirandomisation pour les erreurs de couverture, c'est-à-dire le sous- ou le surdénombrement dans la base de sondage. À la section 8, nous donnons un petit exemple empirique appuyant le nouveau jackknife. Enfin, à la section 9, nous présentons une discussion des diverses approches et des domaines dans lesquels les travaux de recherche doivent se poursuivre.

1. Phillip S. Kott, National Agricultural Statistics Service, 3251 Old Lee Highway, Fairfax, VA 22030, États-Unis. Courriel : pkott@nass.usda.gov.

2. Pondération par calage et l'estimateur GREG

Supposons que nous connaissions la probabilité de sélection, π_k, pour chaque élément d'échantillonnage k dans l'échantillon S. Nous pouvons estimer tout total de population, $T_y = \sum_U y_k$, où U dénote la population, au moyen de l'estimateur à facteur d'extension $t_{y\_E} = \sum_S y_k/\pi_k = \sum_U y_k I_k/\pi_k$, où $I_k = 1$ quand $k \in S$ et 0 autrement. En traitant les I_k comme des variables aléatoires, il est facile de voir que t_{y_E} est un estimateur sans biais de T_y. Les propriétés qui découlent du fait que les I_k sont traitées comme des variables aléatoires sont dites *fondées sur la randomisation*. Nous pouvons également écrire $t_{y\_E} = \sum_U a_k y_k = \sum_S a_k y_k$, où $a_k = I_k/\pi_k$ est le *poids d'échantillonnage* de l'élément k.

Deville et Särndal (1992) ont inventé l'expression « estimateur par calage » pour décrire un estimateur de la forme $t_{y\_\mathrm{CAL}} = \sum_S w_k y_k$, où $\sum_S w_k \mathbf{x}_k = \sum_U \mathbf{x}_k = T_\mathbf{x}$ pour un certain vecteur ligne de variables auxiliaires, $\mathbf{x}_k = (x_{1k}, \ldots, x_{Pk})$, au sujet duquel $T_\mathbf{x}$ est connu. Puisqu'il existe généralement un continuum d'ensembles $\{w_k | k \in S\}$ qui satisfont l'*équation de calage* :

$$\sum_{k\in S} w_k \mathbf{x}_k = T_\mathbf{x}, \tag{1}$$

Deville et Särndal ont imposé comme condition que la différence entre l'ensemble de poids, $\{w_k | k \in S\}$, satisfaisant l'équation (1) et $\{a_k | k \in S\}$ minimise une fonction de perte.

Une autre approche de l'échantillonnage consiste à traiter les y_k comme des variables aléatoires satisfaisant le modèle prédictif linéaire :

$$y_k = \mathbf{x}_k \boldsymbol{\beta} + \varepsilon_k, \tag{2}$$

où $E(\varepsilon_k | \{\mathbf{x}_g, I_g | g \in U\}) = 0$ pour tout $k \in U$. En conditionnant cette espérance sur les I_g, nous supposons que l'on peut ignorer le mécanisme d'échantillonnage. Il s'agit d'un aspect critique, et parfois déraisonnable, du cadre (prédictif) *fondé sur un modèle*.

Il est facile de voir que t_{y_CAL} est un estimateur sans biais de T_y sous le modèle en ce sens que $E_\varepsilon(t_{y\_\mathrm{CAL}} - T_y) = 0$ (en supprimant le conditionnement pour simplifier la notation); l'indice ε fait référence au traitement des ε_k comme des variables aléatoires (et des I_k comme des constantes prédéterminées).

Aux fins de notre étude, l'estimateur par la régression généralisée ou GREG a la forme :

$$t_{y\_\mathrm{GREG}} = t_{y\_E} + \left(T_\mathbf{x} - \sum_{k\in S} a_k \mathbf{x}_k\right)\left(\sum_{k\in S} c_k a_k \mathbf{x}_k' \mathbf{x}_k\right)^{-1} \sum_{k\in S} c_k a_k \mathbf{x}_k' y_k, \tag{3}$$

où c_k est une constante arbitraire qui peut ou non être une fonction de $\mathbf{x}_k$, et $\lim_{N\to\infty} \sum_U c_k \mathbf{x}_k' \mathbf{x}_k / N = \boldsymbol{\Lambda}$ est une matrice définie positive, où N est la taille de U. Cette dernière condition signifie que $\sum_S c_k a_k \mathbf{x}_k' \mathbf{x}_k$ sera habituellement inversible en pratique. Par souci de commodité, nous supposerons qu'elle l'est toujours.

L'estimateur GREG de l'équation (3) peut être réécrit sous une forme de calage comme étant $t_{y\_\mathrm{GREG}} = \sum_S w_k y_k$, où

$$w_k = a_k + \left(T_\mathbf{x} - \sum_{j\in S} a_j \mathbf{x}_j\right)\left(\sum_{j\in S} c_j a_j \mathbf{x}_j' \mathbf{x}_j\right)^{-1} c_k a_k \mathbf{x}_k'.$$

Strictement parlant, les w_k sont des fonctions de l'échantillon réalisé, S, et des $c_k a_k$, mais nous supprimons cela dans la notation pour simplifier. Observons que les poids de calage peuvent être exprimés sous la forme

$$w_k = a_k(1 + c_k \mathbf{x}_k \mathbf{q}), \tag{4}$$

où $\mathbf{q} = [(\sum_S a_j c_j \mathbf{x}_j' \mathbf{x}_j)^{-1}]'(T_\mathbf{x} - \sum_S a_j \mathbf{x}_j)'$ est un vecteur colonne, puisque $\mathbf{x}_k \mathbf{q} = \mathbf{q}' \mathbf{x}_k'$.

Supposons que des conditions de régularité raisonnables soient vérifiées (voir, par exemple, Kott 2004a pour un traitement plus approfondi) et que le plan d'échantillonnage est tel que $t_{y\_E} - T_y = O_P(N/\sqrt{n})$, où n est la taille prévue de S (la taille réelle peut être aléatoire), $\sum_S a_k \mathbf{x}_k - T_\mathbf{x} = \mathbf{O}_P(N/\sqrt{n})$, et $\sum_S a_k c_k \mathbf{x}_k' \mathbf{f}_k - \sum_U c_k \mathbf{x}_k' \mathbf{f}_k = \mathbf{O}_P(N/\sqrt{n})$, où $\mathbf{f}_k$ peut être $\mathbf{x}_k$ ou y_k. Soit $e_k = y_k - \mathbf{x}_k (\sum_U c_i \mathbf{x}_i' \mathbf{x}_i)^{-1} \sum_U c_i \mathbf{x}_i' y_i$, de sorte que $\sum_U c_i \mathbf{x}_i' e_i = 0$, et $\sum_S a_k c_k \mathbf{x}_k' e_k = \mathbf{O}_P(N/\sqrt{n})$. Nous pouvons exprimer l'erreur de t_{y_GREG} sous la forme

$$\begin{aligned}
&t_{y_\mathrm{GREG}} - T_y \\
&= \sum_{k\in S} w_k y_k - \sum_{k\in U} y_k \\
&= \sum_{k\in S} w_k e_k - \sum_{k\in U} e_k \quad \left(\text{car } \sum_{k\in S} w_k \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k\right) \\
&= \sum_{k\in S} a_k e_k + \left(T_\mathbf{x} - \sum_{k\in S} a_k \mathbf{x}_k\right)\left(\sum_{k\in S} a_k c_k \mathbf{x}_k' \mathbf{x}_k\right)^{-1} \sum_{k\in S} a_k c_k \mathbf{x}_k' e_k \\
&\qquad - \sum_{k\in U} e_k \\
&= \sum_{k\in S} a_k e_k - \sum_{k\in U} e_k + O_P(N/n).
\end{aligned} \tag{5}$$

Il est maintenant aisé de voir que l'estimateur GREG est convergent sous randomisation; autrement dit, $p\lim_{n\to\infty} [(t_{y\_\mathrm{GREG}} - T_y)/N] = 0$. En outre, le biais relatif et l'erreur quadratique moyenne relative de l'estimateur GREG dans les conditions de randomisation sont d'ordre $1/n$. Puisque l'erreur quadratique moyenne = biais^2 + variance, nous pouvons conclure que le biais sous randomisation de l'estimateur GREG est habituellement un contributeur asymptotiquement non significatif à l'erreur quadratique moyenne de cet estimateur.

3. Redéfinition des poids de calage

Dans leur définition originale des poids de calage, Deville et Särndal (1992) posaient comme condition que l'ensemble des poids de calage, $\{w_k | k \in S\}$, minimisent une certaine fonction de distance entre les membres de l'ensemble et les poids d'échantillonnage originaux, les a_k, sous la contrainte qu'ils satisfassent l'équation de calage. Par conséquent, l'estimateur par calage, $t_{y\_\mathrm{CAL}} = \sum_S w_k y_k$, était à la fois sans biais sous le modèle donné par l'équation (2) et habituellement convergent sous randomisation.

Estevao et Särndal (2002) ont proposé d'éliminer l'exigence que les poids de calage minimisent une fonction de distance. À la place, ils ont essentiellement proposé que les w_k soient seulement obligés de satisfaire l'équation de calage et d'avoir la « forme fonctionnelle » suivante :

$$w_k = a_k(1 + \mathbf{h}_k \mathbf{q}), \qquad (6)$$

où $\mathbf{h}_k$ est un vecteur ligne de même dimension que $\mathbf{x}_k$, tel que $\sum_S a_k \mathbf{h}_k' \mathbf{x}_k$ est inversible, et $\mathbf{q}$ est un vecteur colonne de même dimension. L'équation (6) est une généralisation faible de (4), où $\mathbf{h}_k$ remplace effectivement $c_k \mathbf{x}_k$.

Il n'est pas difficile de voir que $\mathbf{q} = [(\sum_S a_j \mathbf{h}_j' \mathbf{x}_j)^{-1}]' (T_\mathbf{x} - \sum_S a_j \mathbf{x}_j)'$. En outre, sous des contraintes faibles que nous supposons vérifiées, $t_{y\_\mathrm{CAL}} = \sum_S w_k y_k = \sum_S a_k y_k + (T_\mathbf{x} - \sum_S a_j \mathbf{x}_j)(\sum_S a_j \mathbf{h}_j' \mathbf{x}_j)^{-1} \sum_S a_k \mathbf{h}_k' y_k$ est convergent sous randomisation quand t_{y_E} l'est. Il est sans biais sous le modèle prédictif linéaire donné par l'équation (2) quand $E(\varepsilon_k | \{\mathbf{x}_g, \mathbf{h}_g | g \in S\}, \{I_g | g \in U\}) = 0$ pour tout $k \in U$.

Cela suggère une définition de rechange des poids de calage : un ensemble de poids, $\{w_k | k \in S\}$, tel que

i. les w_k satisfassent l'équation de calage pour $\{\mathbf{x}_k | k \in U\}$ et,

ii. $t_{y\_\mathrm{CAL}} = \sum_S w_k y_k$ soit convergent sous randomisation quand t_{y_E} l'est sous des contraintes faibles.

Cette définition est celle que nous utiliserons. Cette définition élargie de la pondération par calage s'avérera fort utile lors du calage pour la correction de la non-réponse ou des erreurs de couverture.

Il découle de notre nouvelle définition que le calage à forme fonctionnelle d'Estevao et Särdnal est, en réalité, une forme de pondération par calage. En nous inspirant de la théorie économétrique, nous donnons aux composantes de $\mathbf{h}_k$ qui ne sont pas des combinaisons linéaires des composantes de $\mathbf{x}_k$ le nom de « variables instrumentales ».

4. Calage éventuellement non linéaire

En partant des idées de Deville et Särndal (1992), nous pouvons généraliser la forme linéaire des poids de calage donnée par l'équation (6) à

$$w_{k\_\mathrm{GEN}} = a_k f(\mathbf{h}_k \mathbf{q}^*), \qquad (7)$$

où f est une fonction monotone, dérivable deux fois avec $f(0) = 1, f'(0) = 1$ ($f'(0)$ est la dérivée première de f évaluée à 0) et $\mathbf{q}^*$ est choisi de sorte que l'équation de calage soit vérifiée. Contrairement à l'équation des poids de calage susmentionnés, l'équation de calage proprement dite, $\sum_S w_k \mathbf{x}_k = T_\mathbf{x}$, demeure linéaire. Notons que, puisque $f(0) = 1, f'(0) = 1, f(\mathbf{h}_k \mathbf{q}^*) \approx 1 + \mathbf{h}_k \mathbf{q}^*$.

Strictement parlant, nous devrions utiliser un symbole supplémentaire sur w_{k_GEN} (et plus tard sur w_{k_LIN}) pour dénoter le choix particulier de $\mathbf{h}_k$. Nous l'avons laissé tomber pour simplifier.

Une solution, $\mathbf{q}^*$, de l'équation (7) peut souvent être obtenue de façon itérative. On peut partir de $\mathbf{q}^{(0)} = \mathbf{0}$; c'est-à-dire $\sum_S w_k^{(0)} y_k$, où $w_k^{(0)} = a_k f(0)$. Pour $r = 1, 2, \ldots$, on fixe alors $\mathbf{q}^{(r)} = \mathbf{q}^{(r-1)} + \{[\sum_s f'(\mathbf{h}_k \mathbf{q}^{(r-1)}) a_k \mathbf{x}_k' \mathbf{h}_k]^{-1}\}' \times (T_\mathbf{x} - \sum_S w_k^{(r-1)} \mathbf{x}_k)'$, et $w_k^{(r)} = a_k f(\mathbf{h}_k \mathbf{q}^{(r)})$. L'itération s'arrête à r^* quand $T_\mathbf{x} = \sum_S w_k^{(r^*)} \mathbf{x}_k$, à toutes fins utiles. Cependant, il faut se souvenir qu'*il pourrait ne pas exister d'ensemble de poids pouvant être exprimé sous la forme de l'équation* (7) *tout en satisfaisant l'équation de calage.*

Soulignons que $\mathbf{q}^{(1)}$ susmentionné est égal à $\mathbf{q}$ dans $w_{k\_\mathrm{LIN}} = a_k (1 + \mathbf{h}_k \mathbf{q})$. Un développement en série de Taylor autour de zéro révèle $f(\mathbf{h}_k \mathbf{q}^{(1)}) = 1 + \mathbf{h}_k \mathbf{q}^{(1)} + O_P(1/n)$ sous des contraintes faibles, de sorte que $\sum_S w_k^{(1)} y_k = \sum_S w_{k\_\mathrm{LIN}} y_k + O_P(N/n) = T_y [1 + O_P(1/n)]$. En outre, il n'est pas difficile de voir que $w_{k\_\mathrm{GEN}} = w_{k\_\mathrm{LIN}}[1 + O_P(1/n)]$, une égalité qui s'avère utile dans l'estimation de la variance.

L'exemple le plus courant en pratique d'une fonction f non linéaire est $f(\mathbf{h}_k \mathbf{q}) = \exp(\mathbf{x}_k \mathbf{q})$, où les valeurs de chaque composante de $\mathbf{x}_k$, dénotées $x_{1k}, \ldots, x_{Pk}$, sont 0 ou 1. Cela est effectivement la forme des poids de calage sur marges (API) de Deming et Stephan (1940) calculés par ajustement proportionnel itératif. De nombreux auteurs ont constaté que la routine itérative décrite plus haut peut être utilisée même si les composantes de $\mathbf{x}_k$ ne sont pas binaires, comme elles le sont dans Deming et Stephan. Soulignons que les poids de calage par *raking généralisé* résultants sont systématiquement non négatifs.

5. Estimation de la variance

Särndal, Swensson et Wretman (1989) ont proposé cet estimateur de la variance ou de l'erreur quadratique moyenne sous randomisation à modèle *plug-in* pour t_{y_GREG} sous un plan d'échantillonnage arbitraire :

$$v_{\mathrm{SSW}} = \sum_{k\in S}\sum_{j\in S}[(\pi_{kj}-\pi_k\pi_j)/\pi_{kj}](w_k r_k)(w_j r_j). \quad (8)$$

Le terme *plug-in* vient du fait que r_k est « introduit dans » (plugged into) $v_{\mathbf{SSW}}$ à la place des $e_k = y_k - \mathbf{x}_k(\sum_U \mathbf{h}_i'\mathbf{x}_i)^{-1}\sum_U \mathbf{h}_i' y_i$ inconnus pour l'estimation de l'erreur quadratique moyenne sous randomisation.

En utilisant des arguments parallèles à ceux de Deville et Särndal (1992), $v_{\mathbf{SSW}}$ s'applique aussi de façon générale à t_{y_CAL} avec les poids de calage définis par l'équation (7) avec

$$r_k = y_k - \mathbf{x}_k\left(\sum_{j\in S} a_j\mathbf{h}_j'\mathbf{x}_j\right)^{-1}\sum_{j\in S} a_j\mathbf{h}_j'\mathbf{x}_j. \quad (9)$$

Il en est ainsi parce que $w_{k\_\mathrm{GEN}} = w_{k\_\mathrm{LIN}}[1+O_P(1/n)]$, donc $\sum_S w_{k\_\mathrm{GEN}}\, e_k = \sum_S w_{k\_\mathrm{LIN}}\, e_k + O_P(N/n) = \sum_S a_k e_k + O_P(N/n)$. La dernière étape s'appuie sur le raisonnement exprimé dans l'équation (5), avec $\mathbf{h}_j$ remplaçant les $c_j\mathbf{x}_j$.

Dans leur article, Deville et Särndal ont, en réalité, remplacé les a_j de l'équation (9) par $w_j = a_j f(\mathbf{h}_j\mathbf{q}^*)$. Une version différente est donnée dans Demanti et Rao (2004), où les a_j de l'équation sont remplacés par $a_j f'(\mathbf{h}_j\mathbf{q}^*)$. Ces auteurs soulignent dans un commentaire accompagnant cette dernière expression que les trois versions des r_k sont asymptotiquement identiques puisque $f(0) = f'(0) = 1$ et $\mathbf{q}^*$ est asymptotiquement égal à $\mathbf{0}$. Ces identités asymptotiques pourraient ne plus être vérifiées lorsque la pondération par calage est utilisée pour corriger pour la non-réponse, comme nous le verrons à la section suivante.

L'établissement de propriétés asymptotiques pour $v_{\mathbf{SSW}}$ sous échantillonnage aléatoire simple stratifié est un exercice simple. Dans le présent contexte, $v_{\mathbf{SSW}}$ se réduit à

$$v_{\mathrm{ST1}} = \sum_{\alpha=1}^{A}(n_\alpha/[n_\alpha-1])\sum_{k\in S_\alpha}(1-n_\alpha/N_\alpha) \times \left(w_k r_k - \sum_{j\in S_\alpha} w_j r_j/n_\alpha\right)^2,$$

où S_α dénote l'échantillon de n_α unités dans la strate $\alpha(\alpha = 1,\ldots,A)$, et U_α, la population de la strate contenant N_α éléments.

Dans le cas d'un échantillon à plusieurs degrés, il est logique de permettre que, dans le modèle prédictif, ε_k et ε_j soient corrélés quand k et j sont des éléments de la même UPE, mais autrement pas. Si l'on peut ignorer la correction pour population finie, la variance fondée sur un modèle d'un estimateur par calage est approximativement $V_m = \sum_{i\in S'} E_\varepsilon[(\sum_{k\in S(i)} w_k\varepsilon_k)^2]$ sous des contraintes faibles, où $S(i)$ est l'ensemble d'éléments échantillonnés dans l'UPE i, et S' est l'ensemble d'UPE sélectionnées au premier degré de l'échantillonnage.

L'estimateur de la variance qui suit, qui n'est pas strictement égal à $v_{\mathbf{SSW}}$, a souvent de bonnes propriétés sous randomisation et sous un modèle (lorsque les probabilités de sélection de premier degré sont toutes faibles) :

$$v_{\mathrm{ST2}} = \sum_{\alpha=1}^{A}(n_\alpha/[n_\alpha-1]) \times \left\{\sum_{j\in S_\alpha} - \left(\sum_{k\in S_{\alpha j}} w_k r_k\right)^2 \frac{\left(\sum_{j\in S_\alpha}\sum_{k\in S_{\alpha j}} w_k r_k\right)^2}{n_\alpha}\right\}, \quad (10)$$

où α dénote une strate d'UPE de premier degré, $n_{1\alpha}$ est le nombre d'UPE échantillonnées dans la strate α, S_α est l'ensemble d'UPE échantillonnées dans α, et $S_{\alpha j}$ est l'ensemble d'éléments sous-échantillonnés dans l'UPE j de la strate α. Le nombre de degrés d'échantillonnage peut être élevé.

Il n'est pas difficile de montrer que v_{ST2} est asymptotiquement indistinguable de l'estimateur de la variance par le jackknife :

$$v_J = \sum_{\alpha=1}^{A}([n_\alpha-1]/n_\alpha)\left\{\sum_{j\in S_\alpha}(t_{y\_\mathrm{CAL}(\alpha j)} - t_{y\_\mathrm{CAL}})^2\right\}, \quad (11)$$

où $t_{y\_\mathrm{CAL}(\alpha j)} = \sum_{k\in S} w_{k(\alpha j)} y_k$, et les *poids de calage par rééchantillonnage jackknife* sont

$$w_{k(\alpha j)} = w_k a_{k(\alpha j)}/a_k + \left(\sum_{m\in U}\mathbf{x}_m - \sum_{m\in S} w_m[a_{m(\alpha j)}/a_m]\mathbf{x}_m\right) \times \left(\sum_{m\in S} a_{m(\alpha j)}\mathbf{h}_m'\mathbf{x}_m\right)^{-1} a_{k(\alpha j)}\mathbf{h}_k', \quad (12)$$

où $a_{k(\alpha j)} = 0$ quand k est dans l'UPE j de la strate α, $a_{k(\alpha j)} = a_k$ quand k n'est pas dans la strate α du tout, et $a_{k(\alpha j)} = (n_\alpha/[n_\alpha-1])a_k$ autrement. Les $w_{k(\alpha j)}$ sont contraints de telle sorte que $\sum_{k\in S} w_{k(\alpha j)}\,\mathbf{x}_k = \sum_{k\in U}\mathbf{x}_k$ pour tout αj.

Soit $S(\alpha+)$ l'ensemble d'*éléments* dans la strate α (et non les UPE comme S_α) et $S(\alpha j)$, l'ensemble d'éléments dans l'UPE j de la strate α. Sous des contraintes faibles que nous supposons vérifiées,

$$\sum_U \mathbf{x}_m - \sum_S w_m[a_{m(\alpha j)}/a_m]\mathbf{x}_m = (n_\alpha/[n_\alpha-])\left(\sum_{S(\alpha j)} w_k\mathbf{x}_k - \sum_{S(\alpha+)} w_k\mathbf{x}_k/n_\alpha\right) = \mathbf{O}_P(N/n), \sum_S a_{m(\alpha j)}\mathbf{h}_m'\mathbf{x}_m = \mathbf{O}_P(N),$$

et

$$\sum_S a_{m(\alpha j)}\mathbf{h}'_m e_m = \mathbf{O}_P(N/\sqrt{n}).$$

Par conséquent,

$$t_{y\_\mathrm{CAL}} - t_{y\_\mathrm{CAL}} = \sum_s w_{k(\alpha j)}e_k - \sum_s w_k e_k$$
$$= (n_\alpha/[n_\alpha - 1])\left(\sum_{s(\alpha+)} w_k e_k / n_\alpha - \sum_{S(\alpha j)} w_k e_k\right)$$
$$+ O_P(N/n^{3/2}),$$

et $v_J = v_{\mathrm{ST2}}[1 + O_P(1/\sqrt{n})]$ quand $p\lim_{n\to\infty}(nv_{\mathrm{ST2}}/N^2) > 0$.

Les poids de rééchantillonnage définis par l'équation (12) ne nécessitent pas d'itération même si les poids de calage sont eux-mêmes produits de cette façon, ce qui est fort intéressant du point de vue informatique. Cela permet non seulement d'économiser du temps d'ordinateur, mais aussi d'éviter qu'une solution itérative puisse exister pour les w_k, mais non pour les poids de rééchantillonnage.

6. Non-réponse totale

6.1 Modèle de quasi-randomisation et modèle prédictif

À la présente section, nous examinons le traitement de la non-réponse totale (unité complète) en tant que phase supplémentaire de l'échantillonnage de Poisson. Il s'agit essentiellement d'un modèle de *quasi-randomisation*. Nous supposons que chaque élément k de l'échantillon original, maintenant dénoté F, a une probabilité de réponse, p_k. La probabilité que les éléments k et j répondent conjointement est $p_k p_j$, et le fait que l'élément k réponde (sachant un vecteur de covariables) est indépendant du fait qu'il soit choisi à partir de l'échantillon original.

Il est souvent possible de construire un ensemble de poids tel que l'estimateur par calage soit convergent par rapport au plan d'échantillonnage sous le modèle de quasi-randomisation. Nous recherchons ici un moyen particulier de construire ces poids. Pour cela, nous supposons que le modèle de quasi-randomisation est correct. Chaque élément est relié à un vecteur ligne de variables auxiliaires, $\mathbf{x}_k$, pour lequel $T_{\mathbf{x}} = \sum_U \mathbf{x}_j$ est connu. Enfin, nous supposons que chaque p_k est de forme :

$$p_k = 1/f(\mathbf{h}_k\boldsymbol{\phi}), \qquad (13)$$

où $\boldsymbol{\phi}$ est un vecteur colonne inconnu, $\mathbf{h}_k$ est un vecteur ligne de même dimension que $\mathbf{x}_k$, et $\sum_S a_k \mathbf{h}'_k \mathbf{x}_k / N$, où S représente maintenant le « sous-échantillon » de répondants, est inversible à la fois pour la taille de population réalisée, N, et pour la limite de probabilité.

La fonction $f(\cdot)$ de l'équation (13) est supposée être monotone et dérivable deux fois. Sa forme fonctionnelle est connue, mais la valeur du paramètre qui la régit, $\boldsymbol{\phi}$, ne l'est pas. Lorsqu'elle est introduite dans l'équation des poids de calage, $w_k = a_k f(\mathbf{h}_k\mathbf{q})$, de sorte que l'équation de calage proprement dite, $\sum_s w_k \mathbf{x}_k = T_{\mathbf{x}}$ soit vérifiée, $f(\mathbf{h}_k\mathbf{q})$ estime implicitement l'inverse de la probabilité de réponse de l'élément. Contrairement à la situation où le calage est utilisé pour la correction de l'écart de $\sum_s a_k\mathbf{x}$ par rapport à $T_{\mathbf{x}}$ dû purement à l'erreur d'échantillonnage, $f(0)$ et $f'(0)$ n'ont pas à valoir 1 et $\mathbf{h}_k\boldsymbol{\phi}$ n'a pas à valoir zéro.

Le choix le plus évident pour $\mathbf{h}_k$ lorsqu'on postule le modèle de réponse donné par l'équation (13) est $\mathbf{x}_k$ proprement dit. Dans un exemple courant de pondération par calage pour corriger pour la non-réponse, les composantes de $\mathbf{x}_k$ sont des variables indicatrices : $x_{gk} = 1$ quand k est dans le groupe g et zéro sinon. Si les groupes sont mutuellement exclusifs, la pondération par calage équivaut à une repondération dans les classes de poststratification. Voir, par exemple, Särndal, Swensson et Wretman (1992, page 585). Le modèle prédictif qui sous-tend habituellement le calage (le qualificatif « prédictif » est nécessaire pour distinguer ce modèle du modèle de quasi-randomisation) suppose que chaque élément k du groupe g, qu'il réponde ou non, a une moyenne courante : $E_\varepsilon(y_k) = \beta_g$. Le modèle de réponse quasi aléatoire est analogue : $p_k = 1/\phi_g$. Les deux modèles sont toutefois conceptuellement très différents.

Si les groupes ne sont pas mutuellement exclusifs, l'ajustement proportionnel itératif (API) est une méthode de détermination des poids de calage. Il en existe d'autres qui dépendent de la forme exacte de la fonction de réponse hypothétique $f(\cdot)$. Le modèle prédictif demeure linéaire, $E_\varepsilon(y_k) = \mathbf{x}_k\boldsymbol{\beta}$, tandis que le modèle de réponse qui donne lieu à l'API, $p_k = \exp\{-\mathbf{x}_k\boldsymbol{\phi}\}$, ne l'est pas. Berry, Flatt et Pierce (1996) donnent un exemple d'utilisation de l'API pour ajuster pour la non-réponse.

Dans de nombreuses applications de la pondération par calage, les composantes de $\mathbf{x}_k$ sont continues ou semi-continues, plutôt que dichotomiques. Dans une enquête annuelle sur les récoltes, par exemple, soit x_{1k} la quantité de maïs récoltée lors du recensement de l'agriculture précédent par l'exploitation agricole k, x_{2k}, la quantité de blé récoltée par cette exploitation, x_{3k}, la quantité de pommes de terre récoltées, et ainsi de suite. L'enquête annuelle sur les récoltes possède un modèle prédictif hypothétique pour la superficie consacrée à la culture du maïs par l'exploitation agricole k, y_{1k}, de la forme $y_{1k} = \mathbf{x}_k\beta_{1k} + \varepsilon_{1k}$. L'indice 1 désigne le maïs. Il existe d'autres valeurs d'enquête d'intérêt, comme la superficie consacrée à la culture du blé, et éventuellement des modèles prédictifs hypothétiques pour chacune.

Le modèle de réponse quasi aléatoire pour l'enquête sur les récoltes dépend des hypothèses émises au sujet de $f(\cdot)$ et de $\mathbf{h}_k$ dans l'équation (13) avec $\mathbf{h}_k$ éventuellement égal

à $\mathbf{x}_k$. Contrairement au modèle prédictif, le même modèle de quasi-randomisation hypothétique s'applique à toutes les variables de l'enquête.

Des choix prometteurs pour $f(\cdot)$ sont $\exp(\cdot)$ et $1+\exp(\cdot)$, ce dernier correspondant à un modèle de probabilité de réponse ajusté au moyen d'une fonction logistique de $\mathbf{h}_k\boldsymbol{\phi}$. Il pourrait également être raisonnable de supposer que $h_{gk} = x_{gk}^{\lambda}$ pour $\lambda < 1$. En particulier, fixer $\lambda = 0$ signifie que la probabilité que l'exploitation agricole k réponde à l'enquête annuelle sur les récoltes dépend uniquement du fait qu'elle ait déclaré du maïs, du blé ou des pommes de terre lors du recensement de l'agriculture précédent, plutôt que du volume déclaré de ces récoltes.

Dans l'exemple de l'enquête sur les récoltes, les composantes de $\mathbf{x}_k$ provenant du recensement précédent étaient les meilleurs prédicteurs disponibles des valeurs correspondantes pour l'enquête annuelle *avant* l'échantillonnage. Le fait que l'entreprise agricole k réponde à l'enquête est toutefois davantage une fonction de la superficie courante consacrée à la culture du maïs, si tant est qu'il y en ait, que d'une approximation prédéterminée de cette valeur. Par conséquent, il est tentant d'introduire les valeurs d'enquête dans $\mathbf{h}_k$, plutôt que les valeurs de recensement correspondantes. Comme nous allons voir, cette procédure pose un problème théorique.

Sachant une $f(\cdot)$, la méthode itérative décrite à la section 4 permettra souvent de découvrir un vecteur ligne $\mathbf{q}$ tel que $T_{\mathbf{x}} = \sum_s a_k f(\mathbf{h}_k\mathbf{q})\mathbf{x}_k$. Le cas échéant, l'estimation de T_y au moyen de $t_{y\_\mathrm{CAL}} = \sum_s w_k y_k$, où $w_k = a_k f(\mathbf{h}_k\mathbf{q})$, aura de bonnes propriétés sous le modèle prédictif linéaire $y_k = \mathbf{x}_k\boldsymbol{\beta} + \varepsilon_k$, où $E(\varepsilon_k|\{\mathbf{x}_g, \mathbf{h}_g, I_g|g \in U\}) = 0$ pour tout $k \in U$, $I_k = 1$ si l'élément k est présent dans l'échantillon original et qu'il répond, et 0 autrement.

L'absence de biais dans le modèle prédictif est simplement due au fait que les poids satisfont l'équation de calage. Notons toutefois que, si des composantes de $\mathbf{h}_k$ proviennent de l'enquête plutôt que de $\mathbf{x}_k$, l'hypothèse du modèle prédictif voulant que $E(\varepsilon_k|\mathbf{h}_k) = 0$ peut être problématique. Dans les conditions extrêmes, considérons le cas où une telle composante est y_k proprement dit. Habituellement, $E(\varepsilon_k|y_k)$ n'est pas nulle. Dans l'exemple de l'enquête sur les récoltes décrit plus haut, y_k peut être la superficie annuelle consacrée à la culture du maïs de l'exploitation agricole k. L'introduction de cette valeur dans $\mathbf{h}_k$ rend biaisé l'estimateur par calage connexe pour le modèle prédictif pour le maïs.

Cependant, lorsque le modèle prédictif est correct (en traitant $E(\varepsilon_k|\{\mathbf{x}_g, \mathbf{h}_g, I_g|g \in U\}) = 0$ comme une partie intégrante du modèle), la pondération par calage fondée sur tout choix de $f(\cdot)$ produira des estimateurs ayant de bonnes propriétés fondées sur le modèle prédictif. Ces estimateurs auront aussi de bonnes propriétés fondées sur le modèle de quasi-randomisation lorsque le modèle de réponse de l'équation (13) est correct pour ce choix de $f(\cdot)$. Dans un certain sens, un modèle protège contre l'échec de l'autre. Voir Kott (1994).

Comme nous l'avons souligné, le modèle prédictif a plus de chance de tenir lorsque $\mathbf{h}_g = \mathbf{x}_g$. Même ainsi, il arrive que les ε_k du modèle donné par l'équation (2) satisfassent $E(\varepsilon_k|\{\mathbf{x}_g|g \in U\}) = 0$, mais non $E(\varepsilon_k|\{\mathbf{x}_g I_g|g \in U\}) = 0$; autrement dit, le mécanisme d'échantillonnage, y compris la réponse, n'est pas ignorable en ce qui concerne le modèle prédictif.

Nous pouvons décomposer I_k en $I_{k1}I_{k2}$, où $I_{k1} = 1$ si, et uniquement si, k est présent dans l'échantillon original, et $I_{k2} = 1$ si, et uniquement si, k répond s'il est échantillonné. Le lecteur que cela intéresse peut confirmer que la pondération par calage offre une certaine protection contre le biais si le modèle prédictif de l'équation (2) est vérifié quand $E(\varepsilon_k|\{\mathbf{x}_g, \mathbf{h}_g, I_{g2}|g \in U\}) = 0$; c'est-à-dire quand le mécanisme de réponse est ignorable en ce qui concerne le modèle prédictif, mais pas nécessairement le mécanisme d'échantillonnage original.

6.2 Estimation de l'erreur quadratique moyenne sous quasi-randomisation

Que l'on puisse déclarer raisonnablement ou non que t_{y_CAL} est sans biais par rapport au modèle prédictif n'a aucun effet sur ses propriétés sous quasi-randomisation. Notons que $\mathbf{h}_k\boldsymbol{\phi}$ et $\mathbf{h}_k\mathbf{q}$ sont des valeurs scalaires et non des vecteurs. Puisque $T_{\mathbf{x}} = \sum_s a_k f(\mathbf{h}_k\mathbf{q})\mathbf{x}_i$, nos hypothèses et le théorème de la valeur moyenne ($f(\mathbf{h}_k\boldsymbol{\phi}) = f(\mathbf{h}_k\mathbf{q}) + f'(\theta_k)(\mathbf{h}_k\boldsymbol{\phi} - \mathbf{h}_k\mathbf{q})$) révèlent

$$T_{\mathbf{x}} - \sum_{k\in S} a_k f(\mathbf{h}_k\boldsymbol{\phi})\mathbf{x}_k = \sum_{k\in S} a_k[f'(\theta_k)\mathbf{h}_k(\mathbf{q} - \boldsymbol{\phi})]\mathbf{x}_k$$
$$= \mathbf{O}_P(N/\sqrt{n})$$

pour une grandeur scalaire θ_k comprise entre chaque $\mathbf{h}_k\mathbf{q}$ et $\mathbf{h}_k\boldsymbol{\phi}$. Il découle de cela que, si $\sum_s a_j f'(\mathbf{h}_j\boldsymbol{\phi})_j \mathbf{h}'_j\mathbf{x}_j/N$ est inversible à la fois pour la population réalisée N et la limite de probabilité (rappelons que f est monotone, donc que f' n'est jamais nulle), alors

$$\mathbf{q} - \boldsymbol{\phi} = \left\{\left[\sum_{j\in S} a_j f'(\mathbf{h}_j\mathbf{q})_j \mathbf{h}'_j\mathbf{x}_j\right]^{-1}\right\}' \left[T_{\mathbf{x}} - \sum_{i\in S} a_i f(\mathbf{h}_i\boldsymbol{\phi})\mathbf{x}_i\right]'$$
$$= \mathbf{O}_P(1/\sqrt{n})$$
$$= \left\{\left[\sum a_j f'(\mathbf{h}_j\boldsymbol{\phi})_j \mathbf{h}'_j\mathbf{x}_j\right]^{-1}\right\}' \left[T_{\mathbf{x}} - \sum a_i f(\mathbf{h}_i\boldsymbol{\phi})\mathbf{x}_i\right]'$$
$$+ \mathbf{O}_P(1/n).$$

L'estimateur t_{y_CAL} a une erreur de

$$t_{y\_\mathrm{CAL}} - T_y = \sum_{k\in S} a_k f(\mathbf{h}_k\mathbf{q}) y_k - \sum_{k\in U} y_k$$
$$= \sum a_k f(\mathbf{h}_k\mathbf{q}) e_k - \sum e_k,$$

où

$$e_k = y_k - \mathbf{x}_k \left(\sum\nolimits_U f'(\mathbf{h}_j\boldsymbol{\phi}) p_j \mathbf{h}_j' \mathbf{x}_j\right)^{-1} \sum\nolimits_U f'(\mathbf{h}_j\boldsymbol{\phi}) p_j \mathbf{h}_j' y_j,$$

et $p_j = 1/f(\mathbf{h}_j\boldsymbol{\phi})$. Les termes e_k sont de nouveau inconnus. Ils ont été conçus de sorte que $\sum_S a_k f'(\mathbf{h}_k\boldsymbol{\phi})\mathbf{h}_k' e_k = \mathbf{O}_P(N/\sqrt{n})$. En poursuivant :

$$\begin{aligned} & t_{y\_\mathrm{CAL}} - T_y \\ &= \sum_{k\in S} a_k f(\mathbf{h}_k\boldsymbol{\phi}) e_k - \sum_{k\in U} e_k + \sum_{k\in S} a_k \{f(\mathbf{h}_k\mathbf{q}) - f(\mathbf{h}_k\boldsymbol{\phi})\} e_k \\ &= \sum a_k f(\mathbf{h}_k\boldsymbol{\phi}) e_k - \sum e_k + \sum a_k f'(\mathbf{h}_k\boldsymbol{\phi})\mathbf{h}_k(\mathbf{q}-\boldsymbol{\phi}) e_k \\ &\quad + O_P(N/n) \\ &= \sum a_k f(\mathbf{h}_k\boldsymbol{\phi}) e_k - \sum e_k + (\mathbf{q}-\boldsymbol{\phi})' \sum a_k f'(\mathbf{h}_k\boldsymbol{\phi})\mathbf{h}_k' e_k \\ &\quad + O_P(N/n) \\ &= \sum a_k f(\mathbf{h}_k\boldsymbol{\phi}) e_k - \sum e_k + O_P(N/n). \end{aligned} \tag{14}$$

Donc, t_{y_CAL} est convergent sous quasi-randomisation sous des contraintes faibles quand $t = \sum_S a_k f(\mathbf{h}_k\boldsymbol{\phi}) y_k$ l'est.

Pour estimer l'erreur quadratique moyenne sous quasi-randomisation de t_{y_CAL} (c'est-à-dire, l'erreur quadratique moyenne de l'estimateur dans les conditions de randomisation sous le modèle de réponse), nous commençons par noter que la probabilité que les éléments k et $j, k \neq j$, soient tous deux compris dans le sous-échantillon de répondants est $\pi_{kj}^* = \pi_{kj} p_k p_j$. Soit $\pi_k^* = \pi_k p_k$, et rappelons que $a_k = 1/\pi_k$ et $1/p_k = f(\mathbf{h}_k\boldsymbol{\phi})$. Partant de l'équation (14), nous voyons que l'erreur quadratique moyenne sous quasi-randomisation de t_{y_CAL} est approximativement

$$\begin{aligned} & E_1[(t_{y\_\mathrm{CAL}} - T_y)^2] \\ &\approx \sum_{k\in U}\sum_{j\in U} (\pi_{kj}^* - \pi_k^*\pi_j^*)(e_k/\pi_k^*)(e_j/\pi_j^*) \\ &= \sum_{k\in U} (1-\pi_k^*) e_k^2/\pi_k^* \\ &\quad + \sum_{\substack{k\in U \\ k\neq j}}\sum_{j\in U} (\pi_{kj} - \pi_k\pi_j)(e_k/\pi_k)(e_j/\pi_j). \end{aligned} \tag{15}$$

Si l'échantillon original est de type Poisson, alors $v_m = \sum_s (w_k^2 - w_k) r_k^2$ avec

$$r_k = y_k - \mathbf{x}_k \left[\sum_{j\in S} a_j f'(\mathbf{h}_j\mathbf{q})\mathbf{h}_j'\mathbf{x}_j\right]^{-1} \sum_{j\in S} a_j f'(\mathbf{h}_j\mathbf{q})\mathbf{h}_j' y_j, \tag{16}$$

sert à la fois d'estimateur raisonnable de la variance fondée sur le modèle prédictif et de l'erreur quadratique moyenne fondée sur le modèle de quasi-randomisation sous des contraintes faibles, puisque $w_k \approx 1/\pi_k^*$ et $r_k \approx e_k$. Un proche parent du résidu d'échantillon non intuitif dans l'équation (16) est donné dans Folsom et Singh (2000). Voir Kott (2004a) pour une discussion plus approfondie de v_m dans un contexte d'échantillonnage pur.

Pour un plan de sondage général, nous pouvons nous approcher d'un bon estimateur de la variance/erreur quadratique moyenne avec

$$\begin{aligned} v_{\mathrm{com}} &= \sum_{k\in S} (w_k^2 - w_k) r_k^2 \\ &\quad + \sum_{k\in S}\sum_{\substack{j\in S \\ k\neq j}} [(\pi_{kj} - \pi_k\pi_j)/\pi_{kj}](w_k r_k)(w_j r_j). \end{aligned} \tag{17}$$

Le deuxième membre de l'équation (17) estime le deuxième membre de l'équation (15) avec r_k remplaçant e_k. Notons que $\sum_U (1-\pi_k^*) e_k^2/\pi_k^*$ dans l'équation (15) est estimé par $\sum_S (w_k^2 - w_k) r_k^2$ plutôt que par $\sum_S w_k^2 (1-\pi_k^*) r_k^2$, ce qui rendrait v_{com} plus convergent avec v_{SSW} de l'équation (8). Cette substitution donne un estimateur de variance ayant de bonnes propriétés basées sur le modèle prédictif quand les ε_k sont non corrélées, et $\sigma_k^2 = \mathbf{x}_k\boldsymbol{\zeta}$, pour un certain $\boldsymbol{\zeta}$. Elle peut être faite même en l'absence de non-réponse.

Lorsque l'échantillon réel comprend plusieurs degrés et que les probabilités de sélection de premier degré sont suffisamment faibles pour être ignorées, v_{ST2} de l'équation (10) peut être utilisée comme estimateur de la variance/erreur quadratique moyenne avec r_k de nouveau défini par l'équation (16).

Quand f est linéaire, $f'(\theta) = 1$, et les r_k de l'équation (16) sont calculés comme s'il n'y avait aucune non-réponse. Cela est également vrai pour l'estimateur de la variance/erreur quadratique moyenne v_{ST2}. Malheureusement, cette f correspond à une fonction de probabilité de réponse de forme peu maniable : $p_k = 1/\mathbf{h}_k\boldsymbol{\phi}$. Fuller, Loughin et Baker (1994) ont fait ces constatations pour le cas où $\mathbf{h}_k = c_k\mathbf{x}_k$.

Dans l'équation (11), le jackknife, v_J, peut être calculé à l'aide des poids de rééchantillonnage jackknife :

$$\begin{aligned} w_{k(\alpha j)} &= w_k a_{k(\alpha j)}/a_k + \left(\sum_{m\in U} \mathbf{x}_m - \sum_{m\in S} w_m [a_{m(\alpha j)}/a_m]\mathbf{x}_m\right) \\ &\quad \times \left(\sum_{m\in S} a_{m(\alpha j)} f'(\mathbf{h}_m\mathbf{q})\mathbf{h}_m'\mathbf{x}_m\right)^{-1} a_{k(\alpha j)} f'(\mathbf{h}_k\mathbf{q})\mathbf{h}_k', \end{aligned} \tag{18}$$

ce qui est une généralisation évidente des poids de rééchantillonnage jackknife de l'équation (12). De nouveau, si $f'(\theta) = 1$, v_J peut être calculé comme s'il n'y avait pas de non-réponse.

7. Modélisation de la couverture

Folsom et Singh (2000) ont fait remarquer que le traitement de la non-réponse au moyen de la pondération par calage peut aussi être appliqué pour corriger pour le sous-dénombrement. Dans le contexte, la phase quasi

aléatoire, de l'échantillonnage a lieu conceptuellement avant que l'échantillon réel soit tiré. Nous supposons que la population associée à la base de sondage est un échantillon de Poisson provenant d'une population complète hypothétique pour laquelle le vecteur $T_{\mathbf{x}}$ doit être connu. La population de la base de sondage devient F, tandis que la population complète hypothétique est U. Nous supposons que la probabilité que l'élément $k \in U$ soit dans F est modélisé correctement par l'équation (13). Si la première (de U à F) et la deuxième (de F à S) phases d'échantillonnage sont indépendantes, alors toute la théorie élaborée pour l'utilisation de la pondération par calage en vue de traiter la non-réponse peut être transposée au traitement du sous-dénombrement.

Il convient de souligner que la correction de l'erreur de couverture par calage est une extension de la pratique bien connue de correction par poststratification souvent utilisée dans les enquêtes téléphoniques. Comme dans le cas particulier de la poststratification, il faut utiliser comme cible de calage pour U des quantités que l'on peut supposer dépourvues d'erreur ou ayant une erreur quadratique moyenne très faible comparativement aux estimateurs par calage proprement dit.

Folsom et Singh ont fait remarquer que le surdénombrement (dénombrement multiple) ou une combinaison de sous- et de surdénombrement peuvent être traités en suivant leur méthode. La définition de p_k dans l'équation (13) devient le nombre prévu de fois que k est présent dans la base de sondage, nombre qui peut maintenant être supérieur à 1 à cause du dénombrement multiple éventuel.

Folsom et Singh proposent en outre de donner à $f(\cdot)$ la forme flexible :

$$f(\mathbf{x}_k\boldsymbol{\phi}) = \frac{U(C-L)\exp(\mathbf{x}_k\boldsymbol{\phi}) + L(U-C)}{(U-C) + (C-L)\exp(\mathbf{x}_k\boldsymbol{\phi})}, \qquad (19)$$

où $L \geq 0, 1 < U \leq \infty,$ et $L < C \leq U$ sont des constantes prédéterminées. Ils donnent à cette expression le nom de « modèle exponentiel général » ou « MEG ». Observons que, si $C = 1, U = \infty,$ et $L = 0, p_k = 1/f(\mathbf{x}_k\boldsymbol{\phi}) = \exp(-\mathbf{x}_k\boldsymbol{\phi})$. Similairement, si $C = 2,$ $U = \infty,$ et $L = 1,$ $p_k = [1 + \exp(\mathbf{x}_k\boldsymbol{\phi})]^{-1}$; autrement dit, la probabilité de couverture (ou de réponse) est logistique. Les valeurs L et U servent de bornes sur l'*ajustement par calage*, $f(\cdot),$ tandis que $C = f(0)$ est effectivement son centre.

Les auteurs ont rendu l'ajustement par calage dans le MEG encore plus souple en postulant trois classes d'unités d'échantillonnage, chacune ayant son propre ensemble de valeurs U, C et L. Ils ont proposé son utilisation pour la correction de l'erreur de couverture ainsi que de la non-réponse totale.

8. Un petit exemple empirique

Puisque les poids de rééchantillonnage jackknife exprimés par l'équation (18) sont nouveaux, il est prudent de chercher à savoir s'ils fonctionnent effectivement avec des données réelles. Pour ce faire, nous avons pris les données MU281 de Särndal, Swensson et Wretman (1992) et les avons répétées 20 fois (de sorte que $N = 5\,620$). Par échantillonnage aléatoire simple stratifié, nous avons sélectionné 16 unités dans chacune des huit strates de taille inégale. La variable RMT85 a servi de y_k et la variable P75, de x_k dans $\mathbf{x}_k = (1, x_k)$. À chacune des 128 unités échantillonnées, nous avons attribué une probabilité d'être présente dans le sous-échantillon de répondants, S, qui diminuait avec la taille de x_k; en particulier, $p_k = \exp(-0{,}35\, x_k / M_x),$ où M_x était la moyenne de population de x_k. Dans les 1 600 simulations, la taille de S variait de 78 à 110, avec une moyenne d'environ 93,8.

Le total T_y a été estimé de deux façons, avec $t_{y\_\mathrm{LIN}} = \sum_S a_k(1 + \mathbf{x}_k\mathbf{q})y_k$ et avec $t_{y\_\mathrm{EXP}} = \sum_S a_k \exp(\mathbf{x}_k\mathbf{q}^{(\mathrm{exp})})y_k,$ où $\mathbf{q}$ et $\mathbf{q}^{(\mathrm{exp})}$ étaient, respectivement, sélectionnés de sorte que l'équation de calage soit vérifiée. Le premier était un estimateur GREG, tandis que le second était un estimateur par ajustement proportionnel itératif généralisé. Les deux estimateurs étaient sans biais sous le modèle prédictif sous-entendu $(y_k = \mathbf{x}_k\boldsymbol{\beta} + \varepsilon_k),$ mais seul t_{y_EXP} était convergent dans des conditions de randomisation sous le modèle de réponse correcte. L'estimateur GREG supposait implicitement que $p_k = 1/(\phi_0^{(\mathrm{LIN})} + \phi_1^{(\mathrm{LIN})}\mathbf{x}_k)$ pour $\phi_0^{(\mathrm{LIN})}$ et $\phi_1^{(\mathrm{LIN})}$ inconnus.

La petite taille de l'échantillon comparativement à la population de chaque strate a permis d'ignorer la correction pour population finie dans l'estimation de la variance/erreur quadratique moyenne (appelée dans la suite « estimation de la variance »). Nous avons estimé les variances en utilisant *i*) l'estimateur par linéarisation, $v_{\mathrm{ST2}},$ dans l'équation (10) avec r_k défini par l'équation (16) et *ii*) le jackknife proposé, $v_J,$ dans l'équation (11) avec les poids de rééchantillonnage définis par l'équation (18). Pour rendre le calcul du jackknife plus simple, les 16 sous-échantillons dans chaque strate ont été attribués aléatoirement à l'une de quatre grappes, de sorte que 32 répliques jackknife seulement ont dû être calculées.

Aux fins de comparaison, une meilleure version de l'estimateur de variance par linéarisation, dénotée $v_{\mathrm{ST2}(e)},$ a également été calculée avec r_k remplacé par $e_k = y_k - \mathbf{x}_k(\sum_U f'(\mathbf{x}_j\boldsymbol{\phi})\ p_j\mathbf{x}'_j\mathbf{x}_j)^{-1}\sum_U f'(\mathbf{x}_j\boldsymbol{\phi})p_j\mathbf{x}'_j y_j,$ où $\boldsymbol{\phi}$ et p_j étaient connus. En pratique, e_k est rarement connu, mais le calcul de $v_{\mathrm{ST2}(e)}$ est utile ici pour les comparaisons.

Il convient de souligner que les calculs de r_k et e_k différaient légèrement selon que l'on voulait calculer l'estimateur de la variance pour t_{y_LIN} ou pour t_{y_EXP}.

Pour t_{y_LIN}, $f'(\mathbf{x}_j\boldsymbol{\phi}) = f'(\mathbf{x}_j\mathbf{q}) = 1$; pour t_{y_EXP}, $f'(\mathbf{x}_j\mathbf{q}^{(\exp)}) = \exp(\mathbf{x}_j\mathbf{q}^{(\exp)})$ et $f'(\mathbf{x}_j\boldsymbol{\phi}) = 1/p_j$.

Le tableau 1 donne les moyennes empiriques (la moyenne sur les 1 600 simulations) des deux estimateurs pour T_y normalisés de sorte que $T_y = 100$. Bien que tous deux soient pratiquement sans biais, t_{y_LIN} diffère significativement de 100 au seuil de signification de 0,05; ce n'est pas le cas pour t_{y_EXP}. Cela n'est pas étonnant, parce que seul le dernier est fondé sur le modèle de réponse correcte.

Pour chaque estimateur, les estimateurs de variance et les erreurs quadratiques moyennes empiriques ont été normalisés de sorte que les moyennes empiriques des $v_{\mathrm{ST2}(e)}$ respectifs soient égales à 100. Aucun $v_{\mathrm{ST2}(e)}$ n'avait une moyenne empirique significativement différente de l'erreur quadratique moyenne empirique (EQME) de l'estimateur associé. Ce résultat était un peu décevant. Il semble que, bien que t_{y_LIN} ait un biais empirique significatif, celui-ci était une composante tellement faible de l'erreur quadratique moyenne de l'estimateur que la différence entre l'EQME de cet estimateur et la moyenne empirique de $v_{\mathrm{ST2}(e)}$ n'était pas significative.

Les $v_{\mathrm{ST2}(e)}$ ont été choisis comme valeurs de référence pour le tableau plutôt que les erreurs quadratiques moyennes empiriques, parce que chaque $v_{\mathrm{ST2}(e)}$ avait environ la moitié de l'erreur-type empirique de l'EQME correspondante (qui, elle-même, correspondait à la moyenne des 1 600 carrés des écarts) et était corrélée plus fortement avec les estimateurs de variance. Les valeurs t pour cette partie du tableau ont également été calculées par rapport aux $v_{\mathrm{ST2}(e)}$.

Les deux estimateurs de variance par linéarisation avaient un biais par défaut étonnamment grand. Il semble que les w_{k_LIN} et w_{k_EXP} inhabituellement grands avaient tendance à rendre les r_k associés appréciablement plus faibles que les e_k en valeur absolue. Les problèmes associés à des valeurs inhabituellement grandes de w_{k_LIN} et w_{k_EXP} paraissent être plus atténués dans le cas des estimateurs jackknife.

Pour accélérer l'asymptotique des estimateurs de variance par linéarisation (c'est-à-dire, réduire l'écart entre r_k et e_k), nous avons calculé un ajustement ponctuel de v_{ST2} en remplaçant chaque r_k par $r_{k(\mathrm{ajusté})} = r_k/\omega_k$, où $\omega_k^2 = 1 - \mathbf{x}_k(\sum_S a_j f'(\mathbf{x}_j\mathbf{q})\mathbf{x}_j'\mathbf{x}_j)^{-1} a_k f'(\mathbf{x}_k\mathbf{q})\mathbf{x}_k' = 1 + O_P(1/n)$. Notons que, sous le modèle prédictif avec les ε_k non corrélés et $E(\varepsilon_k^2) = \sigma_k^2$, $E(r_{k(\mathrm{ajusté})}^2) \approx \sigma_k^2$. La quasi-égalité est exacte lorsque tous les $a_j f'(\mathbf{x}_j\mathbf{q})$ et σ_j, sont, respectivement, égaux.

Le v_{ST2} ajusté pour t_{y_LIN} ainsi que t_{y_EXP} demeurait entaché d'un biais par défaut, tandis que le v_J présentait un biais par excès d'une valeur légèrement plus faible. Bien que ces biais soient significatifs, ils étaient raisonnablement petits (de 4,5 à 11,2 %) et donnent à penser que les estimateurs de variance étaient peut-être effectivement asymptotiquement sans biais, comme nous l'avons démontré théoriquement aux sections précédentes.

Tableau 1
Moyennes empiriques des estimateurs basées sur 1 600 simulations*

Moyenne empirique (erreur-type)		Valeur t (test de signification bilatéral)	
Estimateurs pour $T_y (T_y = 100)$			
t_{y_LIN}	99,84 (0,06)	−2,79 (0,02)	différent de
t_{y_EXP}	100,04 (0,06)	0,58 (0,56)	T_y
Estimateurs de variance pour $t_{y\_\mathrm{LIN}} (E_{\mathrm{EMP}}(v_{\mathrm{ST2}(e)}) = 100)$			
v_{ST2}	83,59 (1,53)	−19,96 (< 0,0001)	différent de
$v_{\mathrm{ST2(ajusté)}}$	95,53 (1,80)	−6,09 (< 0,0001)	$v_{\mathrm{ST2}(e)}$
v_J	104,69 (2,28)	3,60 (0,0003)	
EQME	99,35 –	−0,18 (0,85)	
Estimateurs de variance pour $t_{y\_\mathrm{EXP}} (E_{\mathrm{EMP}}(v_{\mathrm{ST2}(e)}) = 100)$			
v_{ST2}	73,12 (1,54)	−18,22 (< 0,0001)	différent de
$v_{\mathrm{ST2(ajusté)}}$	88,79 (1,98)	−8,57 (< 0,0001)	$v_{\mathrm{ST2}(e)}$
v_J	107,00 (2,73)	4,09 (< 0,0001)	
EQME	101,21 –	0,33 (0,74)	
Autres statistiques			
relvar $(v_{\mathrm{ST2}(e)[\mathrm{LIN}]})$	0,051 –	–	
relvar $(v_{\mathrm{ST2}(e)[\mathrm{EXP}]})$	0,059 –	–	
$\dfrac{(v_{\mathrm{ST2}(e)[\mathrm{LIN}]} - v_{\mathrm{ST2}(e)[\mathrm{EXP}]})}{(E_{\mathrm{EMP}}(v_{\mathrm{ST2}(e)[\mathrm{EXP}]})}$	−0,1340 (0,010)	−13,87 (< 0,0001)	

* Dans quatre simulations supplémentaires, la convergence n'a pas été atteinte en dix itérations pour t_{y_EXP}. Ces simulations ont été exclues de l'analyse.

Lors de l'utilisation de $v_{\mathrm{ST2}(e)}$ comme approximation efficace de l'EQME, l'erreur quadratique moyenne empirique de t_{y_EXP}, qui intégrait le modèle de réponse correcte, était plus de 13 % plus importante que celle de t_{y_LIN}, qui n'intégrait pas ce modèle. Toutefois, il ne convient pas de faire de grandes généralisations à partir d'un ensemble de données comportant deux variables de calage seulement. Voir Crouse et Kott (2004) pour un ensemble différent de résultats.

Qu'il soit préférable ou non d'intégrer le modèle de réponse correcte dans l'estimateur de calage, si on le fait, alors les estimateurs de variance discutés à la section précédente, peut-être avec l'estimateur par linéarisation corrigé comme il est suggéré à la présente section, semblent utilisables.

Un deuxième ensemble de 1 600 simulations (non présentées) ont été exécutées en utilisant la même population et un plan d'échantillonnage stratifié, mais en donnant à chaque élément échantillonné 70 % de chances de faire partie de l'échantillon de répondants (la taille moyenne de l'échantillon de répondants était d'environ 89,8). Dans cet ensemble de simulations, les deux estimateurs de T_y sont convergents par rapport au plan d'échantillonnage (randomisation) sous le modèle de réponse. Par conséquent, il n'est pas étonnant que les moyennes empiriques de t_{y_LIN} et de t_{y_EXP} soient presque identiques (écart d'au plus 0,01 %), comme le sont leurs erreurs quadratiques moyennes empiriques (écart d'au plus 1 %). Les moyennes empiriques de chaque paire d'estimateurs de variance (par exemple $\mathrm{var}_{\mathrm{ST2}}$ pour t_{y_LIN} et t_{y_EXP}) étaient aussi très proches (écart d'au plus 1 %). Le biais relatif de l'estimateur v_{ST2} ajusté (comparativement à $\mathrm{var}_{\mathrm{ST2}(e)}$) était de –1,3 % lors de l'estimation de la variance de t_{y_LIN} et de –2,2 % lors de l'estimation de la variance de t_{y_EXP}. Le biais relatif des variances par linéarisation non corrigées était de –9,0 % et de –10,3 %, respectivement. Le biais relatif des deux estimateurs jackknife était de 3,6 %.

9. Discussion

9.1 Estimation explicite d'un modèle de réponse

En présence de non-réponse totale, nombreux sont ceux qui ont essayé d'estimer les probabilités de réponse individuelles, $p_k = 1/f(\mathbf{h}_k\boldsymbol{\phi})$, directement. Cette méthode requiert de l'information sur $\mathbf{h}_k$ pour chaque unité échantillonnée, qu'elle réponde à l'enquête ou non, mais $\mathbf{h}_k$ ne doit pas avoir la même dimension que $\mathbf{x}_k$. La méthode d'ajustement directe n'est généralement pas disponible pour le traitement des erreurs de couverture.

Fuller (2002) a souligné qu'un terme supplémentaire peut figurer dans l'erreur quadratique moyenne sous quasi-randomisation de $t_{y\_\mathrm{GREG}} = \sum_S a_k^* y_k + (T_{\mathbf{x}} - \sum_S a_j^* \mathbf{x}_j)\times (\sum_S c_j a_j^* \mathbf{x}_j'\mathbf{x}_j)^{-1}\sum_S c_k a_k^* \mathbf{x}_k'\mathbf{x}_k$, où S est le sous-échantillon de répondants, $a_k^* = a_k[1+f(\mathbf{h}_k\mathbf{q})]$, et $\mathbf{q}$ est un estimateur direct convergent pour le paramètre du modèle de quasi-randomisation, $\boldsymbol{\phi}$. Cela ne sous-entend pas que l'estimation directe du modèle de réponse fondée sur une $f(\cdot)$ et un $\mathbf{h}_k$ donnés est moins efficace que le calage analogue lorsque $\mathbf{h}_k$ a la même dimension que $\mathbf{x}_k$. Voir Kim (2004) pour une suggestion du contraire. Néanmoins, la commodité qu'offre l'intégration de la correction pour la non-réponse dans le calage est séduisante, lorsque les estimations de variance doivent être produites.

Un compromis raisonnable consiste à choisir la forme de $f(\cdot)$ et de $\mathbf{h}_k$ en modélisant le comportement de réponse de l'échantillon complet, puis à estimer le paramètre de $f(\cdot)$ implicitement par calage. Ce compromis permet aussi de contourner une faiblesse frappante de l'utilisation de la pondération par calage pour corriger de la non-réponse (ainsi que pour les erreurs de couverture). Les choix de $f(\cdot)$ et $\mathbf{h}_k$ sont motivés principalement par la vraisemblance et la commodité et non par une analyse statistique des données.

9.2 Groupes à homogénéité de réponse

Afin de contrôler l'importance de la repondération due à la non-réponse, Little (1986) a recommandé que l'on estime $\mathbf{q}$ explicitement, puis qu'on divise l'échantillon en C groupes mutuellement exclusifs en se fondant sur les tailles des valeurs ajustées de $f(\mathbf{h}_k\mathbf{q})$. On calcule ensuite le poids corrigé pour chaque élément k dans le groupe c comme dans le cas de la poststratification :

$$w_{k\_\mathrm{ADJ}} = \left(\sum_{F(c)} w_g / \sum_{S(c)} w_g\right) w_k,$$

où $F_{(c)}$ est la partie de l'échantillon original comprise dans le groupe c, $S(c)$ est le sous-échantillon de $F(c)$ qui répond, et w_k est le poids d'échantillonnage attribué à l'élément k après échantillonnage, mais avant sous-échantillonnage quasi aléatoire. Cette approche suppose que chaque élément d'un groupe a (approximativement) la même probabilité de réponse, d'où l'expression « groupe à homogénéité de réponse ».

Un autre moyen d'intégrer les valeurs ajustées de $f(\mathbf{h}_k\mathbf{q})$ dans l'estimation fondé sur la méthodologie développée dans le texte est décrit ci-après. Répartir les valeurs ajustées en P groupes d'après leur taille, où P est de nouveau la dimension de $\mathbf{x}_k$, et soit $\mathbf{d}_k$, un vecteur ligne de variables indicatrices pour les P cellules. En fixant chaque $w_k = a_k[1+(T_{\mathbf{x}} - \sum_S a_j\mathbf{x}_j)\ (\sum_S a_j\mathbf{d}_j'\mathbf{x}_j)^{-1}\mathbf{d}_k']$, on calcule un ensemble de poids pour le sous-échantillon de répondants qui, contrairement à $\{w_{k\_\mathrm{ADJ}}\}$ ci-dessus, satisfait l'équation de calage pour l'échantillon de répondants. Étant donné la nature de $\mathbf{d}_k$, cette méthode linéaire produit le même ensemble de poids de calage que celui que donnerait l'ajustement de $w_k = a_k \exp(\mathbf{d}_k\mathbf{f})$ – si les deux produisent

un ensemble de poids. Notons que, puisque les poids de calage peuvent être négatifs dans le cas de la méthode linéaire, celle-ci pourrait produire un ensemble que la méthode par ajustement proportionnel itératif généralisé ne pourrait pas produire. La méthode linéaire rééchelonne effectivement les a_k, c'est-à-dire la valeur de chaque élément dans le même groupe, d'une quantité fixe. Donc, elle pourrait ne pas produire des poids étonnamment petits ou étonnamment grands lorsque la dimension de $\mathbf{x}_k$ est faible comparativement à la taille d'échantillon.

9.3 Calage de l'échantillon et calage pour la non-réponse

À la section précédente, nous avons indiqué qu'il est possible que les composantes de $\mathbf{h}_k$ dans l'équation (13), c'est-à-dire le modèle de réponse quasi aléatoire, soient inconnues avant le recensement. Lorsqu'on utilise un tel $\mathbf{h}_k$ dans le calage, il n'est peut-être plus raisonnable d'affirmer que l'estimateur t_{y_CAL} résultant est sans biais par rapport au modèle prédictif. Cela est particulièrement ennuyeux lorsque la non-réponse est modérée, comparativement à la taille de l'échantillon. Une idée intéressante consiste à faire le calage en deux phases. La première phase, le calage de l'échantillon, consiste à corriger pour la différence entre $T_{\mathbf{x}}$ et $\sum_F a_k \mathbf{x}_k$, et ne comprendrait aucune composante de $\mathbf{h}_k$ inconnue au moment de l'échantillonnage. La deuxième phase, le calage pour la non-réponse, corrige pour la différence entre $\sum_F a_k \mathbf{x}_k$ et $\sum_S a_k \mathbf{x}_k$ et comprendrait uniquement les variables composantes disponibles après que le sous-échantillon de répondants soit recensé.

Une analyse plus approfondie de cette idée doit être reportée à une autre occasion.

9.4 Travaux avec le NASS

Le National Agricultural Statistics Service (NASS) a utilisé des variantes de l'approche de Fuller et coll. (1994) pour traiter le sous-dénombrement au Recensement de l'agriculture de 2002 (voir Fetter et Kott 2003) et pour la correction d'une enquête sur l'économie agricole avec non-réponse importante, de façon à faire concorder les totaux à ceux d'enquêtes plus fiables (voir Crouse et Kott 2004). Dans cette approche, $f(\cdot)$ est de la forme :

$$f(\mathbf{x}_k\boldsymbol{\phi}) = \begin{cases} L & \text{si} \quad \mathbf{x}_k\boldsymbol{\phi} < L \\ \mathbf{x}_k\boldsymbol{\phi} & \text{si} \quad \mathrm{L} \le \mathbf{x}_k\boldsymbol{\phi} \le U \\ U & \text{si} \quad \mathbf{x}_k\boldsymbol{\phi} > U, \end{cases} \tag{20}$$

qui tronque le calage linéaire à des valeurs préétablies, L et U, pour contrôler l'importance de l'ajustement des poids. Notons que, quand $f(\cdot) = U$ ou L, $f'(\cdot) = 0$. Contrairement à l'ajustement par calage de l'équation (19), $f(\cdot)$ de l'équation (20) n'est pas dérivable deux fois aux valeurs L ou U. Cela ne cause pas de problème en pratique.

La justification originale du calage offerte par l'organisme dans ce contexte était fondée sur la modélisation prédictive. L'équation (20) est simple à appliquer et semble produire des poids qui se situent dans une fourchette acceptable plus fréquemment que d'autres solutions facilement disponibles.

Le NASS étudie les questions suivantes : quelle est la sensibilité de t_{y_CAL} au choix de $f(\cdot)$ en pratique? Un choix différent pour $f(\cdot)$ produirait-il un biais plus faible et, le cas échéant, la réduction du biais absolu se traduirait-elle par une erreur quadratique moyenne plus faible? Quel serait l'effet du remplacement de certaines composantes du vecteur de variables et de calage par un meilleur prédicteur de la non-réponse ou du sous-dénombrement?

Bibliographie

Berry, C.C., Flatt, S.W. et Pierce, J.P. (1996). Correcting unit nonresponse via nonresponse modeling and raking in the California Tobacco Survey. *Journal of Official Statistics*, 12, 349-363.

Crouse, C., et Kott, P.S. (2004). Evaluation alternative calibration schemes for an economic survey with large nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

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

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

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

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

Fetter, M.J., et Kott., P.S. (2003). Developing a coverage adjustment strategy for the 2002 Census of Agriculture. Présenté à 2003 Federal Committee on Statistical Methodology Research Conference, http://www.fcsm.gov/03papers/fetter_kott.pdf.

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

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

Fuller, W.A., Loughin, M.M. et Baker, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey 1987–88. *Techniques d'enquête*, 20, 79-89.

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

Kott, P.S. (1990). The design consistent regression estimator and its conditional variance. *Journal of Statistical Planning and Inference*, 24, 287-296.

Kott, P.S. (1994). A note on handling nonresponse in surveys. *Journal of the American Statistical Association*, 89, 693-696.

Kott, P.S. (2004a). Randomization-assisted model-based survey sampling. *Journal of Statistical Planning and Inference*, 48, 263-277.

Kott, P.S. (2004b). Commentaire. *Techniques d'enquête*, 30, 28-29.

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

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

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

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

Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 161-168
Statistique Canada, N° 12-001 au catalogue

L'importance de la modélisation du plan d'échantillonnage dans l'imputation multiple pour les données manquantes

Jerome P. Reiter, Trivellore E. Raghunathan et Satkartar K. Kinney [1]

Résumé

La théorie de l'imputation multiple pour traiter les données manquantes exige que l'imputation soit faite conditionnellement du plan d'échantillonnage. Cependant, comme la plupart des progiciels standard utilisés pour l'imputation multiple fondée sur un modèle reposent sur l'hypothèse d'un échantillonnage aléatoire simple, de nombreux praticiens sont portés à ne pas tenir compte des caractéristiques des plans d'échantillonnage complexes, comme la stratification et la mise en grappes, dans leurs imputations. Or, la théorie prédit que l'analyse d'ensembles de données soumis de telle façon à une imputation multiple peut produire des estimations biaisées du point de vue du plan de sondage. Dans le présent article, nous montrons au moyen de simulations que i) le biais peut être important si les caractéristiques du plan sont reliées aux variables d'intérêt et que ii) le biais peu être réduit en tenant compte de l'effet des caractéristiques du plan dans les modèles d'imputation. Les simulations montrent aussi que l'introduction de caractéristiques non pertinentes du plan comme contraintes dans les modèles d'imputation peut donner lieu à des inférences conservatrices, à condition que les modèles contiennent aussi des variables explicatives pertinentes. Ces résultats portent à formuler la prescription qui suit à l'intention des imputeurs : le moyen le plus sûr de procéder consiste à inclure les variables du plan de sondage dans la spécification des modèles d'imputation. À l'aide de données réelles, nous donnons une démonstration d'une approche simple d'intégration des caractéristiques d'un plan de sondage complexe qui peut être suivie en utilisant certains progiciels standard pour créer des imputations multiples.

Mots clés : Plan de sondage complexe; imputation multiple; non-réponse; enquêtes.

1. Introduction

En général, dans les grandes enquêtes, les unités échantillonnées ne répondent pas toutes complètement au questionnaire. Certaines n'y répondent pas du tout et d'autres ne répondent qu'à certaines questions. Une approche pour traiter ce genre de non-réponse est l'imputation multiple des données manquantes (Rubin 1987). Elle a été utilisée, par exemple, dans le Fatality Analysis Reporting System (Heitjan et Little 1991), la Consumer Expenditures Survey (Raghunathan et Paulin 1998), la National Health and Nutrition Examination Survey (Schafer, Ezzati-Rice, Johnson, Khare, Little et Rubin 1998), la Survey of Consumer Finances (Kennickell 1998) et la National Health Interview Survey (Schenker, Raghunathan, Chiu, Makuc, Zhang et Cohen 2005). L'imputation multiple a également été proposée pour assurer la protection des renseignements personnels dans les fichiers de données à grande diffusion (Rubin 1993; Little 1993; Raghunathan, Reiter et Rubin 2003; Reiter 2003, 2004, 2005). Pour une revue d'autres applications, voir Rubin (1996), ainsi que Barnard et Meng (1999).

En théorie, lors de l'établissement de méthodes d'inférence d'après des ensembles de données ayant subi une imputation multiple, cette dernière est rendue conditionnelle au plan d'échantillonnage (Rubin 1987). Toutefois, les imputeurs tiennent rarement compte des caractéristiques des plans d'échantillonnage complexes, comme la stratification et la mise en grappes, lorsqu'ils utilisent les progiciels disponibles pour construire des modèles d'imputation. Ils se servent plutôt de modèles normaux ou de modèles de localisation généraux multivariés (par exemple, le logiciel NORM rédigé par Joe Schafer), ou de modèles de régression séquentielle (Raghunathan, Lepkowshi, van Hoewyk et Solenberger 2001). Bien que ces méthodes puissent être modifiées afin d'intégrer les caractéristiques du plan, cela se fait rarement.

L'objectif du présent article est double. En premier lieu, nous illustrons le biais qui peut se produire lorsque les imputeurs omettent de tenir compte des caractéristiques du plan de sondage complexe dans les modèles d'imputation. Pour cela, nous simulons une imputation multiple dans des échantillons à deux degrés, stratifiés et mis en grappes. Les simulations indiquent que le biais peut être important, même si l'on applique des estimateurs fondés sur le plan de sondage à des ensembles de données soumis à l'imputation multiple ne présentant qu'une quantité modérée de données manquantes. En deuxième lieu, nous proposons deux approches simples en vue de tenir compte des caractéristiques du plan dans les modèles d'imputation. La première, qui est relativement facile à mettre en œuvre, comprend des variables nominales pour les effets de strate ou de grappes dans les modèles d'imputation. La deuxième, qui requiert des calculs plus compliqués que la première,

1. Jerome P. Reiter et Satkartar K. Kinney, Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708, États-Unis; Trivellore E. Raghunathan, Department of Biostatistics and Institute for Social Research, University of Michigan, Ann Arbor, MI 48106, États-Unis.

s'appuie sur des modèles hiérarchiques où i) les effets de la mise en grappes sont intégrés en utilisant des effets aléatoires et ii) les effets de la stratification sont intégrés en utilisant des effets fixes. Les simulations montrent que tenir compte du plan de sondage de cette façon peut réduire le biais. Elles illustrent aussi le fait qu'introduire des caractéristiques du plan qui ne sont pas reliées aux variables de l'enquête peut donner lieu à des inférences inefficaces, mais prudentes, comparativement à celles faites d'après des modèles dans lesquels ce genre de caractéristiques ne sont pas intégrées comme contraintes, à condition que les modèles incluent les variables explicatives requises pour que l'hypothèse selon laquelle les données manquent au hasard (Rubin 1976) soit plausible. Nous démontrons la première approche d'intégration des caractéristiques du plan en imputant des données manquantes dans le cas de la National Health and Nutrition Examination Survey selon une méthode de régression séquentielle.

2. Inférences d'après des ensembles de données multi-imputés

Afin de décrire la construction d'ensembles de données multi-imputés et les inférences d'après ces derniers, nous utilisons la notation de Rubin (1987). Pour une population finie de taille N, soit $I_j = 1$ si l'unité j est sélectionnée dans l'enquête originale, et $I_j = 0$ autrement, où $j = 1, 2, \ldots, N$. Soit $I = (I_1, \ldots, I_N)$. Soit n la taille d'échantillon obtenue au moyen d'un plan d'échantillonnage complexe. Pour simplifier la notation, supposons qu'une seule variable de l'enquête est sujette à la non-réponse. Soit $R_j = 1$, si l'unité j répond à l'enquête originale, et $R_j = 0$, autrement. Soit $R = (R_1, \ldots, R_N)$. La notation peut être étendue afin de traiter la non-réponse partielle multivariée, mais ce genre de complication n'est pas nécessaire aux fins de notre exposé.

Soit Y la matrice de données d'enquête de dimensions $N \times p$ pour toutes les unités de la population. Soit $Y_{\text{inc}} = (Y_{\text{obs}}, Y_{\text{mis}})$ la matrice de données d'enquête de dimensions $n \times p$ pour les unités pour lesquelles $I_j = 1$; Y_{obs} est la portion de Y_{inc} qui est observée, et Y_{mis} est la portion de Y_{inc} qui manque à cause de la non-réponse. Soit Z la matrice de variables du plan de dimensions $N \times d$ pour les N unités de la population, par exemple, des indicateurs de strates ou de grappes ou des mesures de taille. Nous supposons que ce genre d'information sur le plan est connue au moins approximativement, par exemple d'après les dossiers du recensement ou les bases de sondage.

Les valeurs de Y_{mis} sont habituellement construites d'après des tirages à partir d'une approximation de la loi bayésienne prédictive a posteriori de $(Y_{\text{mis}} | Z, Y_{\text{obs}}, I, R)$. Ces tirages sont répétés indépendamment $l = 1, \ldots, M$ fois pour obtenir M ensembles de données complets, $D^{(l)} = (Z, Y_{\text{obs}}, Y_{\text{mis}}^{(l)}, I, R)$.

D'après ces ensembles de données multi-imputés, un utilisateur des données veut faire des inférences au sujet d'un paramètre $Q = Q(Z, Y)$. Par exemple, Q pourrait être une moyenne de population ou un coefficient de régression de population. Dans chaque ensemble de données imputé, $D^{(l)}$, l'analyste estime Q au moyen d'un estimateur q, et la variance de q, au moyen d'un estimateur u. Nous supposons qu'il spécifie q et u en agissant comme si chaque $D^{(l)}$ était, en fait, formé de données recueillies auprès d'un échantillon aléatoire de (Z, Y) en se fondant sur le plan d'échantillonnage original I, c'est-à-dire que q et u sont des estimateurs sur données complètes.

Pour $l = 1, \ldots, M$, posons que $q^{(l)}$ et $u^{(l)}$ sont, respectivement, les valeurs de q et de u dans l'ensemble de données $D^{(l)}$. Sous les hypothèses décrites dans (Rubin 1987), l'analyste peut obtenir les inférences valides pour le scalaire Q en combinant les $q^{(l)}$ et $u^{(l)}$. Plus précisément, les quantités qui suivent sont nécessaires pour les inférences :

$$\bar{q}_M = \sum_{l=1}^{M} q^{(l)} / M \tag{1}$$

$$b_M = \sum_{l=1}^{M} (q^{(l)} - \bar{q}_M)^2 / (M - 1) \tag{2}$$

$$\bar{u}_M = \sum_{l=1}^{M} u^{(l)} / M. \tag{3}$$

L'analyste peut alors utiliser $\bar{q}_M$ pour estimer Q et $T_M = (1 + \frac{1}{M}) b_M + \bar{u}_M$ pour estimer la variance de $\bar{q}_M$. Quand n et M sont grands, les inférences pour le scalaire Q peuvent être fondées sur des lois normales, de sorte qu'un intervalle de confiance à $(1 - \alpha)\,\%$ pour Q est $\bar{q}_M \pm z(\alpha/2)\sqrt{T_M}$. Pour une valeur modérée de M, les inférences peuvent être fondées sur des lois t avec $v_M = (M - 1)(1 + r_M^{-1})^2$ degrés de liberté, où $r_M = (1 + M^{-1}) b_M / \bar{u}_M$, de sorte qu'un intervalle de confiance à $(1 - \alpha)\,\%$ pour Q est $\bar{q}_M \pm t_{vM}(\alpha/2)\sqrt{T_M}$. Des perfectionnements de ces règles de combinaison fondamentales ont été proposés par plusieurs auteurs, y compris Li, Raghunathan et Rubin (1991a), Li, Meng et Rubin (1991b), Raghunathan et Siscovick (1996), ainsi que Barnard et Rubin (1999).

3. Simulations illustratives

À la présente section, nous utilisons des simulations pour illustrer les biais/inefficacités associés à l'intégration des caractéristiques du plan dans les modèles d'imputation. Nous simulons trois populations cibles de $N = 100\,000$ unités, qui sont stratifiées et mises en grappes dans les strates. Dans la première population, Y dépend à la fois des effets de strate et de grappe. Dans la deuxième population, Y dépend des effets de strate, mais non des effets de grappe.

Dans la troisième population, Y n'est relié ni aux indicateurs de strate ni aux indicateurs de grappe. Nous utilisons la première population pour démontrer qu'il importe d'inclure toutes les variables du plan pertinentes, et les deuxième et troisième populations, pour examiner l'effet de l'inclusion de variables du plan non pertinentes. Les populations simulées sont stylisées afin d'illustrer l'importance de la modélisation du plan de sondage; par conséquent, la grandeur des biais/inefficacités n'est pas nécessairement généralisable à d'autres conditions.

Chaque population est divisée en cinq strates de taille égale comprenant chacune $N_h = 200$ grappes, pour $h = 1, \ldots, 5$. Chaque grappe c dans la strate h comprend N_{hc} unités. Dans chaque strate, 10 grappes ont $N_{hc} = 300$, 20 grappes ont $N_{hc} = 200$, 60 grappes ont $N_{hc} = 100$, 60 grappes ont $N_{hc} = 75$, et cinquante grappes ont $N_{hc} = 50$. Nous faisons varier les tailles de grappe afin de grossir les effets du plan lors du tirage d'échantillons en grappes à plusieurs degrés. Pour chaque population cible, il existe deux variables d'enquête, X et Y. Dans les trois populations, par souci de simplicité, nous générons chaque X_{hcj}, où l'indice j indique une unité dans la strate et la grappe hc, à partir de $X_{hcj} \sim N(0, 10^2)$. Pour générer Y, nous utilisons différentes méthodes pour chaque population, comme nous le décrirons aux sections qui suivent.

Nous échantillonnons aléatoirement les unités à partir de chaque population en utilisant un plan d'échantillonnage en grappes à plusieurs degrés. Pour commencer, nous tirons un échantillon aléatoire simple de $n_1 = 40$ grappes à partir de la strate 1, $n_2 = 20$ grappes à partir de la strate 2, $n_3 = 30$ grappes à partir de la strate 3, $n_4 = 10$ à partir de la strate 4 et $n_5 = 15$ grappes à partir de la strate 5. Les tailles des échantillons en grappes varient selon la strate afin de grossir les effets de plan comparativement à l'échantillonnage uniforme. Puis, nous tirons un échantillon aléatoire simple de 20 unités dans chaque grappe échantillonnée. Donc, nous obtenons 2 300 unités pour lesquelles $I_{hcj} = 1$.

Dans chaque population, les paramètres estimés sont $Q = \overline{Y}$, la moyenne de population de Y, et les coefficients de population de la régression de Y sur X. L'estimateur en données complètes de $\overline{Y}$ est l'estimateur sans biais fondé sur le plan de sondage habituel,

$$q = \frac{1}{100\,000}\left(\sum_{h=1}^{5} \frac{200}{n_h} \sum_{c \in h} \hat{y}_{hc}\right),$$

où $\hat{y}_{hc} = N_{hc}\overline{y}_{hc}$ est le total estimé dans la grappe hc. L'estimateur en données complètes de la variance de q est

$$u = \frac{1}{100\,000^2}\left(\begin{array}{l} \sum_{h=1}^{5} 200^2 \left(1 - \frac{n_h}{200}\right) s_h^2 / n_h \\ + \sum_{h=1}^{5} \frac{200}{n_h} \sum_{c \in h} N_{hc}^2 \left(1 - \frac{20}{N_{hc}}\right) s_{hc}^2 / 20 \end{array}\right),$$

où s_h^2 est la variance d'échantillon de $\hat{y}_{hc}$ et s_{hc}^2 est la variance d'échantillon de Y dans la grappe hc. Les estimateurs des coefficients dans la régression de Y sur X sont les estimateurs approximativement sans biais fondés sur le plan de sondage habituels, qui sont calculés en utilisant les routines « Survey » (Lumley 2004) du progiciel R. Ces routines estiment les variances selon des techniques de linéarisation par développement en série de Taylor. Ces estimateurs sont utilisés pour tous les ensembles de données multi-imputés dans toutes les simulations.

Pour chaque échantillon, nous posons que X est entièrement observée et que, pour Y, des données manquent pour environ 30 % des unités échantillonnées.

Pour chaque unité, la variable de réponse binaire, R_{hcj}, est tirée à partir d'une loi de Bernoulli :

$$\Pr(R_{hcj} = 1) = \frac{\exp(-0,847 - 0,1 X_{hcj})}{(1 + \exp(-0,847 - 0,1 X_{hcj}))} \qquad (4)$$

Ici, $R_{hcj} = 1$ signifie que la valeur de Y manque pour l'unité en question. L'équation 4 implique que Y_{mis} manque au hasard (Rubin 1976). Nous pouvons ignorer le mécanisme de création des données manquantes à condition que les imputations pour ces données soient conditionnelles à X. Délibérément, nous ne permettons pas que l'absence de données dépende de l'appartenance à la strate ou à la grappe, afin d'illustrer que le biais peut être dû au fait de ne pas tenir compte du plan de sondage, même si le mécanisme de création de données manquantes ignorable ne dépend pas du plan d'échantillonnage. Naturellement, si le plan d'échantillonnage est relié au fait que des données manquent, comme cela est le cas dans de nombreux ensembles de données réels, il faut introduire les contraintes du plan d'échantillonnage afin que le mécanisme de création des données manquantes soit ignorable.

Nous examinons trois stratégies d'imputation de Y_{mis} s'appuyant sur différentes utilisations de l'information sur le plan de sondage. Ces stratégies sont résumées au tableau 1. La première, dénotée EAS, omet entièrement de tenir compte du plan d'échantillonnage. La deuxième, dénotée FX, intègre la stratification et la mise en grappes grâce à l'utilisation d'effets fixes pour chaque grappe dans la strate. La troisième stratégie, dénotée HM, consiste à utiliser des modèles normaux à effets aléatoires dans lesquels sont intégrées la stratification et la mise en grappes. Pour EAS, un modèle est ajusté à l'ensemble de données complet. Pour FX et HM, les modèles sont ajustés séparément à chaque strate. Les trois stratégies comportent la régression sur X, parce que cette variable fait partie du mécanisme de création des données manquantes; ne pas conditionner à X violerait l'ignorabilité et causerait un biais.

Tableau 1
Stratégies d'imputation

Étiquette	Modèle d'imputation pour Y_{hcj} manquante
EAS	$N(\beta_0 + \beta_1 X_{hcj}, \sigma^2)$
FX	$N(\beta_{0h} + \beta_{1h} X_{hcj} + \omega_{hc}, \sigma_h^2)$
HM	$N(\beta_{0h} + \beta_{1h} X_{hcl} + \omega_{hc}, \sigma_h^2), \omega_{hc} \sim N(0, \tau^2)$

Toutes les imputations sont tirées d'après les lois bayésiennes prédictives a posteriori appropriées. Premièrement, nous sélectionnons les paramètres des modèles d'imputation à partir des lois a posteriori sachant les composantes des données observées, $(Z, X, Y_{\text{obs}}, I, R)$, qui sont incluses dans les modèles. Deuxièmement, nous sélectionnons les valeurs des données manquantes à partir des lois données au tableau 1. Nous utilisons des lois a priori diffusés pour tous les paramètres. Pour la stratégie HM, nous tirons les valeurs des paramètres en utilisant un échantillonneur de Gibbs (Gelfand et Smith 1990). Nous exécutons l'échantillonneur pendant une période de rodage pour obtenir la convergence approximative, puis nous utilisons chaque dixième tirage pour les imputations. Enfin, nous utilisons $M = 5$ imputations tirées indépendamment dans chaque ensemble de données pour chaque stratégie.

3.1 Simulation A : Illustration de la non-prise en compte des caractéristiques pertinentes du plan

Dans cette simulation, nous générons une population dans laquelle la distribution de Y diffère selon la strate et la grappe. Nous l'appelons « Population 1 ». Plus précisément, pour l'unité j dans la strate h et la grappe c, nous construisons la valeur de population de Y_{hcj} d'après

$$Y_{hcj} = 10 X_{hcj} + \beta_{0h} + \omega_{hc} + \in_{hcj} \quad (5)$$

où β_{0h} est une constante scalaire pour la strate h, ω_{hc} est une constante scalaire pour la grappe hc, et $\in_{hcj}$ est un terme d'erreur aléatoire tiré à partir de $N(0, 200^2)$. Les valeurs des effets de strate sont $\beta_{01} = 500, \beta_{02} = -250, \beta_{03} = 0, \beta_{04} = 250,$ et $\beta_{05} = -500$. Les valeurs de ω_{hc} sont obtenues en tirant cinq ensembles de $N_h = 200$ valeurs à partir de lois $N(0, 70^2)$ indépendantes. Les effets de strate et de grappe sont fortement dispersés afin de grossir les effets de plan comparativement à l'échantillonnage aléatoire simple, qui, à son tour, grossit les effets de la non-prise en compte du plan d'échantillonnage dans les imputations. Puis, nous tirons un échantillon à partir de cette population selon le plan d'échantillonnage en grappes stratifié décrit antérieurement. Nous créons l'indicateur de données manquantes R en utilisant l'équation 4.

Le tableau 2 montre les résultats de 1 000 répétitions des trois stratégies d'imputation décrites au tableau 1. La ligne supplémentaire annotée « Données complètes » donne les résultats en utilisant les données pour toutes les unités échantillonnées, c'est-à-dire en supposant qu'aucune unité pour laquelle $I_{hcj} = 1$ n'a $R_{hcj} = 0$. La colonne étiquetée « Couv. IC à 95 % » contient le pourcentage des 1 000 intervalles de confiance simulés qui contiennent le paramètre de population. La colonne étiquetée « Est. ponc. » contient les moyennes des 1 000 estimations ponctuelles de Q. La colonne étiquetée « Var. » contient les variances des 1 000 estimations ponctuelles de Q. La colonne étiquetée « Var. est. » contient les moyennes sur les 1 000 répétitions des variances estimées des estimations ponctuelles. Les colonnes étiquetées « Var(var. est.) » et « EQM(var. est.) » donnent la variance et l'erreur quadratique moyenne des 1 000 variances estimées.

Tableau 2
Propriétés des procédures d'imputation lorsque les caractéristiques du plan sont reliées à la variable d'enquête d'intérêt
La moyenne de population est égale à 3,2 et les coefficients de régression de population sont égaux à 3,0 et 10,1

	Méthode	Couv. IC à 95 %	Est. ponc.	Var.	Var. est.	Var. (var. est.)	EQM (var. est.)
Moyenne de Y	Données complètes	94,2	2,0	544,91	527,31	31 626,19	31 936,07
	EAS	38,0	45,8	327,79	360,74	11 927,97	13 013,35
	FX	94,8	2,4	554,09	579,92	37 474,82	38 141,70
	HM	94,5	2,3	551,02	553,16	34 056,39	34 060,99
Ordonnée à l'origine	Données complètes	93,0	2,4	529,51	499,73	18 543,13	19 430,21
	EAS	39,5	46,8	340,09	365,50	9 351,15	9 996,99
	FX	94,5	2,8	539,19	551,68	21 529,16	21 685,33
	HM	93,9	2,7	536,82	524,82	19 256,24	19 400,11
Pente	Données complètes	93,3	10,1	1,24	1,15	0,14	0,15
	EAS	64,8	7,6	2,10	2,20	0,55	0,56
	FX	94,5	10,1	1,45	1,44	0,18	0,18
	HM	95,7	10,1	1,53	1,65	0,29	0,30

Les imputations fondées sur la méthode EAS produisent des estimations gravement biaisées et une couverture très médiocre des intervalles de confiance dans cette population. Ces problèmes existent même si peu d'information manque et malgré le fait que nous utilisons des estimateurs sans biais par rapport au plan de sondage pour les inférences. Les méthodes FX et HM produisent toutes deux des estimations ponctuelles qui concordent approximativement avec les estimations ponctuelles basées sur les données complètes et donnent toutes deux des taux de couverture qui correspondent approximativement aux taux obtenus pour l'inférence d'après les données complètes. FX et HM ont des profils similaires, parce que les modèles à effets fixes et les modèles hiérarchiques produisent des estimations similaires des paramètres dans l'équation 5.

Lors de l'estimation de la moyenne de population, la variance associée à FX ou à HM n'est que légèrement plus grande que celle associée à l'estimateur d'après des données complètes. Il en est ainsi à cause des grands effets de grappe, qui font de la variance dans les cellules d'imputation un facteur dominant relativement à la variance entre cellules d'imputation. Autrement dit, la fraction d'information manquante due aux données manquantes est relativement faible comparativement à l'effet de la mise en grappes.

3.2 Simulation B : Illustration de l'inclusion de variables explicatives non pertinentes

La modélisation des caractéristiques du plan est essentielle quand ces dernières sont reliées aux variables d'enquête d'intérêt. Quelle est l'incidence de la modélisation de caractéristiques non pertinentes du plan sur les inférences? À la présente section, nous présentons les résultats de deux études par simulation réalisées en vue d'étudier cette question.

Premièrement, nous générons la « Population 2 » dans laquelle la distribution de Y diffère selon la strate, mais ne dépend pas des grappes. Pour cela, nous utilisons la même méthode que celle donnée par l'équation 5, en fixant ω_{hc} à zéro. Les $\in_{hcj}$ sont tirées à partir de $N(0, 100^2)$. Nous sélectionnons un échantillon à partir de la Population 2 et générons des données manquantes en utilisant les scénarios décrits antérieurement. Les résultats pour 1 000 répétitions sont présentés au tableau 3.

La méthode EAS continue de produire un biais important et une couverture médiocre des intervalles de confiance, parce qu'elle ne tient pas compte de la stratification. Pour les méthodes FX et HM, les moyennes des estimations ponctuelles se situent dans la marge d'erreur de simulation de la moyenne des estimations ponctuelles pour les données complètes. En outre, les taux de couverture des intervalles de confiance correspondent approximativement aux taux de couverture pour les intervalles obtenus d'après des données complètes. Donc, les méthodes FX et HM sont raisonnables pour ces populations, même si les caractéristiques de grappe non pertinentes sont incluses dans les modèles d'imputation.

Ensuite, nous générons la « Population 3 » dans laquelle la distribution de Y est indépendante des indicateurs d'appartenance aux strates et aux grappes. Plus précisément, pour générer Y, nous soustrayons β_{0h} des valeurs de Y générées dans la Population 2. Ensuite, nous tirons un échantillon à partir de la Population 3 en utilisant le plan d'échantillonnage en grappes stratifié et en créant des données manquantes selon les méthodes décrites antérieurement. Les résultats pour 1 000 répétitions sont présentés au tableau 4.

Tableau 3
Propriétés des procédures d'imputation lorsque la population comprend des effets de strate, mais non des effets de grappe
La moyenne de population est égale à 0,34 et les coefficients de régression de population sont égaux à 0,14 et 10,13

	Méthode	Couv. IC à 95 %	Est. ponc.	Var.	Var. est.	Var. (var. est.)	EQM (var. est.)
Moyenne de Y	Données complètes	93,6	0,37	468,97	461,88	29 301,77	29 352,04
	EAS	31,1	42,90	259,46	303,46	10 228,40	12 164,74
	FX	93,7	0,32	473,86	474,21	30 408,95	30 409,07
	HM	93,4	0,34	476,03	465,53	29 406,61	29 516,85
Ordonnée à l'origine	Données complètes	93,0	0,72	451,46	432,74	14 955,20	15 305,73
	EAS	31,5	43,10	275,22	311,36	8 134,04	9 440,57
	FX	93,2	0,66	456,08	444,88	15 539,21	15 664,64
	HM	92,3	0,68	457,48	436,25	14 941,00	15 391,75
Pente	Données complètes	93,1	10,09	0,99	0,91	0,09	0,10
	EAS	59,0	7,72	1,67	1,77	0,35	0,36
	FX	93,4	10,10	1,03	0,98	0,10	0,10
	HM	93,3	10,10	1,03	0,96	0,10	0,10

Tableau 4
Propriétés des procédures d'imputation lorsque les variables du plan sont entièrement non corrélées à la variable d'enquête d'intérêt
La moyenne de population est égale à 0,34 et les coefficients de régression de population sont égaux à 0,14 et 10,04

	Méthode	Couv. IC à 95 %	Est. ponc.	Var.	Var. est.	Var. (var. est.)	EQM (var. est.)
Moyenne de Y	Données complètes	94,7	0,35	14,61	14,73	32,65	32,66
	EAS	95,7	0,12	16,45	19,22	40,65	48,31
	FX	97,8	0,40	19,64	28,29	97,66	172,38
	HM	95,1	0,26	18,77	19,16	47,29	47,44
Ordonnée à l'origine	Données complètes	93,7	0,12	7,13	7,20	5,31	5,32
	EAS	96,8	−0,10	8,97	11,72	13,59	21,10
	FX	98,6	0,17	12,23	20,62	39,84	110,24
	HM	96,2	0,03	10,45	11,61	15,09	16,45
Pente	Données complètes	94,5	10,04	0,07	0,07	0,001	0,001
	EAS	96,4	10,07	0,10	0,13	0,002	0,003
	FX	96,4	10,04	0,12	0,15	0,003	0,004
	HM	95,2	10,05	0,11	0,12	0,002	0,002

La méthode EAS produit enfin des estimations ponctuelles dont les moyennes sont comprises dans la marge d'erreur de simulation de l'estimation ponctuelle moyenne d'après des données complètes. Il en est ainsi parce que l'imputation sous EAS reflète raisonnablement bien la structure de population. Il semble donc que ne pas tenir compte du plan d'échantillonnage dans les modèles d'imputation peut fournir des inférences acceptables lorsque les variables du plan ne sont que faiblement corrélées aux résultats de l'enquête. Comme dans les simulations antérieures, les méthodes FX et HM produisent des estimations ponctuelles moyennes comprises dans la marge d'erreur de simulation de l'estimation ponctuelle moyenne d'après des données complètes. Si nous comparons les trois stratégies d'imputation, nous voyons que FX et HM sont inefficaces comparativement à EAS, parce que les modèles d'imputation pour les deux premières méthodes estiment des paramètres qui sont approximativement nuls dans la population, tandis que dans la méthode EAS, leur valeur est fixée à zéro. La variance est plus faible pour HM que pour FX, parce que le modèle d'imputation hiérarchique lisse les effets de grappe estimés vers zéro.

Pour la méthode FX, le pourcentage d'intervalles de confiance qui couvrent Q est plus grand que le pourcentage observé pour les intervalles calculés d'après des données complètes et pour la méthode HM. Cela tient au fait que la variance estimée pour FX a tendance à être plus grande que la variance réelle. Ce biais par excès apparent dans T_M existe également dans le cas de la méthode EAS, ce qui donne un pourcentage de couverture plus grand que celui calculé pour les données complètes et la méthode HM.

4. Exemple fondé sur des données réelles

Nous allons maintenant examiner l'effet de la prise en compte de la stratification et de la mise en grappes lors de l'imputation pour traiter les données manquantes dans un ensemble de données réelles. Les données proviennent du fichier à grande diffusion des National Health and Nutrition Examination Surveys réalisées de 1999 à 2002. Les individus sont groupés en 56 grappes réparties entre 28 strates. De 5 % à 10 % de données manquantes sont relevées pour de nombreuses variables.

Nous avons imputé les données manquantes selon deux stratégies, l'une ne tenant pas compte des variables du plan (comme EAS) et l'autre intégrant les variables du plan au moyen d'effets fixes pour les indicateurs de grappes (comme FX). Dans le modèle d'imputation, nous avons inclus 27 variables nominales pour représenter 28 strates et une variable nominale dans chaque strate pour représenter les deux grappes emboîtées dans chaque strate. Autrement dit, nous avons inclus, en tout, 55 variables nominales à titre de variables explicatives. Nous avons utilisé une procédure de sélection séquentielle des variables pour repérer les interactions statistiquement significatives entre les variables nominales et les variables d'enquête, et nous avons également inclus ces interactions comme variables explicatives dans le modèle d'imputation. Nous avons imputé les valeurs suivant la méthode de régression séquentielle implémentée dans le progiciel IVEWARE (www.isr.umich.edu/src/smp/ive). Nous avons généré $M = 10$ ensembles de données pour chaque stratégie.

Nous considérons l'estimation de trois paramètres. Le premier est le pourcentage des personnes dans la population qui ont déjà fait vérifier leur taux de cholestérol (BPQ060). La proportion de données manquantes pour cette variable est d'environ 15 %. Les deuxième et troisième sont les coefficients de régression de population dans une régression logistique de BPQ060 sur le ratio revenu-seuil de pauvreté de la famille (INDFMPIR), variable continue pour laquelle la proportion de valeurs manquantes est d'environ 12 %. Ces paramètres sont estimés par des méthodes fondées sur

le plan de sondage au moyen des routines « Survey » du progiciel R.

Le tableau 5 donne les résultats pour les deux stratégies d'imputation. Pour toutes les analyses, les deux ensembles d'estimations sont fort semblables. Dans ce cas, l'intégration des variables du plan dans le modèle d'imputation n'a presque pas d'effets sur les résultats. Cela tient, en partie, aux faibles fractions d'information manquante et à l'insignifiance relative des effets de strate et de grappe. Cependant, la pénalité pour l'inclusion des caractéristiques du plan dans le modèle d'imputation est minime. À la lumière des résultats des simulations présentés à la section 3, nous intégrerions les caractéristiques du plan dans ce modèle d'imputation.

Tableau 5
Comparaison des résultats des données réelles lorsque les caractéristiques du plan sont incluses dans le modèle d'imputation et lorsque les caractéristiques du plan sont ignorées

	Est. ponc.	E.-t.	IC à 95 %
Moyenne de BPQ060			
Variables du plan	0,319	0,010	(0,299, 0,339)
Pas de variable du plan	0,319	0,011	(0,296, 0,341)
Ordonnée à l'origine : régression logistique			
Variables du plan	0,362	0,054	(0,256, 0,467)
Pas de variable du plan	0,352	0,052	(0,251, 0,454)
Pente : régression logistique			
Variables du plan	−0,409	0,020	(−0,449, −0,369)
Pas de variable du plan	−0,407	0,019	(−0,444, −0,371)

5. Conclusion

Quoique limitées, les études par simulation donnent à penser que ne pas tenir compte du plan d'échantillonnage dans l'imputation multiple peut être une pratique risquée. Lorsque les variables du plan sont corrélées aux variables d'enquête, comme dans notre simulation A, omettre de les inclure peut donner lieu à un biais important. Par ailleurs, l'inclusion de variables du plan non pertinentes, comme dans notre simulation B et dans l'exemple des enquêtes NHANES, produit, au pire, des inférences inefficaces et prudentes, lorsque les modèles d'imputation sont par ailleurs spécifiés correctement.

Inclure des variables nominales pour les effets de grappe réduit considérablement le biais comparativement à la non-prise en compte totale du plan. Cependant, l'introduction aveugle de variables nominales n'est pas une solution automatique. Lorsque la pente de la régression ou les variances diffèrent selon la grappe, l'utilisation de la méthode FX ou HM peut produire des estimations biaisées, puisque des caractéristiques importantes du plan sont omises. Les imputeurs qui soupçonnent l'existence de relations de ce genre devraient inclure les interactions appropriées avec les variables nominales pour les caractéristiques du plan, comme nous l'avons fait dans l'exemple des enquêtes NHANES. Dans le cas de certaines enquêtes, le plan peut être si complexe qu'il est impossible d'inclure des variables nominales pour chaque grappe. Le cas échéant, les imputeurs peuvent simplifier le modèle en ce qui concerne les variables du plan, par exemple en regroupant des catégories de grappes ou en incluant des variables de substitution (par exemple, taille de grappe) qui sont corrélées à la variable d'enquête d'intérêt.

Les simulations donnent à penser qu'il pourrait être avantageux d'utiliser des modèles hiérarchiques plutôt que des modèles à effets fixes pour l'imputation des données manquantes, particulièrement lorsque les effets de grappe sont semblables. Toutefois, les modèles hiérarchiques sont plus difficiles à ajuster que les modèles à effets fixes. Ainsi, l'ajustement de modèles hiérarchiques dans le cas de plans d'échantillonnage complexes lorsque des données manquent pour plusieurs variables continues et catégoriques est une tâche redoutable. Des modèles hiérarchiques séquentiels pourraient peut-être être ajustés dans un esprit semblable aux imputations par régression séquentielle de Raghunathan et coll. (2001). Il s'agit d'un domaine dans lequel devraient se poursuivre les travaux de recherche. Un autre inconvénient des modèles hiérarchiques est qu'il est plus facile de les spécifier incorrectement que les modèles à effets fixes. Ainsi, si les effets de grappe suivent une loi non normale, le modèle hiérarchique normal utilisé dans le présent article pourrait donner des imputations non plausibles.

Dans le cas de l'imputation multiple, la clé du succès réside dans la spécification d'un modèle d'imputation qui décrit raisonnablement la loi conditionnelle des valeurs manquantes sachant les valeurs observées. Souvent, les caractéristiques du plan sont corrélées aux variables d'enquête, de sorte que leur inclusion dans les modèles d'imputation réduit les risques d'erreur de spécification du modèle. Nous pensons que, dans de nombreux cas, les biais que peut causer l'exclusion de variables importantes, du plan ou d'autres variables reliées au mécanisme de création des données manquantes, surpassent les inefficacités qui pourraient résulter de l'estimation de petits coefficients. Cela renforce le conseil général fréquemment donné concernant l'imputation multiple : inclure toutes les variables qui sont reliées aux données manquantes dans les modèles d'imputation afin de rendre ignorable le mécanisme de création des données manquantes (par exemple, Meng 1994; Little et Raghunathan 1997; Schafer 1997; Collins, Schafer et Kam 2001).

Remerciements

La présente étude a été financée par la bourse ITR-0427889 de la National Science Foundation. Les auteurs remercient le rédacteur adjoint et les examinateurs de leurs commentaires et suggestions.

Bibliographie

Barnard, J., et Meng, X. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research*, 8, 17-36.

Barnard, J., et Rubin, D.B. (1999). Small-sample degrees of freedom with multiple-imputation. *Biometrika*, 86, 948-955.

Collins, L.M., Schafer, J.L. et Kam, C.K. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods*, 6, 330-351.

Gelfand, A.E., et Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Heitjan, D.F., et Little, R.J.A. (1991). Multiple imputation for the Fatal Accident Reporting System. *Applied Statistics*, 40, 13-29.

Kennickell, A.B. (1998). Multiple imputation in survey of consumer finances. Dans *Proceedings of the Section on Business and Economic Statistics*, American Statistical Association, 11-20.

Li, K.H., Raghunathan, T.E. et Rubin, D.B. (1991a). Large-sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, 86, 1065-1073.

Li, K.H., R.T.E., Meng, X.L. et Rubin, D.B. (1991b). Significance levels from repeated $p-$values with multiply-imputed data. *Statistica Sinica*, 1, 65-92.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Little, R.J.A., et Raghunathan, T.E. (1997). Should imputation of missing data condition on all observed variables? Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 617-622.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9, 8.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). *Statistical Science*, 9, 538-558.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J. et Solenberger, P. (2001). Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression. *Techniques d'enquête*, 27, 91-103.

Raghunathan, T.E., and Paulin, G.S. (1998). Multiple imputation of income in the Consumer Expenditure Survey: Evaluation of statistical inference. Dans *Proceedings of the Section on Business and Economic Statistics*, American Statistical Association, 1-10.

Raghunathan, T.E., Reiter, J.P. et Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1-16.

Raghunathan, T.E., et Siscovick, D.S. (1996). A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Applied Statistics*, 45, 335-352.

Reiter, J.P. (2003). Inférence pour les ensembles de microdonnées à grande diffusion partiellement synthétiques. *Techniques d'enquête*, 29, 203-211.

Reiter, J.P. (2004). Utilisation simultanée de l'imputation multiple pour les données manquantes et le contrôle de la divulgation. *Techniques d'enquête*, 30, 263-271.

Reiter, J.P. (2005). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society*, Séries A, 168, 185-205.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-590.

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

Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9, 462-468.

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

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

Schafer, J.L., Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A. et Rubin, D.B. (1998). The NHANES III multiple imputation project. Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 28-37.

Schenker, N., Raghunathan, T.E., Chiu, P.-L., Makuc, D.M., Zhang, G. et Cohen, A.J. Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association*, a paraître.

Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 169-175
Statistique Canada, N° 12-001 au catalogue

Bootstrap de type Bernoulli pour l'échantillonnage stratifié à plusieurs degrés

Fumio Funaoka, Hiroshi Saigo, Randy R. Sitter et Tsutom Toida [1]

Résumé

Nous proposons dans cet article une méthode de bootstrap de type Bernoulli facilement applicable à des plans stratifiés à plusieurs degrés où les fractions de sondage sont grandes, à condition qu'un échantillonnage aléatoire simple sans remise soit utilisé à chaque degré. La méthode fournit un ensemble de poids de rééchantillonnage qui donnent des estimations convergentes de la variance pour les estimateurs lisses ainsi que non lisses. La force de la méthode tient à sa simplicité. Elle peut être étendue facilement à n'importe quel nombre de degrés d'échantillonnage sans trop de complications. L'idée principale est de garder ou de remplacer une unité d'échantillonnage à chaque degré d'échantillonnage en utilisant des probabilités prédéterminées pour construire l'échantillon bootstrap. Nous présentons une étude par simulation limitée afin d'évaluer les propriétés de la méthode et, à titre d'illustration, nous appliquons cette dernière à l'Enquête nationale sur les prix menée en 1997 au Japon.

Mots clés : Enquête complexe; linéarisation; quantiles; rééchantillonnage; stratification.

1. Introduction

De nombreuses enquêtes à grande échelle sont réalisées selon un plan d'échantillonnage stratifié à plusieurs degrés. Or, lorsqu'on utilise ce genre de plan, l'estimation de la variance peut être analytiquement complexe, voire même impossible. En outre, pour les ensembles de données à grande diffusion, les formes particulières des estimateurs dont l'utilisateur pourrait souhaiter se servir pour obtenir les estimations de la variance sont inconnues. Par conséquent, des méthodes de rééchantillonnage sont souvent utilisées pour produire un ensemble de poids de rééchantillonnage qui peuvent être fournis avec l'ensemble de données et utilisés en vue d'estimer la variance pour une grande gamme d'estimateurs possibles. Le bootstrap est particulièrement utile, puisqu'il permet de traiter des statistiques d'échantillon lisses ainsi que non lisses sous des plans d'échantillonnage à plusieurs degrés. Un sommaire de plusieurs méthodes du bootstrap pour l'échantillonnage en population finie peut être consulté dans Shao et Tu (1995, pages 232–282) (voir aussi, Gross 1980; Bickel et Freedman 1984; McCarthy et Snowden 1985; Rao et Wu 1988; Kovar, Rao et Wu 1988; Sitter 1992a, b; Booth, Butler et Hall 1994; Shao et Sitter 1996).

Si la fraction de sondage de premier degré est faible, diverses méthodes du bootstrap existent pour traiter l'échantillonnage de premier degré comme s'il avait eu lieu avec remise afin d'estimer la variance. Dans le cas où les fractions de sondage de premier degré ne sont pas négligeables, un moins grand nombre de résultats sont disponibles. Pour le « bootstrappage » sous échantillonnage à deux degrés avec échantillonnage aléatoire simple (EAS) à chaque degré, voir Sitter (1992a, 1992b) et pour celui avec probabilités inégales, voir Rao et Wu (1988). Cependant, si les fractions de sondage de premier degré ne sont pas négligeables, aucune méthode du bootstrap simple n'existe pour trois degrés ou plus d'échantillonnage. Dans le présent article, nous proposons une nouvelle méthode du bootstrap qui permet de traiter facilement les cas pour lesquels un échantillonnage aléatoire simple (EAS) est utilisé à chaque degré. Nous l'appelons bootstrap de type Bernoulli (EBB) à cause de sa ressemblance à l'échantillonnage à partir d'une loi de Bernoulli. Nous utilisons les données de l'Enquête nationale sur les prix (ENP) du Japon pour l'illustrer.

Le plan de l'article est le suivant. À la section 2, nous présentons la notation pour l'échantillonnage stratifié à trois degrés. À la section 3, nous décrivons deux types d'EBB. À la section 4, nous étudions les propriétés de la méthode par simulation. À la section 5, nous décrivons le plan d'échantillonnage de l'ENP de 1997 et illustrons l'application de l'EBB aux données de l'ENP. Enfin, à la section 6, nous présentons nos conclusions.

2. Échantillonnage stratifié à trois degrés

Dans l'échantillonnage aléatoire stratifié, la population finie, constituée de N unités primaires d'échantillonnage (UPE) est fractionnée en H *strates* non chevauchantes contenant $N_1, N_2, \ldots, N_H$ UPE, respectivement; donc, $\sum_{h=1}^{H} N_h = N$. Un échantillon aléatoire simple sans remise (EASSR) d'UPE est tiré indépendamment dans chaque strate. Les tailles d'échantillon dans chaque strate sont dénotées par $n_1, n_2, \ldots, n_H$, et la taille totale de

1. F. Funaoka, professeur, Faculty of Economics, Shinshu University, 3-1-1 Asahi, Matsumoto, Nagano, 390-8621, Japon; H. Saigo, professeur, School of Political Science and Economics, Waseda University, 1-6-1 Nishiwaseda Shinjuku, Tokyo, 169-8050, Japon; R.R. Sitter, professeur, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, B.C., V5A 1S6, Canada; T. Toida, professeur associé, Faculty of Social and Information Studies, Gunma University, 2-4 Aramakicho, Maebashi, Gunma 371-8510, Japon.

l'échantillon d'UPE est $n = \sum_{h=1}^{H} n_h$. Au deuxième degré, un échantillon de m_{hi} unités secondaires d'échantillonnage (USE) est sélectionné à partir de l'UPE i de taille M_{hi} dans la strate h par EASSR. Au troisième degré, un échantillon de l_{hij} unités finales d'échantillonnage (UFE) est sélectionné à partir de l'USE ij de taille L_{hij} dans la strate h par EASSR. Un vecteur de mesures de certaines caractéristiques des unités est représenté par $\mathbf{y}_{hijk} = (y_{1hijk}, y_{2hijk}, y_{\tau hijk})^T$, où les indices inférieurs $hijk$ sont l'étiquette de strate, l'étiquette d'UPE, l'étiquette d'USE et l'étiquette d'UFE, respectivement. Le paramètre de population d'intérêt $\theta = \theta(\mathrm{S})$, où $\mathrm{S} = \{\mathbf{y}_{hijk} : h = 1, 2, \ldots, H; i = 1, 2, \ldots, N_h; j = 1, \ldots, M_{hi}; k = 1, \ldots, L_{hij}\}$, est habituellement estimé par $\hat{\theta} = \hat{\theta}(\mathrm{s})$, où $\mathrm{s} = \{\mathbf{y}_{hijk} : h = 1, 2, \ldots, H; i = 1, 2, \ldots, n_h; j = 1, \ldots, m_{hi}; k = 1, \ldots, l_{hij}\}$. Le vecteur des totaux de population est dénoté $\mathbf{Y} = (Y_1, \ldots, Y_\tau)^T$. Ici, son estimateur sans biais est :

$$\hat{\mathbf{Y}} = \sum_{h=1}^{H} \hat{\mathbf{Y}}_{\mathbf{h}} = \sum_{h=1}^{H} (N_h / n_h) \sum_{i=1}^{n_h} \hat{\mathbf{Y}}_{\mathbf{hi}},$$

où $\hat{\mathbf{Y}}_{\mathbf{hi}} = (M_{hi} / m_{hi}) \sum_{j=1}^{m_{hi}} \hat{\mathbf{Y}}_{\mathbf{hij}}$ et $\hat{\mathbf{Y}}_{\mathbf{hij}} = (L_{hij} / l_{hij}) \sum_{k=1}^{l_{hij}} \mathbf{y}_{\mathbf{hijk}}$, ce qui peut s'écrire sous la forme $\hat{\mathbf{Y}} = \sum_{hijk} w_{hij} \mathbf{y}_{\mathbf{hijk}}$, où $w_{hij} = (N_h / n_h)(M_{hi} / m_{hi})(L_{hij} / l_{hij})$.

Pour $\tau = 1$, une estimation sans biais de $\mathrm{Var}(\hat{Y})$ est $v(\hat{Y}) = \sum_{h=1}^{H} v(\hat{Y}_h)$, où

$$v(\hat{Y}_h) = \frac{N_h^2 (1 - f_{1h}) s_h^2}{n_h} + \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}^2 (1 - f_{2hi}) s_{hi}^2}{m_{hi}} + \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} \frac{L_{hij}^2 (1 - f_{3hij}) s_{hij}^2}{l_{hij}}$$

avec $\bar{\hat{Y}}_{h\cdot} = n_h^{-1} \sum_i \hat{Y}_{hi}$, $\bar{\hat{Y}}_{hi\cdot} = m_{hi}^{-1} \sum_j \hat{Y}_{hij}$, $\bar{y}_{hij\cdot} = l_{hij}^{-1} \sum_k y_{hijk}$, $f_{1h} = n_h / N_h$, $f_{2hi} = m_{hi} / M_{hi}$, $f_{3hij} = l_{hij} / L_{hij}$, $s_h^2 = \sum_i (\hat{Y}_{hi} - \bar{\hat{Y}}_{h\cdot})^2 / (n_h - 1)$, $s_{hi}^2 = \sum_j (\hat{Y}_{hij} - \bar{\hat{Y}}_{hi\cdot})^2 / (m_{hi} - 1)$, et $s_{hij}^2 = \sum_k (y_{hijk} - \bar{y}_{hij\cdot})^2 / (l_{hij} - 1)$ (Särndal, Swensson et Wretman 1992, pages 148–149).

3. Bootstrap de type Bernoulli proposé

Afin de traiter la question de l'échantillonnage à plusieurs degrés dans une strate, nous proposons un bootstrap à plusieurs degrés. Pour simplifier les idées, nous commençons par introduire une version simple, dont l'application présente certaines limites. Puis, nous décrivons une forme plus générale qui permet d'éviter ces difficultés.

EBB abrégé

Étape I. Pour chaque UPE de l'échantillon, hi, dans la strate $h, h = 1, \ldots, H$: a) la garder dans l'échantillon bootstrap avec la probabilité

$$p_h = \sqrt{1 - \frac{(1 - f_{1h})}{(1 - n_h^{-1})}}; \qquad (3.1)$$

ou b) la remplacer par une autre sélectionnée au hasard parmi les n_h UPE. Si l'opiton est a), passer à l'étape II.

Étape II. Pour chaque USE hij dans l'UPE hi de la strate h retenue à l'étape I : c) la garder dans l'échantillon bootstrap avec la probabilité

$$q_{hi} = \sqrt{1 - \frac{f_{1h}}{p_h^{-1}} \frac{(1 - f_{2hi})}{(1 - m_{hi}^{-1})}}; \qquad (3.2)$$

ou d) la remplacer par une autre sélectionnée au hasard parmi les m_{hi} USE dans l'UPE hi de la strate h. Si l'option est c), passer à l'étape III.

Étape III. Pour chaque UFE $hijk$ dans l'USE hij dans l'UPE hi de la strate h : e) la garder dans l'échantillon bootstrap avec la probabilité

$$r_{hij} = \sqrt{1 - \frac{f_{1h}}{p_h^{-1}} \frac{f_{2hi}}{q_{hi}^{-1}} \frac{(1 - f_{3hij})}{(1 - l_{hij}^{-1})}}; \qquad (3.3)$$

ou f) la remplacer par une autre sélectionnée au hasard parmi les l_{hij} UFE dans l'USE hij dans l'UPE hi de la strate h.

Soit K_{hij}^* le nombre de fois que l'unité $hijk$ figure dans la réplique bootstrap; alors, l'estimation du total par le bootstrap est $\hat{\mathbf{Y}}^* = \sum_{hijk} w_{hij}^* \mathbf{y}_{\mathbf{hijk}}$, où $w_{hij}^* = K_{hij}^* w_{hij}$, et l'estimation de $V(\hat{\theta})$ par le bootstrap est $v_B(\hat{\theta}) = V_*(\hat{\theta}^*)$, où $\hat{\theta}^* = \theta(\hat{\mathbf{Y}}^*)$ et V_* représente la variance sous la procédure de rééchantillonnage. Habituellement, l'estimation de la variance par le bootstrap est obtenue par simulation de Monte Carlo. Autrement dit, on répète les étapes I à III un grand nombre de fois, B, pour obtenir $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ et on utilise

$$v_B(\hat{\theta}) = \sum_{b=1}^{B} (\hat{\theta}_b^* - \bar{\theta}_{(\cdot)}^*)^2 / B,$$

où $\bar{\theta}_{(\cdot)}^* = \sum_{b=1}^{B} \hat{\theta}^* / B$. Dans la plupart des cas, il est possible de remplacer $\bar{\theta}_{(\cdot)}^*$ par $\hat{\theta}$. Cela permet au méthodologiste d'enquête de créer un ensemble de poids de rééchantillonnage w_{hij}^* pour chaque réplique bootstrap et de les inclure dans les fichiers de données à grande diffusion.

Il est clair que l'EBB abrégé n'est applicable que si $p_h, q_{hi}, r_{hij} \in [0, 1] \; \forall \; h, i, j$. Par exemple, il est nécessaire que $f_{1h} \geq n_h^{-1}$. Pour traiter les cas arbitraires $n_h, m_{hi}, l_{hij} \geq 2$, nous pouvons modifier chaque étape et changer p_h, q_{hi}, r_{hij} en conséquence.

EBB général

Étape I′. Tirer $(n_h - 1)$ UPE par EAS avec remise parmi les n_h UPE de l'échantillon, $h = 1, \ldots, H$.

Dénoter l'ensemble candidat par $\{\tilde{\text{UPE}}_{hi}: i = 1, 2, \ldots, n_h - 1\}$. Pour chaque UPE i dans l'échantillon de la strate h : a) la garder dans l'échantillon bootstrap avec la probabilité

$$p_h = 1 - \frac{1}{2}\frac{(1 - f_{1h})}{(1 - n_h^{-1})}; \qquad (3.4)$$

ou b) la remplacer par une autre sélectionnée au hasard à partir de $\{\tilde{\text{UPE}}_{hi}: i = 1, 2, \ldots, n_h - 1\}$. Si l'option est a), passer à l'étape II′.

Étape II′. Pour l'unité hi retenue à l'étape **I′**, tirer $(m_{hi} - 1)$ USE par EAS avec remise parmi les m_{hi} USE dans l'UPE hi. Dénoter l'ensemble candidat par $\{\tilde{\text{USE}}_{hij}: j = 1, 2, \ldots, m_{hi} - 1\}$. Pour chaque USE hij dans l'UPE hi retenue à l'étape **I′** : c) la garder dans l'échantillon bootstrap avec la probabilité

$$q_{hi} = 1 - \frac{1}{2}\frac{f_{1h}}{p_h^{-1}}\frac{(1 - f_{2hi})}{(1 - m_{hi}^{-1})}; \qquad (3.5)$$

ou d) la remplacer par une autre sélectionnée au hasard à partir de $\{\tilde{\text{USE}}_{hij}: j = 1, 2, \ldots, m_{hi} - 1\}$. Si l'option est c), passer à l'étape **III′**.

Étape III′. Pour l'unité hij retenue à l'étape **II′**, tirer $l_{hij} - 1$ UFE par EAS avec remise parmi les l_{hij} UFE dans l'USE hij dans l'UPE hi. Dénoter l'ensemble candidat par $\{\tilde{\text{UFE}}_{hijk}: k = 1, 2, \ldots, l_{hij} - 1\}$. Pour chaque UFE $hijk$ dans l'USE hij dans l'UPE hi : e) la garder dans l'échantillon bootstrap avec la probabilité

$$r_{hij} = 1 - \frac{1}{2}\frac{f_{1h}}{p_h^{-1}}\frac{f_{2hi}}{q_{hi}^{-1}}\frac{(1 - f_{3hij})}{(1 - l_{hij}^{-1})}; \qquad (3.6)$$

ou f) la remplacer par une autre sélectionnée au hasard à partir de $\{\tilde{\text{UFE}}_{hijk}: k = 1, 2, \ldots, l_{hij} - 1\}$.

Il est facile de voir que $p_h, q_{hi}, r_{hij} \in [0, 1] \; \forall \; n_h, m_{hi}, l_{hij} \geq 2$.

La raison justifiant la sélection aléatoire d'un ensemble candidat dans l'EBB général est la suivante. Pour fixer les idées, considérons un EASSR à un degré dans une seule strate. Soit $\bar{y}^*$ une moyenne d'échantillon bootstrap sous l'EBB abrégé avec une probabilité arbitraire $p \in [0, 1]$. Alors, nous pouvons montrer que $V_*(\bar{y}^*) = n^{-1}(1 - n^{-1})\, s^2(1 - p^2)$, où $s^2 = \sum_i (y_i - \bar{y})^2/(n-1)$. Notons que $V_*(\bar{y}^*)$ est monotone décroissante par rapport à p dans l'intervalle $[0, 1]$. Donc, $\min_{p \in [0,1]} V_*(\bar{y}^*) = 0$ et $\max_{p \in [0,1]} V_*(\bar{y}^*) = n^{-1}(1 - n^{-1})s^2$. Si $f_1 < n^{-1}$, puis $\max_p V_*(\bar{y}^*) < v(\bar{y})$. La notion clé de l'EBB général est que nous pouvons rendre $\max_p V_*(\bar{y}^*)$ plus grand que $v(\bar{y})$ en introduisant une variation supplémentaire dans le remplacement des unités grâce à la sélection aléatoire d'un ensemble candidat.

Nous pouvons montrer que l'EBB abrégé et l'EBB général produisent une estimation convergente de la variance pour les fonctions lisses des totaux de population estimés. En outre, sous des conditions de régularité appropriées pour la fonction de répartition de la population, ils produisent aussi une estimation convergente de la variance pour les quantiles d'échantillon. Qui plus est, dans les deux méthodes EBB, la taille des répliques est égale à celle de l'échantillon original, propriété qui peut être désirable lorsque l'on a affaire à des données d'enquête imputées (voir Saigo, Shao et Sitter 2001).

Il n'est pas difficile d'étendre l'approche de l'EBB à des plans comportant plus de trois degrés. Par exemple, pour un plan stratifié à quatre degrés, une UFE au quatrième degré dans la strate h est retenue avec la probabilité

$$\sqrt{1 - p_h^{-1} f_{1h} q_{hi}^{-1} f_{2hi} r_{hij}^{-1} f_{3hij} (1 - g_{hijk}^{-1})^{-1} (1 - f_{4hijk})}$$

ou remplacée dans l'EBB abrégé, où g_{hijk} est la taille d'échantillon de quatrième degré et f_{4hijk} est la fraction de sondage de quatrième degré. Les extensions ultérieures sont analogues.

L'EBB général randomise un ensemble candidat simplement pour remédier à l'infaisabilité de l'EBB abrégé. Ce concept présente des similarités avec le bootstrap approximativement bayésien de Rubin et Schenker (1986).

Un inconvénient de l'EBB général comparativement à l'EBB abrégé est que le premier nécessite, en moyenne, $\sum_h \{(n_h - 1) + p_h \sum_i (m_{hi} - 1) + p_h \sum_i q_{hi} \sum_j (l_{hij} - 1)\}$ générations de nombres aléatoires de plus que le second, où $p_h, q_{hi},$ et r_{hij} sont donnés par (3.4), (3.5) et (3.6), respectivement, ce qui peut demander beaucoup de temps lorsque les tailles d'échantillon et(ou) le nombre de strates sont grands. Afin de réduire les générations de nombres aléatoires dans l'EBB général, on peut créer un ensemble candidat en supprimant aléatoirement une unité de l'échantillon original et utiliser

$$p_h = (n_h + 1/2) - \sqrt{(n_h + 1/2)^2 - n_h(1 + f_{1h})}, \qquad (3.7)$$

$$q_{hi} = (m_{hi} + 1/2) - \sqrt{(m_{hi} + 1/2)^2 - f_{h1} p_h^{-1} m_{hi}(1 + f_{2hi})}, \qquad (3.8)$$

$$r_{hij} = (l_{hij} + 1/2) - \sqrt{(l_{hij} + 1/2)^2 - f_{h1} p_h^{-1} f_{2hi} q_{hi}^{-1} l_{hij}(1 + f_{3hij})}, \qquad (3.9)$$

au lieu des trois équations susmentionnées. On peut montrer que $p_h, q_{hi}, r_{hij} \in [0, 1]$. La preuve de cette version modifiée de l'EBB général est similaire.

4. Une étude par simulation

À la présente section, nous décrivons l'exécution de simulations limitées pour étudier l'EBB dans le cas de l'estimation par le ratio et de l'estimation par quantile. Pour simplifier, nous considérons un EASSR à deux degrés et nous nous limitons à une seule strate.

4.1 Description générale de la simulation

Une population finie unistratifiée est générée selon la procédure qui suit et est maintenue fixe pour toutes les exécutions de la simulation afin d'observer les propriétés fondées sur le plan de sondage de l'EBB. Premièrement, la moyenne des variables auxiliaires dans la grappe i est générée par $\mu_i \sim N(\mu, \sigma^2)$ pour $i = 1, 2, \ldots, N$. Puis, la variable auxiliaire x_{ik} de l'unité k dans la grappe i est générée par

$$x_{ik} = \mu_i + \varepsilon_{ik} (k = 1, 2, \ldots, M_i; i = 1, 2, \ldots, N), \quad (4.1)$$

où $\varepsilon_{ik} \sim N(0, (1-\rho)\sigma^2/\rho)$. La variable cible y_{ik} de l'unité k dans la grappe i est obtenue par

$$y_{ik} = a + bx_{ik} + e_{ik} (k = 1, 2, \ldots, M_i; i = 1, 2, \ldots, N), \quad (4.2)$$

où $e_{ik} \sim N(0, \sigma^2/4)$. Les valeurs des paramètres sont fixées à $\mu = 100, \sigma = 10, \rho = 0{,}1(0{,}3), a = 0$ et $b = 1$, et l'EASSR à deux degrés est utilisé tout au long de l'étude par simulation.

4.2 Estimation par le ratio

Soit $N = 50, n = 15, M_i = 20$ et $m_i = 3$, for $i = 1, \ldots, n$. Considérons l'estimateur par le ratio du total de population, Y,

$$\hat{Y}_R = \hat{R}\, X,$$

où $X = \sum_{i=1}^{N} \sum_{k=1}^{M_i} x_{ik}$ est le total de population des x, $\hat{R} = \hat{Y}/\hat{X}, \hat{Y} = \sum_{h=1}^{H} \hat{Y}_h = \sum_{h=1}^{H} (N_h/n_h) \sum_{i=1}^{n_h} \hat{Y}_{hi}, \hat{X} = \sum_{h=1}^{H} \hat{X}_h = \sum_{h=1}^{H} (N_h/n_h) \sum_{i=1}^{n_h} \hat{X}_{hi}, \hat{Y}_{hi} = (M_{hi}/m_{hi}) \sum_{k=1}^{m_{hi}} \hat{Y}_{hik}$ et $\hat{X}_{hi} = (M_{hi}/m_{hi}) \sum_{k=1}^{m_{hi}} \hat{X}_{hik}$.

Aux fins de comparaison, nous considérons un certain nombre d'estimateurs de la variance utilisables dans ce simple contexte :

1) L'estimateur classique de la variance est dénoté

$$v_0(\hat{Y}_R) = N^2 \frac{1-f_1}{n} \frac{\sum_i (\hat{Y}_i - \hat{R}\hat{X}_i)^2}{n-1} + \frac{N}{n} \sum_i \frac{M_i^2 (1-f_{2i}) s_{d'2i}^2}{m_i}, \quad (4.3)$$

où $f_1 = n/N, f_{2i} = m_i/M_i$ et

$$s_{d'2i}^2 = \sum_j (y_{ij} - \hat{R}x_{ij})^2 / (m_i - 1).$$

2) L'estimateur par le jackknife avec suppression d'une UPE à la fois corrigé pour la fraction de sondage de premier degré, parfois utilisé même s'il n'est pas entièrement correct, est dénoté

$$v_{cj}(\hat{Y}_R) = (1-f_1) \frac{n-1}{n} \sum_i (\hat{Y}_{R(i)} - \hat{Y}_{R(\cdot)})^2, \quad (4.4)$$

où $\hat{Y}_{R(i)}$ est l'estimateur recalculé après l'élimination de la i^{e} UPE et $\hat{Y}_{R(\cdot)} = \sum_i \hat{Y}_{R(i)}/n$.

3) Un estimateur par le jackknife pondéré extérieurement (voir Folsom, Bayless et Shah 1971) qui comprend une correction pour les deux degrés d'échantillonnage peut être dérivé sous la forme

$$v_{ewj}(\hat{Y}_R) = (1-f_1) \frac{n-1}{n} \sum_i (\hat{Y}_{R(i)} - \hat{Y}_{R(\cdot)})^2 + f_1 \sum_i (1-f_{2i}) \frac{m_i - 1}{m_i} \sum_j (\hat{Y}_{R(ij)} - \hat{Y}_{R(\cdot\cdot)})^2, \quad (4.5)$$

où $\hat{Y}_{R(i)}$ est la i^{e} pseudo valeur jackknife obtenue par suppression de l'UPE i, $\hat{Y}_{R(ij)}$ est la ij^{e} pseudo valeur jackknife obtenue par suppression de l'unité j dans l'UPE i, $\hat{Y}_{R(\cdot)} = \sum_i \hat{Y}_{R(i)}/n$ et $\hat{Y}_{R(\cdot\cdot)} = \sum_j \hat{Y}_{R(ij)}/m_i$.

4) Il existe aussi un estimateur de la variance assisté par modèle (voir Särndal, Swensson et Wretman (1992), équation (8.10.6)) de la forme

$$v_{ma}(\hat{Y}_R) = (X/\hat{X})^2 v_0(\hat{Y}_R). \quad (4.6)$$

Nous utilisons $B = 100$ répliques bootstrap dans chacune des $S = 1\,000$ exécutions de la simulation. Nous obtenons une approximation des EQM réelles d'après 10 000 exécutions de la simulation et nous utilisons les estimations de Monte Carlo du biais relatif en pourcentage et du coefficient de variation des divers estimateurs de la variance, ainsi que les probabilités empiriques de couverture des intervalles de confiance à 90 %, comme mesures de leur performance relative.

Nous voyons au tableau 1 que $v_{\text{EBB}}, v_0, v_{ewj}$ et v_{ma} donnent des résultats comparables et bons, excepté que le coefficient de variation (cv) des méthodes de rééchantillonnage est un peu plus élevé que celui des méthodes sans rééchantillonnage, ce qui est typique. Le jackknife avec suppression d'une UPE à la fois donne des résultats médiocres.

Tableau 1
Comparaison des estimateurs de la variance pour $\hat{Y}_r$

ρ		Biais en %	CV	Couverture (90 %)
0,1	v_0	−1,70	0,28	89,2
	v_{EBB}	−0,62	0,33	88,9
	v_{ewj}	−0,33	0,30	89,4
	v_{cj}	−26,55	0,39	80,5
	v_{ma}	−0,39	0,30	89,4
0,3	v_0	−0,67	0,28	86,6
	v_{EBB}	−1,63	0,33	86,5
	v_{ewj}	−0,74	0,29	86,5
	v_{cj}	−26,85	0,39	80,2
	v_{ma}	−0,87	0,29	86,4

Afin d'étudier les propriétés conditionnelles, nous avons ordonné les 1 000 exécutions de la simulation selon $X/\hat{X}$ et réparti les exécutions en 20 groupes de taille égale. Pour chaque groupe, nous avons calculé la moyenne de chaque estimateur de la variance. La figure 1 représente ces moyennes groupées pour chaque estimateur de la variance (sauf v_{cj} puisqu'il présente un biais négatif important) en fonction de la moyenne groupée $X/\hat{X}$, pour $\rho = 0,3$. L'EQM réelle est incluse dans le tracé également. Le graphique est semblable à celui utilisé par Royall et Cumberland (1981a, 1981b). Nous voyons que v_{EBB} suit l'EQM réelle, en grande partie comme v_{ewj} et v_{ma}, tandis que v_0 ne le fait pas. Donc, l'EBB semble avoir une propriété conditionnelle désirable.

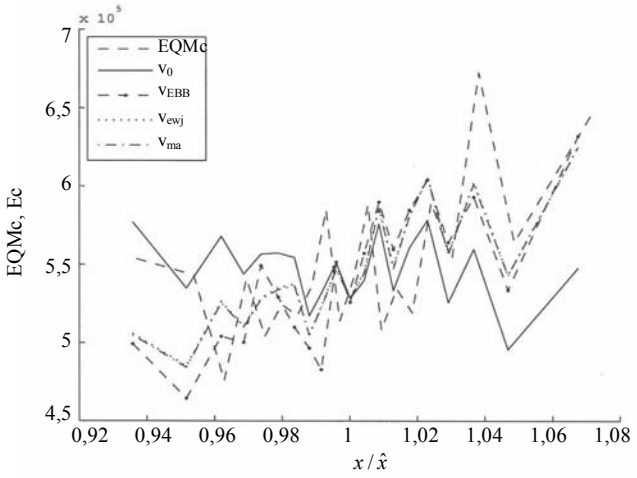

Figure 1. EQMc et Ec(v) pour l'estimation par le ratio.

4.3 Estimation par quantile

Pour l'estimation par quantile, nous posons que $N = 100, n = 30, M_i = 100$ et $m_i = 10$, pour $i = 1, \ldots, n$. Nous utilisons $B = 500$ répliques bootstrap dans chacune des $S = 5\,000$ exécutions de la simulation. Nous obtenons une approximation des EQM réelles au moyen de 50 000 exécutions de la simulation. Seuls les résultats pour v_{EBB} et v_{ewj} quand $\rho = 0,1$ sont résumés au tableau 2, parce que ceux obtenus quand $\rho = 0,3$ sont similaires. Nous voyons que la méthode de l'EBB donne d'assez bons résultats, avec un léger biais par excès, tandis que la méthode du jackknife pondérée extérieurement produit un biais important, à cause de son absence de convergence dans l'estimation de la variance pour les quantiles.

Tableau 2
Propriétés de v_{EBB} et v_{ewj} pour les quantiles 0,10, 0,25, 0,50, 0,75 et 0,90

	v_{BBE}			v_{ewj}		
Quantile	Biais en %	CV	Couverture (90 %)	Biais en %	CV	Couverture (90 %)
0,10	8,40	0,51	87,7	51,87	1,93	81,3
0,25	6,21	0,42	88,2	21,19	1,28	83,3
0,50	2,53	0,37	87,4	14,27	1,00	83,0
0,75	6,23	0,42	87,8	28,07	1,33	83,4
0,90	6,32	0,50	88,0	54,47	2,05	80,3

5. Application à l'Enquête nationale sur les prix menée en 1997 au Japon

L'objectif de l'ENP est d'analyser la formation des prix des principaux biens de consommation, comme les aliments, les vêtements et les appareils électroménagers. L'estimation par quantile joue un rôle essentiel dans cette analyse, et de nombreuses estimations par quantile fondées sur plusieurs stratifications a posteriori sont incluses dans les rapports de l'ENP.

L'échantillonnage stratifié à plusieurs degrés utilisé dans l'ENP de 1997 se résume comme suit :

Stratification. Les municipalités forment les UPE et sont réparties en 537 strates, d'abord en fonction des préfectures et des sphères économiques que constitue chaque municipalité, puis d'après la taille de leur population.

Échantillonnage de premier degré. Ces UPE sont sélectionnées par EASSR indépendamment dans chaque strate. Le tableau 3 donne un aperçu des fractions de sondage de premier degré.

Échantillonnage de deuxième degré. Dans une municipalité sélectionnée, tous les grands points de vente sont dénombrés. Autrement dit, un échantillonnage en grappes à un degré est utilisé pour ces points de vente. Pour les petits points de vente, par contre, une municipalité échantillonnée est subdivisée en régions d'enquête (USE), chacune constituée d'environ 100 points de vente. Un échantillonnage systématique est utilisé pour échantillonner les régions d'enquête. Les fractions de sondage au deuxième degré sont comprises entre 0,1 et 1,0.

Échantillonnage de troisième degré. Dans chaque région d'enquête sélectionnée, 40 points de vente (UFE) sont choisis par échantillonnage systématique ordonné en fonction du type de point de vente et du chiffre de ventes annuel déclaré lors du Recensement du commerce de 1994.

Tableau 3
Fractions de sondage de premier degré dans l'ENP de 1997

Catégorie de région	Taille de la population	N^bre d'UPE	Fractions de sondage	Taille de l'échantillon
Villes	≥ 100 000	221	1/1	221
Villes	50 000 – 99 999	220	2/3	179
Villes	< 50 000	224	1/3	80
Petites villes et villages	≥ 40 000	32	1/5	4
Petites villes et villages	< 40 000	2 536	1/15	187

À proprement parler, il n'existe aucune formule de variance valide pour les données de l'ENP, parce que celles-ci comportent un échantillonnage systématique. Cependant, pour estimer la variance, nous supposons que l'échantillonnage systématique peut être approximé par l'EASSR. Même sous cette condition simplifiée, il n'existe

aucune expression analytique explicite de la variance pour les quantiles d'échantillon. En fait, aucune estimation de la variance n'est associée aux estimations des quantiles de prix dans le rapport de l'ENP, tandis que les prix moyens sont publiés avec l'estimation de leur variance.

À la présente section, nous appliquons l'EBB abrégé aux données de l'ENP, en supposant que l'échantillonnage systématique peut être approximé par l'EASSR. Certaines strates ne contiennent qu'une seule UPE. En outre, $f_{1h} < n_h^{-1}$ dans certaines strates. Les strates de ce genre sont intégrées à des strates adjacentes de sorte que p_h, donnée par (3.1), soit comprise dans l'intervalle $[0, 1]$. Après regroupement, il existe plus de 280 strates. Nous supposons que l'effet de la reformation des strates est négligeable.

Après reformation des strates, nous employons l'EBB abrégé dans les strates composées de grandes villes. Par ailleurs, nous utilisons le bootstrap avec remise (Shao et Tu 1995, page 247) où la taille des répliques est $(n_h - 1)$ dans les strates composées de petites villes et de villages, où les fractions de sondage de premier degré sont faibles. Les estimations par quantile et leurs erreurs-types pour certains produits vendus par les petits points de vente sont présentées au tableau 4. Notons que les prix d'un produit donné sont discrets. Cependant, nous appliquons le bootstrap comme s'ils étaient continus. Cette approximation devrait être acceptable pour de nombreux produits, mais non pour ceux qui sont très bon marché, puisque, dans ce cas, un pourcentage élevé d'observations est concentré sur un prix particulier et l'erreur-type estimée peut être nulle.

6. Conclusion

Le bootstrap est utile pour estimer les variances dans le cas des enquêtes complexes, particulièrement lorsque l'estimation par quantile est importante. Nous avons proposé deux méthodes du bootstrap de type Bernoulli qui permettent de traiter facilement les plans EASSR stratifiés à plusieurs degrés où les fractions de sondage sont grandes : l'EBB abrégé et l'EBB général. Dans les deux méthodes, une unité d'échantillonnage à un degré donné est soit retenue, soit remplacée avec une probabilité prédéterminée, afin de construire un échantillon bootstrap. L'EBB général a l'avantage de permettre le traitement de toute combinaison de tailles d'échantillon ≥ 2, mais il nécessite plus de générations de nombres aléatoires que l'EBB abrégé. À titre d'illustration, nous avons appliqué l'EBB abrégé aux données de l'Enquête nationale sur les prix menée en 1997 au Japon.

Remerciements

Les travaux du deuxième auteur ont été financés par la Japan Statistical Association. Ceux du troisième auteur ont été financés par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada. Les auteurs remercient le Bureau de la statistique, le ministère de la Gestion publique, des Affaires intérieures, des Postes et des Télécommunications, ainsi que le ministère de l'Économie, du Commerce et de l'Industrie du Japon d'avoir fourni les données de l'ENP de 1997.

Tableau 4
Quantiles d'échantillon (erreurs-types) de certains produits pour les petits points de vente dans l'ENP

Produit	p	0,10	0,25	0,5	0,75	0,90
Riz (5kg)[a]	Quantile d'échantillon	239,4	255,2	278,3	299,1	315,0
(10 yens)	(erreur-type)	(0,24)	(0,53)	(0,21)	(0,02)	(0,61)
Café instantané (1 flacon)[b]	Quantile d'échantillon	714	788	859	893	914
(yen)	(erreur-type)	(0,13)	(0,40)	(0,00)	(2,68)	(1,43)
Bière (24 cannettes)[c]	Quantile d'échantillon	467,3	500,0	536,8	549,4	549,4
(10 yens)	(erreur-type)	(1,01)	(0,64)	(0,82)	(0,00)	(0,00)
PC[d]	Quantile d'échantillon	248,8	260,4	299,3	346,5	375,9
(1 000 yens)	(erreur-type)	(2,03)	(0,35)	(3,25)	(7,17)	(1,48)

Marques spécifiées : [a]Koshihikari; [b]Nescafe Gold Blend, 100g; [c]Sapporo (Nama) Black Label, 350ml; [d]NEC PC9821 NW133/D14.

Bibliographie

Bickel, P.J., et Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statististics*, 12, 470-482.

Booth, J.G., Butler, R.W. et Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89, 1282-1289.

Folsom, R.E., Bayless, D.L. et Shah, B.V. (1971). Jackknifing for variance components in complex sample survey designs. *Proceedings of the Social Statistics Section*, American Statistical Association, 36-39.

Gross, S. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 181-184.

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

McCarthy, P.J., et Snowden, C.B. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics*, Série 2, 95, Public Health Service Publication, 85-1369, Washington, DC : U.S. Government Printing Office.

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

Royall, R.M., et Cumberland, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.

Royall, R.M., et Cumberland, W.G. (1981b). The finite population linear regression estimator: An empirical study. *Journal of the American Statistical Association*, 76, 924-930.

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

Saigo, H., Shao, J. et Sitter, R.R. (2001). Bootstrap à demi-échantillon répété et répliques équilibrées répétées en cas d'imputation aléatoire de données. *Techniques d'enquête*, 27, 209-218.

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

Shao, J., et Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.

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

Sitter, R.R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.

Sitter, R.R. (1992b). Comparing three bootstrap methods for survey data. *Canadian Journal of Statististics*, 20, 135-154.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 177-183
Statistique Canada, N° 12-001 au catalogue

Approche de la stratification par une méthode géométrique et par optimisation : Une comparaison de l'efficacité

Marcin Kozak et Med Ram Verma [1]

Résumé

L'article donne une comparaison des approches de la stratification par une méthode géométrique, par optimisation et par la méthode de Lavallée et Hidiroglou (LH). L'approche géométrique de stratification est une approximation, tandis que les deux autres, qui s'appuient sur des méthodes numériques, peuvent être considérées comme des méthodes de stratification optimales. L'algorithme de la stratification géométrique est très simple comparativement à ceux des deux autres approches, mais il ne tient pas compte de la construction d'une strate à tirage complet, qui est habituellement produite lorsque l'on stratifie une population positivement asymétrique. Dans le cas de la stratification par optimisation, on peut prendre en considération toute forme de la fonction d'optimisation et de ses contraintes. Une étude numérique comparative portant sur cinq populations artificielles positivement asymétriques a indiqué que, dans chaque cas étudié, l'approche par optimisation était plus efficace que la stratification géométrique. En outre, nous avons comparé les approches géométrique et par optimisation à l'algorithme LH. Cette comparaison a révélé que la méthode géométrique de stratification était moins efficace que l'algorithme LH, tandis que l'approche par optimisation était aussi efficace que cet algorithme. Néanmoins, les limites de strate déterminées par la stratification géométrique peuvent être considérées comme de bons points de départ pour l'approche par optimisation.

Mots clés : Stratification optimale; stratification géométrique; optimisation numérique; algorithme de Lavallée-Hidiroglou.

1. Introduction

Gunning et Horgan (2004) ont proposé un algorithme de stratification basé sur une progression géométrique. Par souci de simplicité, nous appellerons cette technique « approche géométrique de stratification », « stratification géométrique » ou simplement « approche géométrique ». La stratification géométrique vise à produire des valeurs égales du coefficient de variation d'une variable de stratification dans les diverses strates, en émettant l'hypothèse que la variable suit une loi uniforme dans chaque strate. Gunning et Horgan (2004) ont montré que leur algorithme est nettement plus facile à appliquer et plus efficace que la méthode classique de la fonction cumulative de la racine carrée des fréquences (Dalenius et Hodges 1959) et que l'algorithme de Lavallée et Hidiroglou (LH) (Lavallée et Hidiroglou 1988). Horgan (2006) a comparé la stratification géométrique aux méthodes de Dalenius et Hodges (1959), d'Ekman (1959), et de Lavallée et Hidiroglou (1988); de nouveau, son étude a montré que la stratification géométrique était la plus efficace parmi les méthodes comparées. Gunning, Horgan et Yancey (2004) ont appliqué cette méthode en vue de stratifier des populations comptables.

À l'instar de la méthode de la fonction cumulative de la racine carrée des fréquences, l'approche géométrique est une technique de stratification approximative, si bien que les points de stratification qu'elle fournit peuvent s'écarter considérablement des points de stratification optimaux. Par ailleurs, il existe des approches, particulièrement pour la stratification univariée, qui produisent des stratifications quasi optimales. Ces approches sont fondées sur l'utilisation d'algorithmes auto-appliqués ou de méthodes numériques d'optimisation pour produire les limites de strate (par exemple, Lavallée et Hidiroglou 1988; Lednicki et Wieczorkowski 2003; Kozak 2004). Toutefois, les méthodes de ce genre requièrent habituellement des limites initiales pour lancer le processus d'optimisation; les méthodes de stratification approximatives peuvent être utilisées pour rechercher ces points initiaux. Naturellement, les limites de strate initiales doivent être de haute qualité; sinon, l'optimisation risque de fournir un minimum local (Rivest 2002).

De nombreuses enquêtes comportent des variables d'intérêt positivement asymétriques. Le cas échéant, il est important de tenir compte de cet attribut lors de la stratification d'une population. Nombre de chercheurs ont essayé de créer des méthodes de stratification permettant de construire une strate dite « à tirage complet » (par exemple, Glasser 1962; Hidiroglou 1986) dont tous les éléments sont sélectionnés dans l'échantillon avec une probabilité égale à 1. Dans le contexte de l'échantillonnage stratifié, il s'agit du meilleur moyen de traiter les variables positivement asymétriques. Ces méthodes sont habituellement plus efficaces (de façon certaine, uniquement si une population est

1. Marcin Kozak, département de biométrie, Université agricole de Varsovie, Nowoursynowska 159, 02-776 Varsovie, Pologne. Courriel : marcin.kozak@omega.sggw.waw.pl; Med Ram Verma, Division of Agricultural Economics & Statistics, ICAR Research Complex for N.E.H. Region, Umroi Road, Umiam (Barapani) Meghalaya, Inde, Pin 793 103. Courriel : mrverma19@yahoo.co.in.

positivement asymétrique) que les méthodes de stratification ne comportant pas la construction d'une strate à tirage complet. La stratification géométrique ne comprend pas la création d'une telle strate (Gunning et Horgan 2004).

Le but du présent article est de comparer l'efficacité de la stratification géométrique, proposée par Gunning et Horgan (2004) à celle de deux approches de stratification par optimisation (Lavallée et Hidiroglou 1988; Lednicki et Wieczorkowski 2003; Kozak 2004) fondées sur l'utilisation de méthodes numériques d'optimisation.

2. Approches de stratification comparées

Supposons que nous souhaitions stratifier une population positivement asymétrique de N unités, U, en nous fondant sur un vecteur $\mathbf{x} = (x_1, \ldots, x_N)^T$ de dimension N connu dès le départ (c'est-à-dire avant le début de l'étude) des valeurs d'une variable de stratification X.

Dans le présent article, nous considérons deux problèmes de stratification. Le premier consiste à construire L strates sachant la taille fixe d'échantillon n. Supposons que nous recherchions un vecteur de dimension $(L+1)$ de limites de strate $\mathbf{k} = (k_0, \ldots, k_L)^T$, $(k_0 < k_1 < \ldots < k_L, k_0$ étant la valeur minimale et k_L la valeur maximale de X) qui minimise la variance d'un estimateur de la moyenne de population de X sous échantillonnage stratifié avec échantillonnage aléatoire simple sans remise dans les strates (*STSI*) et combiné à une approche avec strate à tirage complet. (Il convient de souligner que nous traitons la variable de stratification comme étant identique à la variable d'enquête correspondante.) La variance de $\bar{x}_{\text{st}}$ est donnée par

$$V(\bar{x}_{\text{st}}) = \sum_{h=1}^{L-1} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h},$$

$$\bar{x}_{\text{st}} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{x}_h, \bar{x}_h = \frac{1}{n_h} \sum_{k=1}^{n_h} x_{kh} (h = 1, \ldots, L), \qquad (1)$$

où n_h est la taille de l'échantillon provenant de la h^{e} strate, N_h est la taille de la h^{e} strate, S_h^2 est la variance de population de X restreinte à la h^{e} strate, $\bar{x}_{\text{st}}$ est l'estimateur de la moyenne de population de X sous échantillonnage *STSI*, $\bar{x}_h$ est l'estimateur de la moyenne de population de X dans la h^{e} strate sous échantillonnage aléatoire simple sans remise (*SI*) et x_{kh} est la valeur de X pour la k^{e} unité d'échantillonnage de la h^{e} strate et $h = 1, \ldots, L.$.

La répartition optimale de l'échantillon, qui s'obtient, dans le cas de notre problème, par minimisation de la variance (1) sachant la taille d'échantillon n, est donnée par la formule de l'optimum de Neyman adaptée à une approche avec strate à tirage complet (Lednicki et Wieczorkowski 2003) :

$$n_h = (n - N_L) \frac{N_h S_h}{\sum_{h=1}^{L-1} N_h S_h}, h = 1, \ldots, L-1. \qquad (2)$$

L'approche géométrique de stratification a pour objectif de rendre égales les valeurs du coefficient de variation de X dans les L strates. Elle consiste simplement à appliquer la formule qui suit basée sur une progression géométrique (Gunning et Horgan 2004)

$$k_h = ar^h, h = 0, \ldots, L, \qquad (3)$$

où $a = \min(X), k_L = \max(X)$ et $r = (k_L / k_0)^{1/L}$. La formule (3) repose sur l'hypothèse selon laquelle X suit une loi uniforme dans chaque strate.

L'approche par optimisation appliquée à ce problème de stratification particulier s'inspire de l'optimisation numérique du problème suivant : minimiser

$$f(\mathbf{k}) = V(\bar{x}_{\text{st}}), \qquad (4)$$

où $V(\bar{x}_{\text{st}})$ est la variance (1) sous la répartition optimale (2), sous les contraintes

$$N_h \geq 2 \text{ et } 2 \leq n_h \leq N_h \text{ pour } h = 1, \ldots, L-1, \qquad (5)$$

et

$$\sum_{h=1}^{L-1} n_h = n - N_L. \qquad (6)$$

Parfois, si l'on veut que le niveau de précision soit plus ou moins le même dans chaque strate, il est possible d'appliquer une méthode de « répartition avec puissance » (en anglais, *power allocation*) (Bankier 1988; Rivest 2002; Lednicki et Wieczorkowski 2003):

$$n_h = \frac{(n - N_L)(N_h \bar{x}_h)^P}{\sum_{h=1}^{L-1} (N_h \bar{x}_h)^p}, p \in (0,1]; h = 1, \ldots, L-1. \qquad (7)$$

L'approche par optimisation est plus difficile à appliquer que l'approche géométrique, en grande partie parce que cette dernière requiert un algorithme considérablement plus simple. Un choix doit être fait parmi les diverses méthodes d'optimisation disponibles. Lednicki et Wieczorkowski (2003) ont utilisé la méthode du simplexe de Nelder et Mead (1965); cependant, il est également possible d'appliquer des méthodes plus efficaces, qui nécessitent souvent l'auto-application d'algorithmes (par exemple, Kozak 2004).

Il convient de souligner que la stratification géométrique ne tient compte ni de la formule de la variance (1), ni de la répartition de l'échantillon (2), ni des contraintes (5). Or, il peut arriver que l'une des contraintes (5) ne soit pas satisfaite. Par conséquent, la stratification géométrique est une méthode de stratification approximative.

Dans la présente étude, nous avons appliqué l'algorithme proposé par Kozak (2004) pour stratifier plusieurs populations. Il s'agit d'un algorithme de recherche aléatoire adapté au problème de la stratification. Cet algorithme est simple; à chaque étape, une limite de strate est sélectionnée aléatoirement et modifiée aléatoirement. Si le nouvel ensemble de

limites de strate est meilleur que le précédent, il remplace ce dernier. L'annexe décrit en détail l'algorithme basé sur l'article publié par Kozak (2004).

Le deuxième problème examiné dans le présent article est la construction de strates qui minimisent la taille de l'échantillon provenant d'une population sachant le niveau de précision voulu de l'estimation (précision qui est donnée par la variance d'un estimateur de la moyenne ou du total de population). L'algorithme de Lavallée-Hidiroglou (LH) (Lavallée et Hidiroglou 1988) peut être considéré comme une méthode d'optimisation particulière en vue de résoudre ce problème précis de stratification; par contre, il n'est pas applicable à d'autres problèmes, par exemple celui considéré plus haut. Pour des précisions sur l'algorithme, consulter l'article de Lavallée et Hidiroglou (1988). Outre l'algorithme LH, nous avons appliqué la méthode de stratification géométrique et de recherche aléatoire pour construire les strates.

Nous avons utilisé le langage et l'environnement R (R Development Core Team 2005) pour réaliser tous les calculs de la présente étude.

3. Comparaison numérique de l'efficacité des approches de stratification sous taille d'échantillon fixe

À la présente section, nous comparons deux approches de stratification, la stratification géométrique (geom) et l'approche par optimisation (optim), appliquées à un problème de recherche des limites de strate qui minimisent la variance de l'estimateur considéré sachant une taille fixe d'échantillon. Pour réaliser la comparaison, nous avons généré cinq populations artificielles de tailles différentes (allant de 2 000 à 10 000). Les statistiques sommaires de ces populations sont présentées au tableau 1; les histogrammes des variables de stratification dans les populations sont donnés à la figure 1. Dans chaque cas, la variable de stratification était positivement asymétrique (le coefficient d'asymétrie variait de 1,40 pour la première population à 5,02 pour la cinquième). Comme cela est généralement le cas dans les populations réelles, les valeurs des variables de stratification étaient des nombres entiers. La taille d'échantillon, n_i, pour la i^{e} population était $n_i = f N_i$, où $f = 0{,}15$ est une fraction d'échantillonnage hypothétique et N_i est la taille de la i^{e} population.

Pour commencer, nous avons stratifié chaque population par la méthode de stratification géométrique en 4, 5, 6 et 7 strates. Puis, nous avons appliqué l'approche par optimisation; dans cette dernière, nous avons utilisé comme paramètres initiaux les limites de strate déterminées par la méthode de stratification géométrique.

Tableau 1
Statistiques sommaires pour les populations artificielles étudiées

Population	Taille	Étendue	Asymétrie	Moyenne	Variance
1	4 000	3–72	1,40	16,11	45,8
2	4 000	243–28 578	2,66	2 823,95	$4{,}8 \times 10^6$
3	2 000	6–2 793	3,55	224,12	$6{,}0 \times 10^4$
4	10 000	62–74 398	4,20	3 616,41	$2{,}1 \times 10^7$
5	2 000	259–186 685	5,02	9 265,36	$1{,}1 \times 10^8$

Comme Gunning et Horgan (2004), pour comparer l'efficacité des deux approches, nous avons calculé l'efficacité relative en appliquant la formule :

$$\text{eff}_{\text{geom, optim}} = \frac{V_{\text{geom}}(\bar{x}_{\text{st}})}{V_{\text{optim}}(\bar{x}_{\text{st}})}, \qquad (8)$$

où $V_{\text{geom}}(\bar{x}_{\text{st}})$ et $V_{\text{optim}}(\bar{x}_{\text{st}})$ sont les variances (1) sous les approches géométrique et par optimisation, respectivement. En outre, nous avons calculé les coefficients de variation de l'estimateur de la moyenne de population sous les deux approches :

$$\text{cv}_{\text{geom}} = \frac{\sqrt{V_{\text{geom}}(\bar{x}_{\text{st}})}}{\bar{x}_{\text{st}}}; \ \text{cv}_{\text{optim}} = \frac{\sqrt{V_{\text{optim}}(\bar{x}_{\text{st}})}}{\bar{x}_{\text{st}}}. \qquad (9)$$

Le tableau 2 contient les valeurs des efficacités relatives (8) et des coefficients de variation (9) pour chaque combinaison étudiée (population × nombre de strates).

Tableau 2
Coefficients de variation de l'estimateur de la moyenne de population sous les approches de stratification géométrique (CV_{geom}) et par optimisation (CV_{optim}), et efficacité de la stratification géométrique comparativement à l'approche par optimisation ($\text{eff}_{\text{geom,optim}}$)

Nombre de strates L	CV_{geom}	CV_{optim}	$\text{eff}_{\text{geom,optim}}$
Population 1			
4	0,0086	0,0056	1,53
5	0,0070	0,0042	1,66
6	0,0057	0,0034	1,66
7	0,0051	0,0029	1,75
Population 2			
4	0,0116	0,0084	1,37
5	0,0095	0,0065	1,47
6	0,0085	0,0051	1,66
7	0,0073	0,0042	1,72
Population 3			
4	0,0235	0,0133	1,76
5	0,0174	0,0100	1,74
6	0,0146	0,0081	1,80
7	0,0129	0,0067	1,91
Population 4			
4	0,0104	0,0063	1,64
5	0,0089	0,0047	1,88
6	0,0073	0,0038	1,93
7	0,0064	0,0032	2,00
Population 5			
4	0,0235	0,0134	1,76
5	0,0185	0,0100	1,86
6	0,0161	0,0080	2,00
7	0,0134	0,0074	1,82

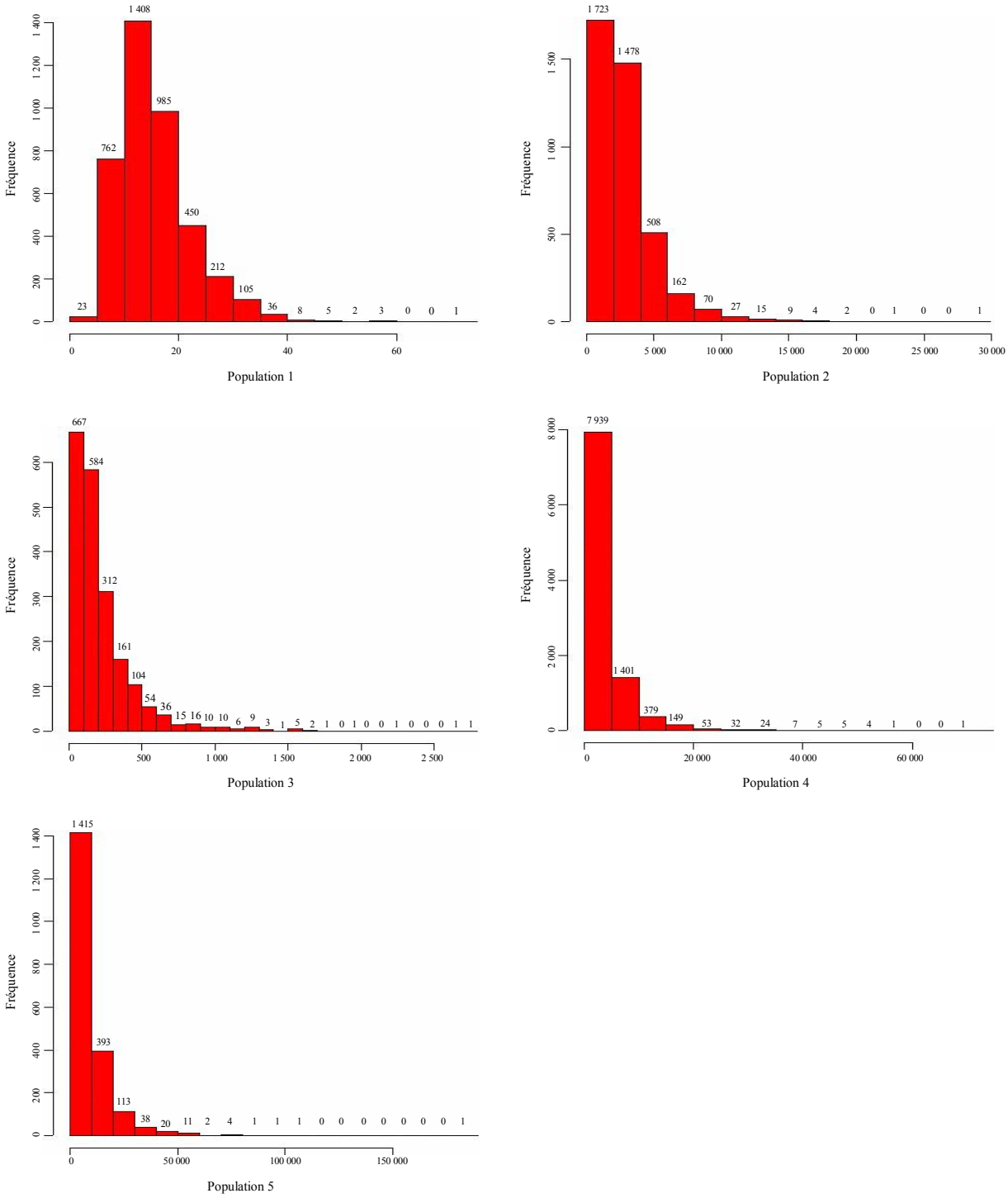

Figure 1. Histogrammes de la variable de stratification dans les populations artificielles étudiées.

Dans chaque cas, l'approche par optimisation a été plus efficace que la stratification géométrique. L'efficacité n'était inférieure à 1,5 que pour deux combinaisons; pour les autres, elle variait entre 1,5 et 2. En général, le gain d'efficacité est d'autant plus important que le nombre de strates construites est grand.

4. Comparaison numérique de l'efficacité des approches de stratification sous niveau de précision fixe de l'estimation

Gunning et Horgan (2004), ainsi que Horgan (2006) ont comparé la stratification géométrique à l'algorithme de Lavallée et Hidiroglou (Lavallée et Hidiroglou 1988) et constaté que la première était généralement plus efficace. À

la présente section, nous comparons les trois approches de stratification, à savoir la stratification géométrique, l'algorithme LH et l'approche par optimisation selon la méthode de recherche aléatoire. Nous avons utilisé pour la présente étude les cinq mêmes populations qu'à la section précédente (voir tableau 1 et figure 1).

Les efficacités relatives des approches comparées ont été évaluées au moyen de la formule

$$\text{eff}_{i,j} = \frac{n_i(\text{cv})}{n_j(\text{cv})}, \qquad (10)$$

où i et j sont les indices des approches de stratification (i, j = geom, optim, LH), et $n_i(\text{cv})$ et $n_j(\text{cv})$ sont les tailles d'échantillon minimales requises pour obtenir un niveau souhaité de précision (cv) sous les i^{e} et j^{e} approches, respectivement.

En suivant ces trois approches, nous avons stratifié chaque population en $L = 4, \ldots, 7$ strates; le niveau fixé de précision était de 0,01 dans chaque cas. Les tailles minimales d'échantillon requises pour ce niveau de précision et les efficacités relatives (10) sont données au tableau 3.

Tableau 3
Tailles d'échantillon minimales requises pour obtenir une valeur égale à 0,01 pour le coefficient de variation de l'estimateur de la moyenne de population, sous la stratification géométrique (n_{geom}), l'approche par optimisation (n_{optim}) et l'algorithme LH (n_{LH}); et efficacité de la stratification géométrique relativement à l'approche par optimisation ($\text{eff}_{\text{geom, optim}}$), de la stratification géométrique relativement à l'algorithme LH ($\text{eff}_{\text{geom, LH}}$) et de l'algorithme LH relativement à l'approche par optimisation ($\text{eff}_{\text{LH, optim}}$)

Nombre de strates L	n_{geom}	n_{optim}	n_{LH}	$\text{eff}_{\text{geom, optim}}$	$\text{eff}_{\text{geom, LH}}$	$\text{eff}_{\text{LH, optim}}$
Population 1						
4	805	496	496	1,63	1,63	1,00
5	613	344	344	1,78	1,78	1,00
6	460	252	252	1,83	1,83	1,00
7	357	192	192	1,86	1,86	1,00
Population 2						
4	483	248	259	1,94	1,86	1,04
5	329	154	163	2,14	2,02	1,06
6	224	113	117	1,98	1,92	1,03
7	180	83	83	2,17	2,17	1,00
Population 3						
4	782	410	411	1,91	1,90	1,00
5	601	303	304	1,98	1,98	1,00
6	495	242	241	2,04	2,05	1,00
7	422	195	195	2,11	2,16	1,00
Population 4						
4	839	409	409	2,05	2,05	1,00
5	650	301	301	2,15	2,15	1,00
6	552	240	242	2,30	2,28	1,01
7	–[1]	200	200	–	–	1,00
Population 5						
4	1 768	894	894	1,98	1,98	1,00
5	1 274	628	628	2,03	2,03	1,00
6	949	459	459	2,07	2,07	1,00
7	758	355	355	2,13	2,13	1,00

[1] L'obtention des limites de strate a posé des problèmes numériques (les tailles d'échantillon provenant de certaines strates étaient supérieures aux tailles de ces strates).

Il découle des résultats que l'approche par optimisation est plus efficace que la stratification géométrique; cette observation a été faite pour chaque population et chaque nombre de strates. L'efficacité relative était systématiquement supérieure à 1,6. En outre, une conclusion intéressante se dégage de la comparaison de l'efficacité des stratifications géométrique et LH. Comme nous l'avons déjà mentionné, Gunning et Horgan (2004), ainsi que Horgan (2006) ont constaté que la stratification géométrique était plus efficace que l'algorithme LH. Par contre, dans notre étude, l'algorithme LH était systématiquement plus efficace que la stratification géométrique, constatation que nous avons également faite pour d'autres populations de taille et d'asymétrie différentes que nous avons générées (les résultats ne sont pas présentés ici). Néanmoins, nous n'affirmons pas que l'algorithme LH est systématiquement plus efficace que la stratification géométrique. Il peut arriver que cette dernière donne de meilleurs résultats, comme Gunning et Horgan (2004) et Horgan (2006) l'ont observé lors de leurs études.

De la comparaison de l'algorithme LH à l'approche par optimisation, il découle que les deux méthodes donnent des points de stratification qui produisent des tailles d'échantillon semblables. Dans certains cas, la stratification LH est un peu meilleure et dans d'autres, un peu moins bonne, que l'approche par optimisation. Néanmoins, ces différences ne nous permettent pas de déclarer que l'une de ces deux approches est plus efficace que l'autre. En fait, elles ont toutes deux le même objectif (dans ce problème de stratification particulier) et diffèrent simplement en ce qui concerne l'algorithme utilisé pour atteindre cet objectif. Brièvement, d'après nos résultats, nous concluons qu'en général, la stratification LH et l'approche par optimisation sont plus efficaces que la stratification géométrique.

5. Conclusion

La méthode de stratification fondée sur une progression géométrique proposée par Gunning et Horgan (2004) possède un avantage significatif; plus précisément, son algorithme est très simple à appliquer comparativement à la méthode de la fonction cumulative de la racine carrée des fréquences de Dalenius et Hodges (1959) et à d'autres méthodes de stratification. Toutefois, il s'agit d'une méthode approximative, si bien que les limites de strate qu'elle produit peuvent mener à des estimations de précision médiocre (ou nécessiter l'utilisation d'un échantillon de grande taille pour obtenir le niveau requis de précision). En outre, il est probable que les strates construites ne satisferont pas toutes aux contraintes (5); autrement dit, il se peut que certaines strates soient vides (de sorte qu'elles ne contiendront aucune unité de population) ou (et) que la taille

des échantillons provenant de certaines strates soit inférieure à deux ou supérieure à la taille de population de la strate.

Dans notre étude, l'approche par optimisation (au moyen des algorithmes LH et de recherche aléatoire) s'est avérée plus efficace que la stratification géométrique pour chaque population étudiée et chaque nombre de strates construites. Néanmoins, les limites de strate données par la stratification géométrique peuvent être considérées comme de bons paramètres initiaux pour l'approche par optimisation; par contre, elles ne devraient pas être regardées comme des limites de strate optimales ou efficaces. De surcroît, nos résultats montrent de façon concluante que la stratification géométrique est moins efficace que celle présentée par Lavallée et Hidiroglou (1988), résultat opposé à celui obtenu par Gunning et Horgan (2004) et par Horgan (2006). L'étude de ce problème doit se poursuivre sur des populations asymétriques réelles; les recherches portant sur des populations artificielles indiquent sans équivoque que l'algorithme LH et l'approche par optimisation sont plus efficaces que la stratification géométrique.

À première vue, on pourrait s'étonner du fait que le gain d'efficacité réalisé en appliquant les approches LH et par optimisation comparativement à la stratification géométrique s'accroît lorsque le nombre de strates augmente. Toutefois, l'explication est simple. Le but de la stratification géométrique est d'égaliser les coefficients de variation de la variable de stratification dans les strates. Par conséquent, il diffère de celui de la stratification consistant à optimiser l'efficacité de l'estimation ou à minimiser la taille d'échantillon. Qui plus est, il n'est pas certain que, sous la stratification optimale, la distribution de la variable de stratification/d'enquête soit uniforme dans les strates. Les deux ensembles de limites de strate (c'est-à-dire ceux fournis par les approches géométrique et par optimisation) ne sont pas nécessairement les mêmes; en fait, il est probable qu'ils diffèrent.

Notons que nous appliqué l'algorithme de recherche aléatoire dans l'approche de stratification par optimisation. Or, l'algorithme de Lavallée et Hidiroglou (1988) est également un représentant des approches par optimisation. Quand le but de la stratification est de minimiser la taille d'échantillon requise pour obtenir un niveau souhaité de précision, il est probable que les deux approches produisent des résultats semblables, comme cela a été le cas lors de notre étude. Néanmoins, l'algorithme de recherche aléatoire peut être appliqué à n'importe quel problème de stratification (c'est-à-dire à toute fonction d'optimisation et ses contraintes), contrairement à l'algorithme LH, qui n'est applicable qu'à la minimisation de la taille d'échantillon pour un niveau de précision donné. Il convient de souligner que l'algorithme de recherche aléatoire fournit, comme méthode d'optimisation globale, des résultats aléatoires.

Notre but n'était pas, toutefois, de promouvoir l'un ou l'autre de ces deux algorithmes en montrant qu'ils sont plus efficaces que la stratification géométrique. Qui plus est, nous avons appliqué la méthode du simplexe de Nelder et Mead (1965) pour stratifier les populations (résultats non présentés ici); les résultats obtenus par cette méthode étaient fort semblables à ceux produits par les algorithmes LH et de recherche aléatoire. Chacune de ces méthodes présente certains inconvénients. Par exemple, des difficultés numériques peuvent survenir lors de l'utilisation de l'algorithme LH (Slanta et Krenzke 1996), tandis que la méthode de recherche aléatoire fournit des résultats aléatoires (Kozak 2004), la méthode de Nelder et Mead (1965) peut être inefficace si le nombre de strates et la taille de la population sont grands (Kozak 2004) et, en fait, l'obtention de points de stratification optimaux n'a été prouvée pour aucune de ces méthodes. Par conséquent, il reste encore à construire un algorithme de stratification produisant des résultats optimaux quelle que soit la situation (par exemple en ce qui concerne la taille de la population ou l'asymétrie de la variable), ainsi que des résultats non aléatoires. Notre objectif principal était de prouver que la stratification géométrique n'est pas optimale, mais que les points de stratification qu'elle produit peuvent être utiles comme paramètres initiaux dans d'autres approches de stratification.

Remerciements

Les auteurs remercient vivement les examinateurs et le rédacteur adjoint de *Techniques d'enquête* de leurs commentaires précieux, qui leur ont permis d'améliorer la première version du présent article.

Annexe

L'algorithme qui suit a été proposé par Kozak (2004) et nous nous sommes bornés à adapter certains de ses détails au problème général de la stratification. Dans l'algorithme, nous ne faisons pas référence au problème particulier de la stratification (autrement dit, nous ne définissons pas la fonction d'optimisation et ses contraintes), puisqu'il fonctionne pour les deux problèmes présentés dans l'article, ainsi que pour d'autres problèmes de stratification. Au besoin, nous faisons référence à la « fonction d'optimisation » (qui peut être la variance d'un estimateur étudié ou la taille d'un échantillon provenant d'une population) et aux « contraintes » (qui, selon la fonction d'optimisation, peuvent être les contraintes (5) et (6), ou les contraintes (5) combinées à la contrainte sur le niveau de précision de l'estimation); d'autres formes de la fonction d'optimisation et de ses contraintes peuvent sans aucun doute être prises en considération.

Définissons un vecteur comme il suit. Il prend des valeurs dans l'intervalle $(1, N)$, N étant la taille de population. À condition qu'une population soit triée en fonction des valeurs d'une variable de stratification X, deux éléments a_{h-1} et a_h du vecteur **a** définissent la strate h de telle façon que cette strate comprenne les éléments d'indice I (qui donne l'ordre d'un élément dans la population triée) tel que $a_{h-1} < I \le a_h, \quad h = 1, \ldots, L, a_0 = 0, a_L = N$. L'algorithme est le suivant.

1. Trier la population en fonction des valeurs de la variable de stratification.
2. Choisir un vecteur initial **a**, c'est-à-dire le vecteur de limites de strate initiales. Des nombres entiers aléatoires qui satisfont les contraintes peuvent être utilisés, mais la pratique révèle que de meilleurs résultats peuvent être obtenus en utilisant les limites de strate approximatives déterminées par une méthode de stratification approximative. Calculer la valeur de la fonction d'optimisation. Vérifier les contraintes; si elles ne sont pas satisfaites, les points initiaux doivent être modifiés.
3. Pour $r = 0, 1, \ldots, R$ répéter l'étape suivante :
   a. Générer le point $\mathbf{a}'$ en tirant une limite de strate a_i puis en la modifiant comme il suit

$$\begin{aligned} a_i' &= a_i + j, \\ a_k' &= a_k \quad \text{for } k = 1, \ldots, L-1, k \ne i, \end{aligned} \tag{11}$$

   où j est le nombre entier aléatoire, $j \in \langle -p; -1 \rangle \cup \langle 1; p \rangle$, p étant un nombre entier donné choisi d'après la taille de population (la valeur de p est d'autant plus élevée que la population est grande); habituellement, p devrait être compris entre 3 et 5.
   b. Calculer la valeur de la fonction d'optimisation.
   c. Si les contraintes sont satisfaites et que la valeur de la fonction d'optimisation sous le vecteur $\mathbf{a}'$ est plus petite que celle obtenue sous le vecteur **a**, accepter le nouveau vecteur, c'est-à-dire $\mathbf{a}_{r+1} = \mathbf{a}'$ (où $\mathbf{a}_{r+1}$ est le vecteur de limites de strate dans une itération suivante); sinon, ne pas accepter le vecteur, c'est-à-dire $\mathbf{a}_{r+1} = \mathbf{a}$.
4. Finir l'algorithme si la règle d'arrêt est satisfaite, c'est-à-dire si $r = R$, où R est le nombre donné d'étapes ou que, lors des m (par exemple, 50) dernières étapes, la valeur de la fonction d'optimisation n'a pas varié. Enfin, calculer le vecteur **k** (le vecteur de limites de strate finales) en fonction des valeurs du vecteur **a**.

Bibliographie

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

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

Ekman, G. (1959). An approximation useful in univariate stratification. *Annals of Mathematical Statistics*, 30, 219-229.

Glasser, G.J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute*, 30, 28-32.

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., Horgan, J.M. et Yancey, W. (2004). Geometric stratification of accounting data. *J. de Contaduria y Administration*, 214, septiembre-diciembre.

Hidiroglou, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.

Horgan, J.M. (2006). Stratification of skewed populations: A review. *Revue Internationale de Statistique*, 74(1): 67-76.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6(5), 797-806.

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

Lednicki, B., et Wieczorkowski, R. (2003). Optimal stratification and sample allocation between subpopulations and strata. *Statistics in Transition*, 6, 287-306.

Nelder, J.A., et Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7, 308-313.

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

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 (http://www.mat.ulaval.ca/pages/lpr/).

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.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 185-196
Statistique Canada, N° 12-001 au catalogue

Sondage indirect : Les fondements de la méthode généralisée du partage des poids

Jean-Claude Deville et Pierre Lavallée [1]

Résumé

Lorsqu'on veut sélectionner un échantillon, il arrive qu'au lieu de disposer d'une base de sondage contenant les unités de collecte souhaitées, on ait accès à une base de sondage contenant des unités liées d'une certaine façon à la liste d'unités de collecte. On peut alors envisager de sélectionner un échantillon dans la base de sondage disponible afin de produire une estimation pour la population cible souhaitée en s'appuyant sur les liens qui existent entre les deux. On donne à cette approche le nom de *sondage indirect*.

L'estimation des caractéristiques de la population cible étudiée par sondage indirect peut poser un défi de taille, en particulier si les liens entre les unités des deux populations ne sont pas bijectifs. Le problème vient surtout de la difficulté à associer une probabilité de sélection, ou un poids d'estimation, aux unités étudiées de la population cible. La méthode généralisée du partage des poids (MGPP) a été mise au point par Lavallée (1995) et Lavallée (2002) afin de résoudre ce genre de problème d'estimation. La MGPP fournit un poids d'estimation pour chaque unité enquêtée de la population cible.

Le présent article débute par une description du sondage indirect, qui constitue le fondement de la MGPP. En deuxième lieu, nous donnons un aperçu de la MGPP dans lequel nous la formulons dans un cadre théorique en utilisant la notation matricielle. En troisième lieu, nous présentons certaines propriétés de la MGPP, comme l'absence de biais et la transitivité. En quatrième lieu, nous considérons le cas particulier où les liens entre les deux populations sont exprimés par des variables indicatrices. En cinquième lieu, nous étudions certains liens typiques spéciaux afin d'évaluer leur effet sur la MGPP. Enfin, nous examinons le problème de l'optimalité. Nous obtenons des poids optimaux dans un sens faible (pour des valeurs particulières de la variable d'intérêt), ainsi que les conditions dans lesquelles ces poids sont également optimaux au sens fort et indépendants de la variable d'intérêt.

Mots clés : Sondage indirect; méthode généralisée du partage des poids; absence de biais; poids optimaux.

1. Introduction

En vue de sélectionner les échantillons nécessaires pour les enquêtes sociales ou économiques, il est utile de disposer de bases de sondage, c'est-à-dire de listes d'unités, offrant un moyen d'atteindre les populations cibles souhaitées. Malheureusement, il arrive qu'au lieu de posséder une liste contenant les unités de collecte souhaitées, on dispose d'une liste d'unités reliée d'une certaine façon à celle des unités de collecte. On peut par conséquent parler de deux populations U^A et U^B liées l'une à l'autre, où l'on souhaite produire une estimation pour U^B. Malheureusement, on ne dispose d'une base de sondage que pour U^A. On peut alors envisager de sélectionner un échantillon s^A dans U^A afin de produire une estimation pour U^B en s'appuyant sur la correspondance qui existe entre les deux populations. On parle alors de *sondage indirect*.

L'estimation des caractéristiques d'une population cible U^B étudiée par sondage indirect peut poser un défi de taille, en particulier si les liens entre les unités des deux populations ne sont pas bijectifs. Le problème vient surtout de la difficulté à associer une probabilité de sélection, ou un poids d'estimation, aux unités de la population cible visées par le sondage. La méthode généralisée du partage des poids (MGPP) a été mise au point par Lavallée (1995) et Lavallée (2002), et également présentée dans Lavallée et Caron (2001), afin de résoudre ce genre de problème d'estimation. La MGPP fournit un poids d'estimation pour chaque unité étudiée de la population cible U^B. Fondamentalement, ce poids d'estimation correspond à une moyenne pondérée des poids de sondage des unités de l'échantillon s^A. La MGPP est une extension de la méthode de partage des poids décrite par Ernst (1989) dans le contexte des enquêtes longitudinales auprès des ménages.

Le but du présent article est de décrire le sondage indirect, c'est-à-dire les fondements de la MGPP, et d'obtenir, par la MGPP, des poids optimaux produisant des estimations sans biais dont la variance est minimale. Nous commencerons par décrire le sondage indirect, ainsi que la MGPP dans un cadre théorique qui fera appel, notamment, à la notation matricielle. L'utilisation de cette notation pour la MGPP a été présentée antérieurement par Deville (1998). Puis, nous utiliserons ce cadre théorique en vue d'énoncer certaines propriétés générales associées à la MGPP, dont l'absence de biais et la transitivité. Cette dernière consiste à passer de la population U^A à une population cible U^C par l'intermédiaire d'une population U^B. En troisième lieu, nous montrerons la correspondance entre la formulation

1. Jean-Claude Deville, Laboratoire de Statistique d'Enquête (ENSAI/CREST), Campus de Ker Lann, rue Blaise Pascal, 35170 Bruz, FRANCE. Courriel : deville@ensai.fr; Pierre Lavallée, Statistique Canada, Ottawa (Ontario), K1A 0T6, CANADA. Courriel : pierre.lavallee@statcan.ca.

matricielle et celle qui a été décrite dans Lavallée (1995), Lavallée (2002), ainsi que Lavallée et Caron (2001). En quatrième lieu, nous étudierons l'effet de diverses matrices de liens établissant la liaison entre U^A et U^B sur la précision des estimations obtenues par la MGPP. Enfin, nous examinerons le problème de l'optimalité. Nous obtiendrons des poids optimaux dans un sens faible (pour des valeurs particulières de la variable d'intérêt), ainsi que les conditions sous lesquelles ces poids sont également optimaux dans un sens fort et indépendants de la variable d'intérêt.

2. Sondage indirect

Comme nous l'avons mentionné dans l'introduction, le sondage indirect consiste à sélectionner un échantillon s^A dans une population U^A afin de produire une estimation pour une population cible U^B, en s'appuyant pour cela sur la correspondance qui existe entre les deux populations. Par exemple, supposons que nous voulions produire des estimations pour une population d'enfants (unités de collecte), mais que nous ne disposions d'une base de sondage que pour les parents. La population cible U^B est celle des enfants, mais nous devons sélectionner un échantillon de parents avant de pouvoir interviewer les enfants. Cette situation est illustrée à la figure 1.

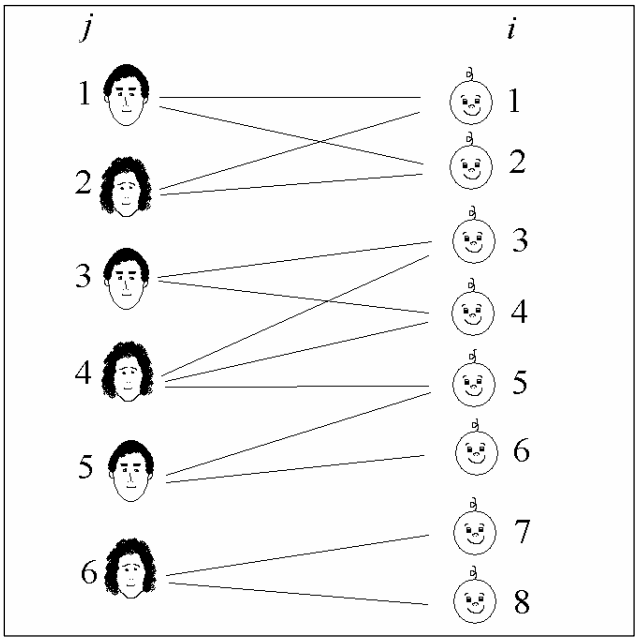

Figure 1. Population U^A de parents et population U^B d'enfants avec les liens entre les deux.

Soit U^A une population de N^A unités, où chaque unité est notée j. De même, soit U^B la population cible de N^B unités, où chaque unité est notée i. La correspondance entre les deux populations U^A et U^B peut être représentée par une *matrice de liens* $\mathbf{\Theta}_{AB} = [\theta_{ji}^{AB}]$ de taille $N^A \times N^B$, où chaque élément est $\theta_{ji}^{AB} \geq 0$. Autrement dit, l'unité j de U^A est reliée à l'unité i de U^B à condition que $\theta_{ji}^{AB} > 0$; sinon, il n'existe aucun lien entre les deux unités. Dans le cas de l'exemple susmentionné, la matrice de liens est donnée par

$$\mathbf{\Theta}_{AB} = \begin{bmatrix} \theta_{11}^{AB} & \theta_{12}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_{21}^{AB} & \theta_{22}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{33}^{AB} & \theta_{34}^{AB} & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{43}^{AB} & \theta_{44}^{AB} & \theta_{45}^{AB} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_{55}^{AB} & \theta_{56}^{AB} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \theta_{67}^{AB} & \theta_{68}^{AB} \end{bmatrix}.$$

En sondage indirect, l'obtention de la *matrice de liens* $\mathbf{\Theta}_{AB} = [\theta_{ji}^{AB}]$ est une question cruciale. Dans le cas où deux unités $j \in U^A$ et $i \in U^B$ ne sont pas liées, nous fixons simplement que $\theta_{ji}^{AB} = 0$. Lorsqu'il existe un lien entre deux unités j et i, le choix de la valeur de $\theta_{ji}^{AB} > 0$ est important. Comme nous le verrons, il influe sur la précision des estimations émanant du sondage indirect. Dans plusieurs applications, les valeurs de θ_{ji}^{AB} pour les unités liées sont simplement fixées à 1. Naturellement, elles pourraient être choisies différentes de 1. Lavallée et Caron (2001) discutent de l'utilisation des poids de couplage obtenus à partir d'un processus de couplage d'enregistrements entre U^A et U^B pour attribuer des valeurs aux éléments θ_{ji}^{AB}. Les poids de couplage sont proportionnels à la probabilité que deux unités $j \in U^A$ et $i \in U^B$ soient liées. Puisque le choix de $\theta_{ji}^{AB} > 0$ pour les deux unités liées j et i peut influencer la précision des estimations, il est naturel de rechercher les valeurs de θ_{ji}^{AB} qui minimiseront la variance des estimations. Ce problème d'optimisation est examiné à la section 6 de l'article.

Dans le sondage indirect, nous sélectionnons l'échantillon s^A de n^A unités à partir de U^A selon un certain plan d'échantillonnage. Soit π_j^A la probabilité de sélection de l'unité j. Nous supposons que $\pi_j^A > 0$ pour tout $j \in U^A$. Pour chaque unité j sélectionnée dans s^A, nous identifions les unités i de U^B pour lesquelles la correspondance n'est pas nulle, c'est-à-dire pour lesquelles $\theta_{ji}^{AB} > 0$. Soit Ω^B l'ensemble des n^B unités de U^B identifié par les unités $j \in s^A$, c'est-à-dire $\Omega^B = \{i \in U^B \mid \exists\, j \in s^A \text{ et } \theta_{ji}^{AB} > 0\}$. Pour chaque unité i de l'ensemble Ω^B, nous mesurons une variable d'intérêt y_i à partir de la population cible U^B. Soit $\mathbf{Y} = \{y_1, \ldots, y_{N^B}\}'$ le vecteur colonne de cette variable d'intérêt. D'un point de vue pratique, il est important de mentionner que, bien que la taille d'échantillon n^A soit habituellement déterminée d'avance, le nombre d'unités n^B est difficile à contrôler, car il dépend de l'échantillon sélectionné s^A et de la matrice de liens $\mathbf{\Theta}_{AB}$. Par conséquent, il s'avère difficile en général d'établir un budget

pour mesurer la variable d'intérêt y_i. Heureusement, dans la plupart des applications (par exemple, le cas parents-enfants susmentionné), le nombre de liens qui ont pour origine une unité donnée j de s^A est plus ou moins prévisible (par exemple, un parent a en général 1, 2 ou 3 enfants), ce qui facilite la détermination du nombre d'unités i de U^B qui, en dernière analyse, seront mesurées.

Nous supposons que pour toute unité j de s^A, il est possible d'obtenir les correspondances pour $i = 1, ..., N^B$. Autrement dit, nous pouvons identifier tous les liens entre les deux populations par interview directe ou grâce à une source administrative pour toute unité j échantillonnée. En outre, pour toute unité i identifiée de U^B, nous supposons qu'il est possible d'obtenir les liens pour $j = 1, ..., N^A$ (comme l'a mentionné Lavallée (2002), il existe des cas où cette dernière contrainte est difficile à satisfaire en pratique. Si nous revenons à l'exemple des parents et des enfants, il pourrait être difficile pour un très jeune enfant, sélectionné par l'entremise de sa mère, de mentionner son père, si les parents sont divorcés. Afin de simplifier la discussion, nous supposerons que ce genre de problème d'identification de liens est négligeable). Par conséquent, il n'est pas nécessaire de connaître les valeurs des liens entre les populations complètes U^A et U^B. En fait, nous ne devons connaître les liens (et, par conséquent, les valeurs de θ_{ji}^{AB}) que pour les lignes j de $\mathbf{\Theta}_{AB}$, où $j \in s^A$, ainsi que pour les colonnes i de $\mathbf{\Theta}_{AB}$, où $i \in \Omega^B$.

Supposons que nous voulions estimer le total Y^B de la population cible U^B, où $Y^B = \sum_{i=1}^{N^B} y_i$. Nous pouvons aussi écrire $Y^B = \mathbf{1}'_B \mathbf{Y}$, où $\mathbf{1}_B$ est le vecteur colonne de 1 de taille N^B (notons que, pour simplifier, nous utilisons la notation $\mathbf{1}_B$ au lieu de $\mathbf{1}_{N^B}$). Maintenant, posons que $\theta_{+i}^{AB} = \sum_{j=1}^{N^A} \theta_{ji}^{AB}$ et que $\tilde{\theta}_{ji}^{AB} = \theta_{ji}^{AB} / \theta_{+i}^{AB}$. Nous avons $\mathbf{1}'_A \mathbf{\Theta}_{AB} = \{\theta_{+1}^{AB}, ..., \theta_{+N^B}^{AB}\}$. Nous définissons alors la *matrice de liens normalisée* $\tilde{\mathbf{\Theta}}_{AB} = \mathbf{\Theta}_{AB} [\text{diag}(\mathbf{1}'_A \mathbf{\Theta}_{AB})]^{-1}$, où $\text{diag}(\mathbf{v})$ est la matrice carrée obtenue en plaçant les éléments du vecteur ligne (ou du vecteur colonne) $\mathbf{v}$ sur la diagonale et 0 ailleurs. Notons que, pour que la matrice $\tilde{\mathbf{\Theta}}_{AB}$ soit bien définie, il faut que $[\text{diag}(\mathbf{1}'_A \mathbf{\Theta}_{AB})]^{-1}$ existe, ce qui n'est le cas que si et autrement si $\theta_{+i}^{AB} > 0$ pour tout $i = 1, ..., N^B$. Dans l'exemple des parents et des enfants, cela signifie que chaque enfant doit être lié à au moins un parent.

Résultat 1 :

La matrice de liens $\tilde{\mathbf{\Theta}}_{AB}$ est une matrice de liens normalisée si et seulement si

$$\tilde{\mathbf{\Theta}}'_{AB} \mathbf{1}_A = \mathbf{1}_B. \tag{2.1}$$

La preuve du résultat 1 découle directement de la définition d'une matrice de liens normalisée. Partant du résultat 1, nous obtenons directement le résultat 2 que l'on trouve aussi dans Deville (1998) :

Résultat 2 :

$$\begin{aligned} Y^B &= \mathbf{1}'_B \mathbf{Y} \\ &= \mathbf{1}'_A \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y} = \sum_{j=1}^{N^A} \sum_{i=1}^{N^B} \frac{\theta_{ji}^{AB}}{\theta_{+i}^{AB}} y_i. \end{aligned} \tag{2.2}$$

Soit le vecteur colonne $\mathbf{Z} = \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y}$ de taille N^A. En considérant chaque ligne de $\mathbf{Z}$, nous définissons la variable $z_j = \sum_{i=1}^{N^B} \tilde{\theta}_{ji}^{AB} y_i$ pour chaque unité j de la population U^A et nous la mesurons pour chaque unité $j \in s^A$.

Pour estimer Y^B, nous voulons utiliser les valeurs de y_i mesurées à partir de l'ensemble Ω^B. Pour cela, nous utiliserons un estimateur de la forme :

$$\hat{Y}^B = \sum_{i=1}^{N^B} w_i y_i \tag{2.3}$$

où w_i est le poids d'estimation de l'unité i de Ω^B, avec $w_i = 0$ pour $i \notin \Omega^B$. Soit $\mathbf{W}' = \{w_1, ..., w_{N^B}\}$. L'estimateur (2.3) peut être réécrit sous la forme

$$\hat{Y}^B = \mathbf{W}'\mathbf{Y}. \tag{2.4}$$

Habituellement, pour obtenir une estimation sans biais de Y^B, il suffit d'utiliser comme poids l'inverse de la probabilité de sélection π_i^B de l'unité i. Comme le mentionne Lavallée (1995) et Lavallée (2002), dans le cas du sondage indirect, il peut être difficile, voire impossible, de calculer cette probabilité. Il propose alors de recourir à la MGPP, qui est définie comme il suit.

Soit $\boldsymbol{\pi}^A = \{\pi_1^A, ..., \pi_{N^A}^A\}'$ et soit $\mathbf{\Pi}_A = \text{diag}(\boldsymbol{\pi}^A)$ la matrice diagonale de taille $N^A \times N^A$ contenant les probabilités de sélection utilisées pour le tirage de l'échantillon s^A. Similairement, soit $\mathbf{t}^A = \{t_1^A, ..., t_{N^A}^A\}'$ où $t_j^A = 1$ si $j \in s^A$, et 0 autrement. Soit $\mathbf{T}_A = \text{diag}(\mathbf{t}^A)$ la matrice diagonale de taille $N^A \times N^A$ contenant les variables indicatrices t_j^A. En partant de $Y^B = \mathbf{1}'_A \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y} = \mathbf{1}'_A \mathbf{Z}$, nous pouvons former directement l'estimateur d'Horvitz-Thompson en fonction du vecteur $\mathbf{Z}$:

$$\hat{Y}^B = \mathbf{1}'_A \mathbf{T}_A \mathbf{\Pi}_A^{-1} \mathbf{Z}. \tag{2.5}$$

Puisque $\mathbf{Z} = \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y}$, nous avons $\hat{Y}^B = \mathbf{1}'_A \mathbf{T}_A \mathbf{\Pi}_A^{-1} \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y}$ et nous pouvons donc définir le vecteur colonne $\mathbf{W}$ de poids :

$$\mathbf{W} = \tilde{\mathbf{\Theta}}'_{AB} \mathbf{T}_A \mathbf{\Pi}_A^{-1} \mathbf{1}_A. \tag{2.6}$$

Le vecteur $\mathbf{W}$ est de taille N^B et, pour chaque $i = 1, ..., N^B$, nous avons $w_i = \sum_{j=1}^{N^A} t_j^A \tilde{\theta}_{ji}^{AB} / \pi_j^A$. Les poids w_i de ce vecteur sont obtenus par la MGPP, comme le décrit Lavallée (2002).

3. Propriétés de la MGPP

3.1 Absence de biais

Comme l'a mentionné Ernst (1989), pour obtenir un estimateur sans biais, il suffit que $E(\mathbf{W}) = \mathbf{1}_B$. Par construction, puisque l'estimateur (2.5) est un estimateur d'Horvitz-Thompson, cette condition est directement satisfaite et, par conséquent, la MGPP produit des estimations sans biais.

Partant de cette discussion, nous pouvons aussi obtenir le résultat suivant :

Résultat 3 :

Le vecteur de poids $\mathbf{W}$ donné par (2.6) fournit des estimations sans biais si et seulement si la matrice $\tilde{\mathbf{\Theta}}_{AB}$ est une matrice de liens normalisée.

Démonstration :

Partant de (2.6), nous avons

$$E(\mathbf{W}) = \tilde{\mathbf{\Theta}}'_{AB}\mathbf{1}_A. \qquad (3.1)$$

En utilisant le résultat 1, nous obtenons directement $E(\mathbf{W}) = \mathbf{1}_B$ et les estimations sont donc sans biais. Maintenant, supposons que $E(\mathbf{W}) = \mathbf{1}_B$. D'après (3.1), nous devons avoir $\tilde{\mathbf{\Theta}}'_{AB}\mathbf{1}_A = \mathbf{1}_B$ et, par conséquent, $\tilde{\mathbf{\Theta}}_{AB}$ est une matrice de liens normalisée.

3.2 Variance

Comme l'estimateur (2.5), est un estimateur d'Horvitz-Thompson, nous obtenons directement le résultat suivant :

Résultat 4 :

La variance de $\hat{Y}^B$ est donnée par

$$\begin{aligned}\mathrm{Var}(\hat{Y}^B) &= \mathbf{Z}'\mathbf{\Delta}_A\mathbf{Z} \\ &= \mathbf{Y}'\mathbf{\Delta}_B\mathbf{Y} \qquad (3.2)\end{aligned}$$

où $\mathbf{\Delta}_A = [(\pi^A_{jj'} - \pi^A_j\pi^A_{j'})/\pi^A_j\pi^A_{j'}]_{N^A\times N^A}$ est une matrice définie non négative de taille $N^A \times N^A$ et où $\pi^A_{jj'}$ est la probabilité de sélection conjointe des unités j et j' dans U^A, et où $\Delta_B = \tilde{\mathbf{\Theta}}'_{AB}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AB}$.

Pour une preuve de la variance de l'estimateur d'Horvitz-Thompson, voir Särndal, Swensson et Wretman (1992).

3.3 Transitivité

Supposons que nous voulions produire des estimations pour une population cible U^C que l'on ne peut atteindre que par l'entremise de la population U^B. Nous supposons que la population cible U^C contient N^C unités, chacune notée k. La correspondance entre les deux populations U^B et U^C peut être représentée par la matrice de liens $\mathbf{\Theta}_{BC} = [\theta^{BC}_{ik}]$ de taille $N^B \times N^C$, où chaque élément $\theta^{BC}_{ik} \geq 0$. Autrement dit, l'unité i de U^B est reliée à l'unité k de U^C à condition que $\theta^{BC}_{ik} > 0$, sinon; il n'existe aucun lien entre les deux unités.

Nous pouvons maintenant utiliser le sondage indirect par *transitivité*. Pour cela, nous sélectionnons un échantillon s^A à partir de la population U^A et commençons par identifier l'ensemble Ω^B de U^B. À partir de cet ensemble Ω^B, nous identifions alors les unités de U^C qui y sont associées, afin de former l'ensemble $\Omega^C = \{k \in U^C \mid \exists\, i \in \Omega^B \text{ et } \theta^{BC}_{ik} > 0\}$ d'unités devant être mesurées à partir de U^C. Une question importante est celle de savoir si, lorsqu'elle est appliquée dans le contexte du sondage indirect par transitivité, la MGPP est également transitive. Autrement dit, l'application de la MGPP de U^A à U^B, puis de U^B à U^C équivaut-elle à son application directe de U^A à U^C ?

Pour commencer, considérons le sondage indirect allant de U^A directement à la population cible U^C. Passer de la population U^A à U^B, puis à U^C revient à définir la matrice de liens $\mathbf{\Theta}_{AC} = [\theta^{AC}_{jk}]$ de taille $N^A \times N^C$ par $\mathbf{\Theta}_{AC} = \mathbf{\Theta}_{AB}\mathbf{\Theta}_{BC}$. Pour chaque unité j sélectionnée dans s^A, nous identifions les unités k de U^C pour lesquelles la correspondance n'est pas nulle, c'est-à-dire pour lesquelles $\theta^{AC}_{jk} > 0$, pour obtenir l'ensemble $\vec{\Omega}^C = \{k \in U^C \mid \exists\, j \in s^A \text{ et } \theta^{AC}_{jk} > 0\}$. Nous mesurons la variable d'intérêt y_k à partir de la population cible U^C. En appliquant la MGPP, nous obtenons, d'après (2.6), les poids suivants :

$$\vec{\mathbf{W}}_C = \tilde{\mathbf{\Theta}}'_{AC}\mathbf{T}_A\mathbf{\Pi}^{-1}_A\mathbf{1}_A \qquad (3.3)$$

où $\tilde{\mathbf{\Theta}}_{AC} = \mathbf{\Theta}_{AC}\,[\mathrm{diag}(\mathbf{1}'_A\mathbf{\Theta}_{AC})]^{-1}$.

Considérons maintenant l'utilisation du sondage indirect en deux étapes. Pour chaque unité j sélectionnée dans s^A, nous identifions les unités i de U^B pour lesquelles la correspondance n'est pas nulle, c'est-à-dire pour lesquelles $\theta^{AB}_{ji} > 0$. Comme auparavant, nous avons $\Omega^B = \{i \in U^B \mid \exists\, j \in s^A \text{ et } \theta^{AB}_{ji} > 0\}$. Pour chaque unité i de l'ensemble Ω^B, nous identifions alors les unités k de U^C pour lesquelles la correspondance n'est pas nulle, c'est-à-dire pour lesquelles $\theta^{BC}_{ik} > 0$. Nous avons alors l'ensemble $\Omega^C = \{k \in U^C \mid \exists\, i \in \Omega^B \text{ et } \theta^{BC}_{ik} > 0\}$. Partant de (2.6), nous obtenons le vecteur colonne $\mathbf{W}_B$ de poids associés aux unités de la population U^B :

$$\mathbf{W}_B = \tilde{\mathbf{\Theta}}'_{AB}\mathbf{T}_A\mathbf{\Pi}^{-1}_A\mathbf{1}_A. \qquad (3.4)$$

Pour chaque unité i de l'ensemble Ω^B, nous avons alors un poids non nul w^B_i. Or, l'ensemble Ω^B peut être considéré comme un échantillon d'unités qui sont utilisées dans un processus de sondage indirect pour identifier l'ensemble Ω^C. Par similarité avec l'échantillonnage indirect allant de l'échantillon s^A à la population cible U^B, l'application de la MGPP dans le contexte du sondage indirect allant de l'ensemble Ω^B à la population cible U^C produit les poids suivants :

$$\mathbf{W}_C = \tilde{\mathbf{\Theta}}'_{BC}\,\mathbf{T}_B\,\mathrm{diag}(\mathbf{W}_B)\mathbf{1}_B \qquad (3.5)$$

où $\tilde{\boldsymbol{\Theta}}_{BC} = \boldsymbol{\Theta}_{BC}[\,\mathrm{diag}(\mathbf{1}'_B \boldsymbol{\Theta}_{BC})]^{-1}$ et $\mathbf{T}_B = \mathrm{diag}(\mathbf{t}_B)$ avec $\mathbf{t}_B = (t_1^B, ..., t_{N^B}^B)'$ et $t_i^B = 1$ si $i \in \Omega^B$, et 0 autrement. Comme les poids $w_i^B = 0$ pour $i \notin \Omega^B$, nous avons $\mathbf{T}_B \,\mathrm{diag}(\mathbf{W}_B) = \mathrm{diag}(\mathbf{W}_B)$. Par conséquent, nous obtenons

$$\mathbf{W}_C = \tilde{\boldsymbol{\Theta}}'_{BC}\,\mathrm{diag}(\mathbf{W}_B)\mathbf{1}_B. \tag{3.6}$$

En remplaçant $\mathbf{W}_B$ par (3.4) dans l'équation (3.6), nous obtenons

$$\begin{aligned}\mathbf{W}_C &= \tilde{\boldsymbol{\Theta}}'_{BC}\,\mathrm{diag}(\tilde{\boldsymbol{\Theta}}'_{AB}\,\mathbf{T}_A\,\boldsymbol{\Pi}_A^{-1}\mathbf{1}_A)\mathbf{1}_B \\ &= \tilde{\boldsymbol{\Theta}}'_{BC}\,\tilde{\boldsymbol{\Theta}}'_{AB}\,\mathbf{T}_A\,\boldsymbol{\Pi}_A^{-1}\mathbf{1}_A. \end{aligned}\tag{3.7}$$

Puisque $\tilde{\boldsymbol{\Theta}}'_{BC}\tilde{\boldsymbol{\Theta}}'_{AB}\mathbf{1}_A = \tilde{\boldsymbol{\Theta}}'_{BC}\mathbf{1}_B = \mathbf{1}_C$, d'après le résultat 1, la matrice $\tilde{\boldsymbol{\Theta}}_{AB}\tilde{\boldsymbol{\Theta}}_{BC}$ est une matrice de liens normalisée. Par conséquent, la MGPP est transitive, du moins dans un certain sens. Autrement dit, les poids $\mathbf{W}_C$ peuvent être obtenus en une seule étape en utilisant la matrice de liens normalisée $\tilde{\boldsymbol{\Theta}}_{AB}\tilde{\boldsymbol{\Theta}}_{BC}$ dans la MGPP. Maintenant, pour que la MGPP soit parfaitement transitive, les poids $\mathbf{W}_C$ donnés par (3.7) devraient être exactement les mêmes que les poids $\vec{\mathbf{W}}_C$ donnés par (3.3). En comparant les équations (3.3) et (3.7), nous obtenons le résultat suivant :

Résultat 5 :

L'application de la MGPP de U^A à U^B, puis de U^B à U^C est transitive si et seulement si

$$\tilde{\boldsymbol{\Theta}}_{AC} = \tilde{\boldsymbol{\Theta}}_{AB}\tilde{\boldsymbol{\Theta}}_{BC}. \tag{3.8}$$

Malheureusement, la condition (3.8) n'est pas vérifiée en général. En fait, il est relativement facile de produire des exemples où $\tilde{\boldsymbol{\Theta}}_{AC} \neq \tilde{\boldsymbol{\Theta}}_{AB}\tilde{\boldsymbol{\Theta}}_{BC}$.

4. Une propriété structurelle de la MGPP

À la présente section, nous insistons sur le fait que, dans le cas du sondage indirect, le processus d'échantillonnage dépend uniquement des liens entre les deux populations U^A et U^B. Outre le fait d'être nulles ou non, les valeurs des θ_{ji}^{AB} proprement dites n'interfèrent pas avec le processus d'échantillonnage. Par ailleurs, les valeurs des θ_{ji}^{AB} **jouent** un rôle dans les poids (et donc l'estimateur) produits par la MGPP. Nous développons cette notion dans les paragraphes qui suivent.

L'échantillonnage indirect associe à chaque échantillon s^A dans U^A un échantillon Ω^B dans U^B, nommément $\Omega^B = \{i \in U^B \mid \exists\, j \in s^A \text{ et } \theta_{ji}^{AB} > 0\}$. Donc, une fonction $f : s^A \to \Omega^B$ qui établit la correspondance entre l'échantillon s^A et l'échantillon Ω^B est déterminée de façon unique par l'ensemble de couples (j, i) avec $\theta_{ji}^{AB} > 0$. Soit $l_{ji}^{AB} = 1$ si $\theta_{ji}^{AB} > 0$, et 0 autrement. Il s'agit des éléments de la matrice d'incidence du graphe reliant U^A à U^B.

Supposons qu'on nous donne une fonction ϕ partant de l'ensemble de sous-ensembles de U^A vers l'ensemble de sous-ensembles de U^B. Comme f, supposons que ϕ satisfait la « propriété d'union » : $\phi(s_1^A \cup s_2^A) = \phi(s_1^A) \cup \phi(s_2^A)$, où s_1^A et s_2^A sont deux sous-ensembles de U^A.

Résultat 6 :

La fonction ϕ est déterminée sans équivoque par une *matrice de liens zéro-un*.

Démonstration :

Nous pouvons le démontrer comme il suit : prenons $s_j^A = \{j\}$ pour une unité j dans U^A. Alors, $\phi(s_j^A)$ est un ensemble dans U^B. Soit $l_{ji}^{AB} = 1$ si l'unité i de U^B appartient à $\phi(s_j^A)$, et 0 autrement. En vertu de la propriété d'union, $\phi(s^A) = \bigcup_{j \in s^A} \phi(s_j^A)$ et l'ensemble de l_{ji}^{AB} définit la *matrice de liens zéro-un* $\mathbf{L}_{AB} = [l_{ji}^{AB}]$ de taille $N^A \times N^B$, qui définit précisément la fonction ϕ.

Cela nous donne une relation d'équivalence entre les matrices de liens, associées à une propriété plus profonde. Soit p^A un plan d'échantillonnage sur U^A (c'est-à-dire une loi de probabilité sur l'ensemble de sous-ensembles de U^A). La fonction f induit un plan d'échantillonnage sur U^B par $p^B(\Omega^B) = \sum_{s^A : \Omega^B = f(s^A)} p^A(s^A)$. Comme le plan est induit par f, il ne dépend pas de la matrice de liens particulière $\boldsymbol{\Theta}_{AB}$ définissant la fonction, mais est plutôt une caractéristique de la classe d'équivalence par la voie de la matrice de liens zéro-un $\mathbf{L}_{AB}$. Par conséquent, l'estimateur d'Horvitz-Thompson en U^B dépend uniquement de cette classe. Il y a donc un certain intérêt à choisir dans cette classe une matrice $\boldsymbol{\Theta}_{AB}$ ayant, dans un certain sens, une caractéristique optimale (voir la section 6).

5. Matrices de liens spéciales

Comme le montre les sections précédentes, la matrice de liens $\boldsymbol{\Theta}_{AB}$ dicte la forme de l'estimateur (2.4) donnée par la MGPP. À la présente section, nous décrivons certaines matrices de liens spéciales $\boldsymbol{\Theta}_{AB}$ qui correspondent à des cas extrêmes. Il est probable que tous ces cas ne seront pas observés en pratique, mais ils illustrent l'effet de la matrice de liens sur l'estimateur (2.4).

5.1 Matrice identité

Supposons que la matrice de liens $\boldsymbol{\Theta}_{AB}$ soit donnée par la matrice identité $\mathbf{I}$. En pratique, cela signifie que la relation entre la population U^A et la population cible U^B est bijective. Naturellement, cela implique que $N^A = N^B = N$ et que la matrice identité $\mathbf{I}$ est de taille $N \times N$.

Comme premier résultat, nous avons $\tilde{\boldsymbol{\Theta}}_{AB} = \mathbf{I}$. Par conséquent, le vecteur de poids (2.6) est donné par $\mathbf{W}' = (t_1^A / \pi_1^A, ..., t_{N^A}^A / \pi_{N^A}^A)$ et nous avons aussi $\mathbf{Z} = \tilde{\boldsymbol{\Theta}}_{AB}\mathbf{Y} = \mathbf{Y}$. Donc, l'estimateur $\hat{Y}^B$ donné par (2.5) n'est autre que l'estimateur d'Horvitz-Thompson $\hat{Y}^B = \mathbf{1}'_A \mathbf{T}_A \boldsymbol{\Pi}_A^{-1} \mathbf{Y}$.

5.2 Un pour tous (à l'intérieur des grappes)

Considérons le cas où la population U^B est divisée en Γ grappes γ, chacune de taille N_γ^B. Ces grappes sont telles que chaque grappe γ de U^B est associée à exactement une unité j de U^A. Par conséquent, nous pouvons utiliser la lettre γ pour les unités j de U^A ainsi que pour les grappes de U^B. Notons aussi que $\Gamma = N^A$.

Cette situation correspond à une matrice de liens $\boldsymbol{\Theta}_{AB}$ de forme diagonale par blocs où chaque sous-matrice ne contient qu'une seule ligne. Soit le vecteur ligne $\mathbf{1}'_{B\gamma}$ de taille N_γ^B et contenant uniquement des 1. La matrice de liens $\boldsymbol{\Theta}_{AB}$ est alors définie comme étant

$$\boldsymbol{\Theta}_{AB} = \begin{bmatrix} \mathbf{1}'_{B1} & & \mathbf{0} & \cdots & \mathbf{0} \\ & \ddots & & \ddots & \vdots \\ \mathbf{0} & & \mathbf{1}'_{B\gamma} & & \mathbf{0} \\ \vdots & \ddots & & \ddots & \\ \mathbf{0} & \cdots & \mathbf{0} & & \mathbf{1}'_{B\Gamma} \end{bmatrix}. \qquad (5.1)$$

Nous pouvons aussi écrire $\boldsymbol{\Theta}_{AB} = \text{diag}(\{\mathbf{1}'_{B1}, ..., \mathbf{1}'_{B\Gamma}\})$. En utilisant cette expression, nous avons $\text{diag}(\mathbf{1}'_A \boldsymbol{\Theta}_{AB}) = \text{diag}(\mathbf{1}'_A \text{diag}(\{\mathbf{1}'_{B1}, ..., \mathbf{1}'_{B\Gamma}\})) = \text{diag}(\{\mathbf{1}'_{B1}, ..., \mathbf{1}'_{B\Gamma}\})$ et donc $\tilde{\boldsymbol{\Theta}}_{AB} = \boldsymbol{\Theta}_{AB}$. À partir de l'équation (2.6), nous obtenons le vecteur colonne de poids $\mathbf{W}' = (t_1^A / \pi_1^A \mathbf{1}'_{B1}, ..., t_\Gamma^A / \pi_\Gamma^A \mathbf{1}'_{B\Gamma})$. Comme nous pouvons le voir, les éléments du vecteur colonne $\mathbf{W}$ ont les valeurs $t_\gamma^A / \pi_\gamma^A$ répétées dans chaque grappe γ de U^B. Partant de (2.4), nous obtenons

$$\hat{Y}^B = \sum_{\gamma=1}^{\Gamma} \frac{t_\gamma^A}{\pi_\gamma^A} Y_\gamma^B \qquad (5.2)$$

où $Y_\gamma^B = \sum_{i=1}^{N_\gamma^B} y_i$.

5.3 Tous pour un (à l'intérieur des grappes)

Considérons le cas où la population U^A est divisée en Γ grappes γ, chacune de taille N_γ^A. Ces grappes sont telles que chaque grappe γ de U^A est associée à exactement une unité i de U^B. Par conséquent, nous pouvons utiliser la lettre γ pour les grappes de U^A ainsi que les unités i de U^B. Notons aussi que $\Gamma = N^B$.

Cette situation correspond à une matrice de liens $\boldsymbol{\Theta}_{AB}$ de forme diagonale par blocs où chaque sous-matrice ne contient qu'une seule colonne. Soit le vecteur colonne $\mathbf{1}_{A\gamma}$ de taille N_γ^A et contenant uniquement des 1. La matrice de liens $\boldsymbol{\Theta}_{AB}$ est alors définie comme étant

$$\boldsymbol{\Theta}_{AB} = \begin{bmatrix} \mathbf{1}_{A1} & & \mathbf{0} & \cdots & \mathbf{0} \\ & \ddots & & \ddots & \vdots \\ \mathbf{0} & & \mathbf{1}_{A\gamma} & & \mathbf{0} \\ \vdots & \ddots & & \ddots & \\ \mathbf{0} & \cdots & \mathbf{0} & & \mathbf{1}_{A\Gamma} \end{bmatrix}. \qquad (5.3)$$

Nous pouvons aussi écrire $\boldsymbol{\Theta}_{AB} = \text{diag}(\{\mathbf{1}_{A1}, ..., \mathbf{1}_{A\Gamma}\})$. En utilisant cette expression, nous avons $\tilde{\boldsymbol{\Theta}}_{AB} = \text{diag}(\{1/N_1^A \mathbf{1}_{A1}, ..., 1/N_\Gamma^A \mathbf{1}_{A\Gamma}\})$. D'après (2.6), nous obtenons le vecteur colonne des poids $\mathbf{W}' = (1/N_1^A \sum_{j=1}^{N_1^A} t_j^A / \pi_j^A, ..., 1/N_\Gamma^A \sum_{j=1}^{N_\Gamma^A} t_j^A / \pi_j^A)$. Donc, les éléments γ (ou i) du vecteur colonne $\mathbf{W}$ ont les valeurs moyennes $\sum_{j=1}^{N_\gamma^A} t_j^A / \pi_j^A N_\gamma^A$, $\gamma = 1, ..., \Gamma$. Partant de (2.4), nous obtenons $\hat{Y}^B = \sum_{\gamma=1}^{\Gamma} y_\gamma / N_\gamma^A \sum_{j=1}^{N_\gamma^A} t_j^A / \pi_j^A$.

5.4 Échantillonnage inefficace

Supposons que certaines lignes de la matrice de liens $\boldsymbol{\Theta}_{AB}$ ne contiennent que des zéros. Cela signifie que certaines unités de la population U^A ne sont associées à aucune unité de la population U^B. Alors, si de telles unités sont sélectionnées dans l'échantillon s^A, elles ne permettront d'identifier aucune unité de U^B, ce qui peut être considéré comme inefficace du point de vue de l'échantillonnage. De façon plus formelle, supposons que chacune des N^{1A} premières lignes de la matrice de liens $\boldsymbol{\Theta}_{AB}$ contient au moins un $\theta_{ji} > 0$, et qu'elles forment la sous-matrice $\boldsymbol{\Theta}_1$. Supposons que les N^{0A} autres lignes de $\boldsymbol{\Theta}_{AB}$ ont $\theta_{ji} = 0$ pour $i = 1, ..., N^B$. Par conséquent, nous avons

$$\boldsymbol{\Theta}_{AB} = \begin{bmatrix} \boldsymbol{\Theta}_1 \\ \mathbf{0} \end{bmatrix}.$$

Comme premier résultat, nous obtenons

$$\tilde{\boldsymbol{\Theta}}_{AB} = \begin{bmatrix} \boldsymbol{\Theta}_1 [\text{diag}(\mathbf{1}'_{1A} \boldsymbol{\Theta}_1)]^{-1} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{\Theta}}_1 \\ \mathbf{0} \end{bmatrix} \qquad (5.4)$$

où $\mathbf{1}_{1A}$ est le vecteur colonne de 1 de taille N^{1A}. Partant de l'équation (2.6), nous obtenons le vecteur colonne de poids $\mathbf{W} = [\tilde{\boldsymbol{\Theta}}'_1 \ \mathbf{0}'] \mathbf{T}_A \boldsymbol{\Pi}_A^{-1} \mathbf{1}_A$. Soit $\boldsymbol{\Pi}_{1A} = \text{diag}(\{\pi_1^A, ..., \pi_{N^{1A}}^A\})$ la matrice diagonale de taille $N^{1A} \times N^{1A}$ et, de plus, soit $\mathbf{T}_{1A} = \text{diag}(\{t_1^A, ..., t_{N^{1A}}^A\})$ la matrice diagonale de taille $N^{1A} \times N^{1A}$. Nous obtenons alors

$$\begin{aligned} \mathbf{W} &= [\tilde{\boldsymbol{\Theta}}'_1 \ \mathbf{0}'] \mathbf{T}_A \boldsymbol{\Pi}_A^{-1} \mathbf{1}_A \\ &= \tilde{\boldsymbol{\Theta}}'_1 \mathbf{T}_{1A} \boldsymbol{\Pi}_{1A}^{-1} \mathbf{1}_{1A}. \end{aligned} \qquad (5.5)$$

Comme le montre (5.5), les poids dépendent uniquement des probabilités de sélection π_j^A des unités de U^A qui ont au moins un $\theta_{ji} > 0$ pour $i = 1, ..., N^B$. À partir de (2.4), nous obtenons finalement $\hat{Y}^B = \mathbf{1}'_{1A} \mathbf{T}_{1A} \boldsymbol{\Pi}_{1A}^{-1} \tilde{\boldsymbol{\Theta}}_1 \mathbf{Y}$.

5.5 Estimateur biaisé

Supposons que certaines colonnes de la matrice de liens $\boldsymbol{\Theta}_{AB}$ ne contiennent que des zéros. Cela signifie que certaines unités de la population U^B ne sont associées à aucune unité de la population cible U^A. Rappelons que, pour que la matrice $\tilde{\boldsymbol{\Theta}}_{AB}$ soit bien définie, il faut que $\text{diag}(\mathbf{1}'_A\boldsymbol{\Theta}_{AB})^{-1}$ existe. Comme nous le verrons, le cas qui nous occupe ne satisfait pas cette condition, ce qui mène à un estimateur biaisé du total Y^B.

De façon plus formelle, supposons que chacune des N^{1B} premières colonnes de la matrice de liens $\boldsymbol{\Theta}_{AB}$ contient au moins un $\theta_{ji} > 0$, et supposons qu'elles forment la sous-matrice $\boldsymbol{\Theta}_1$, différentes de celles de la section précédente. Supposons que les N^{0B} autres colonnes de $\boldsymbol{\Theta}_{AB}$ ont $\theta_{ji} = 0$ pour $j = 1, ..., N^A$. Nous avons par conséquent $\boldsymbol{\Theta}_{AB} = [\boldsymbol{\Theta}_1, \mathbf{0}]$.

De cette définition, il découle directement que

$$[\text{diag}(\mathbf{1}'_A\boldsymbol{\Theta}_{AB})]^{-1} = [\text{diag}([\mathbf{1}'_A\boldsymbol{\Theta}_1, \mathbf{1}'_A\mathbf{0}])]^{-1} = \begin{bmatrix} \text{diag}(\mathbf{1}'_A\boldsymbol{\Theta}_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}^{-1}. \quad (5.6)$$

Puisque cette matrice est singulière, $[\text{diag}(\mathbf{1}'_A\boldsymbol{\Theta}_{AB})]^{-1}$ n'existe pas. Il serait peut-être possible d'utiliser une *inverse généralisée* comme solution de ce problème. Rappelons que, pour une matrice carrée donnée $\mathbf{A}$, la matrice $\mathbf{A}^-$ est une inverse généralisée de $\mathbf{A}$ à condition que $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$ (Searle 1971). Une inverse généralisée possible de (5.6) est

$$[\text{diag}(\mathbf{1}'_A\boldsymbol{\Theta}_{AB})]^- = \begin{bmatrix} [\text{diag}(\mathbf{1}'_A\boldsymbol{\Theta}_1)]^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \quad (5.7)$$

Avec cette inverse généralisée, nous avons la matrice de liens normalisée suivante $\tilde{\boldsymbol{\Theta}}_- = \boldsymbol{\Theta}_{AB}[\text{diag}(\mathbf{1}'_A\boldsymbol{\Theta}_{AB})]^- = [\tilde{\boldsymbol{\Theta}}_1, \mathbf{0}]$. Partant de l'équation (2.6), nous pouvons obtenir le vecteur colonne $\mathbf{W}_-$ de poids :

$$\mathbf{W}_- = \begin{bmatrix} \tilde{\boldsymbol{\Theta}}'_1\mathbf{T}_A\boldsymbol{\Pi}_A^{-1}\mathbf{1}_A \\ \mathbf{0}' \end{bmatrix}. \quad (5.8)$$

Comme le montre l'expression (5.8), les poids sont nuls pour les unités i de la population cible U^B pour lesquels $\boldsymbol{\Theta}_{AB}$ contient $\theta_{ji} = 0$ pour $j = 1, ..., N^A$. Partant de (2.4) et en utilisant $\mathbf{W}_-$ au lieu de $\mathbf{W}$, nous obtenons $\hat{Y}^B_- = \mathbf{1}_A\mathbf{T}_A\boldsymbol{\Pi}_A^{-1}\tilde{\boldsymbol{\Theta}}_1\mathbf{Y}_1$ où $\mathbf{Y}_1 = \{y_1, ..., y_{N^{1B}}\}'$ est le sous-vecteur construit d'après les N^{1B} premiers éléments de $\mathbf{Y}$. Puisqu'en général, $E(\hat{Y}^B_-) = \mathbf{1}'_A\tilde{\boldsymbol{\Theta}}_1\mathbf{Y}_1 \neq \mathbf{1}'_A\mathbf{Y} = Y^B$, cet estimateur est biaisé pour le total Y^B.

6. Optimalité

L'optimalité est un aspect important de la MGPP. Comme nous l'avons montré au résultat 3, l'estimateur $\hat{Y}^B$ obtenu par cette méthode fournit des estimations sans biais à condition que la matrice $\tilde{\boldsymbol{\Theta}}_{AB}$ soit une matrice de liens normalisée. Étant donné que la variance (3.2) de cet estimateur dépend de cette matrice, il devrait exister au moins une matrice $\tilde{\boldsymbol{\Theta}}_{AB,\text{opt}}$ telle que la variance de l'estimateur $\hat{Y}^B$ soit minimale. Autrement dit, nous aimerions trouver les valeurs que les éléments θ_{ji}^{AB} plus grands que 0 devraient prendre pour obtenir l'estimateur de $\hat{Y}^B$ le plus précis.

Kalton et Brick (1995) ont été les premiers à examiner ce problème d'optimalité. Ils ont obtenu des résultats pour la situation simplifiée où $N^A = 2$ et où s^A est obtenu par échantillonnage avec probabilité égale. Ils ont conclu qu'il fallait utiliser $\theta_{ji}^{AB,\text{opt}} = 1$ lorsque $\theta_{ji}^{AB} > 0$ et $\theta_{ji}^{AB,\text{opt}} = 0$ lorsque $\theta_{ji}^{AB} = 0$. Lavallée (2002) et Lavallée et Caron (2001) ont obtenu des résultats du même genre par des simulations. Dans cette section, nous présentons de nouveaux résultats sur l'optimalité de la MGPP.

6.1 Factorisation

La factorisation est le problème inverse de la transitivité. Elle consiste à trouver une population U^G et des matrices de liens normalisées $\tilde{\boldsymbol{\Theta}}_{AG}$ et $\tilde{\boldsymbol{\Theta}}_{GB}$ telles que $\tilde{\boldsymbol{\Theta}}_{AB} = \tilde{\boldsymbol{\Theta}}_{AG}\tilde{\boldsymbol{\Theta}}_{GB}$. Cet exercice simplifie considérablement la recherche d'une matrice de liens normalisée optimale $\tilde{\boldsymbol{\Theta}}_{AB,\text{opt}}$.

Considérons que la population U^G est formée de grappes et que la factorisation est réalisée dans les contextes « un pour tous (à l'intérieur des grappes) » (de U^A à U^G) et « tous pour un (à l'intérieur des grappes) » (de U^G à U^B) présentés aux sections 5.2 et 5.3. Nous pouvons décrire cette situation de façon très générale comme il suit. Soit une population U^G contenant autant d'unités qu'il y a de liens partant des unités j de U^A. La taille de la population N^G est alors donnée par le nombre d'éléments θ_{ji}^{AB} de $\boldsymbol{\Theta}_{AB}$ dont la valeur est supérieure à 0. Chaque unité g de U^G peut être conceptualisée comme étant l'extrémité d'une « flèche » partant d'une unité j de U^A. Partant de ce graphe, il n'existe qu'une seule matrice de liens $\boldsymbol{\Theta}_{AG}$ de taille $N^A \times N^G$ assurant l'absence de biais, à savoir $\boldsymbol{\Theta}_{AG} = [\theta_{jg}^{AG}]$, où $\theta_{jg}^{AG} = 1$ s'il existe un lien (ou une « flèche ») partant de l'unité j de U^A vers l'unité g de U^G, et $\theta_{jg}^{AG} = 0$ autrement. Notons que, par construction, chaque unité g de U^G est liée, au plus, à une unité j de U^A et, donc, que $\tilde{\boldsymbol{\Theta}}_{AG} = \boldsymbol{\Theta}_{AG}$. Cela correspond à la situation « un pour tous dans les grappes » présentée à la section 5.2. Le sondage indirect de U^A à U^G est en fait un sondage en grappes type et fait aboutir la MGPP à l'estimateur d'Horvitz-Thompson habituel (voir Lavallée 2002). Dans le cas de l'exemple des parents et des enfants, le résultat de cette factorisation serait donné par la figure 2.

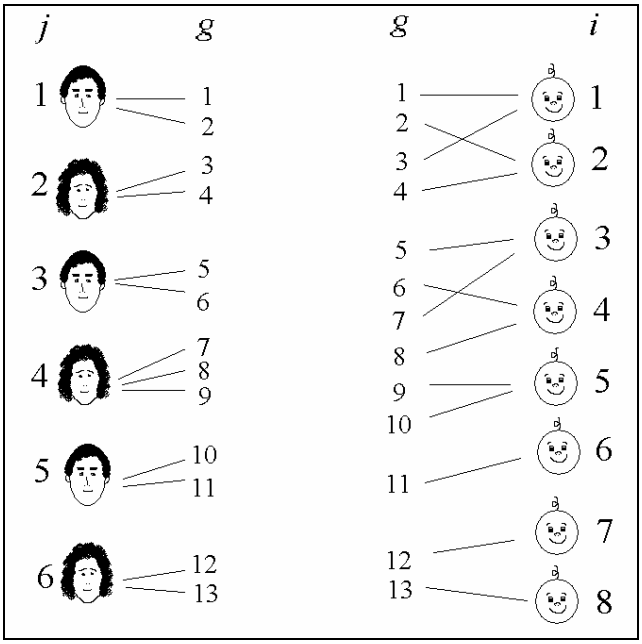

Figure 2. Résultat de la factorisation des populations parents-enfants.

Si nous considérons le graphe allant de U^G à U^B, nous pouvons construire la matrice de liens $\mathbf{\Theta}_{GB}$ de taille $N^G \times N^B$ comme suit. Étant donné la définition de la population U^G, chaque unité g de U^G est liée à exactement une unité i de U^B. Notons que le sondage indirect dans ce contexte peut être considéré comme un échantillonnage de grappes (c'est-à-dire les unités i de U^B) à partir de leurs éléments (c'est-à-dire les unités g de U^G). Il peut également être considéré comme étant le cas « tous pour un à l'intérieur des grappes » présenté à la section 5.3. Soit $\tilde{\mathbf{\Theta}}_{GB} = \mathbf{\Theta}_{GB}[\text{diag}(\mathbf{1}'_G\mathbf{\Theta}_{GB})]^{-1}$ la matrice de liens normalisée obtenue à partir de $\mathbf{\Theta}_{GB}$. Nous avons $\text{diag}(\mathbf{1}'_G\mathbf{\Theta}_{GB}) = \text{diag}(\mathbf{1}'_A\mathbf{\Theta}_{AB})$, et, par conséquent, $\tilde{\mathbf{\Theta}}_{GB} = \mathbf{\Theta}_{GB}[\text{diag}(\mathbf{1}'_A\mathbf{\Theta}_{AB})]^{-1}$.

Or,

$$\begin{aligned}\tilde{\mathbf{\Theta}}_{AG}\tilde{\mathbf{\Theta}}_{GB} &= \mathbf{\Theta}_{AG}\tilde{\mathbf{\Theta}}_{GB} \\ &= \mathbf{\Theta}_{AG}\mathbf{\Theta}_{GB}[\text{diag}(\mathbf{1}'_A\mathbf{\Theta}_{AB})]^{-1} \\ &= \mathbf{\Theta}_{AB}[\text{diag}(\mathbf{1}'_A\mathbf{\Theta}_{AB})]^{-1} \\ &= \tilde{\mathbf{\Theta}}_{AB}. \end{aligned} \tag{6.1}$$

Donc, en utilisant cette construction, la matrice de liens normalisée $\tilde{\mathbf{\Theta}}_{AB}$ reliant U^A à U^B peut toujours être factorisée en deux matrices $\tilde{\mathbf{\Theta}}_{AG}$ et $\tilde{\mathbf{\Theta}}_{GB}$.

6.2 Optimalité forte : énoncé du problème

Comme nous l'avons mentionné plus haut, le problème d'optimalité examiné ici consiste à minimiser la variance (3.2) par rapport à la matrice de liens normalisée $\tilde{\mathbf{\Theta}}_{AB}$. Par la factorisation présentée à la section 6.1, nous obtenons

$$\begin{aligned}\text{Var}(\hat{Y}^B) &= \mathbf{Y}'\tilde{\mathbf{\Theta}}'_{AB}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AB}\mathbf{Y} \\ &= \mathbf{Y}'\tilde{\mathbf{\Theta}}'_{GB}\tilde{\mathbf{\Theta}}'_{AG}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AG}\tilde{\mathbf{\Theta}}_{GB}\mathbf{Y} \\ &= \mathbf{Y}'\tilde{\mathbf{\Theta}}'_{GB}\mathbf{\Delta}_G\tilde{\mathbf{\Theta}}_{GB}\mathbf{Y} \end{aligned} \tag{6.2}$$

où $\mathbf{\Delta}_G = \tilde{\mathbf{\Theta}}'_{AG}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AG}$.

Pour toute matrice de liens normalisée $\tilde{\mathbf{\Theta}}_{AB}$, la factorisation présentée à la section 6.1 produit systématiquement le même premier facteur $\tilde{\mathbf{\Theta}}_{AG}$. Par conséquent, si nous recherchons une matrice optimale $\tilde{\mathbf{\Theta}}_{AB,\text{opt}}$ qui minimise la variance (3.2), il suffit d'optimiser le deuxième facteur $\tilde{\mathbf{\Theta}}_{GB}$. Nous aimerions aussi que la matrice optimale $\tilde{\mathbf{\Theta}}_{AB,\text{opt}}$ produise des estimations sans biais.

Soit U_i^G la sous-population de U^G contenant les N_i^G liens vers l'unité i de U^B. Notons que les sous-populations U_i^G sont disjointes. Donc, sans perte de généralité, nous pouvons classer les liens allant de U^A à U^B de façon que, pour tout i, les liens vers l'unité i dans U^B soient indicés consécutivement. Soit, ensuite, $\tilde{\mathbf{\theta}}_{GB,\,i}$ le i^{e} vecteur colonne de la matrice $\tilde{\mathbf{\Theta}}_{GB}$, $i = 1, \ldots, N^B$. Par construction, le vecteur $\tilde{\mathbf{\theta}}_{GB,\,i}$ ne contient que des éléments non nuls pour les N_i^G liens vers l'unité i de U^B. Donc, si nous représentons par $\tilde{\mathbf{\theta}}_{GB,\,i}$ un vecteur colonne de taille N_i^G contenant les éléments non nuls de $\tilde{\mathbf{\theta}}_{GB,\,i}$, nous obtenons

$$\tilde{\mathbf{\theta}}_{GB,\,i} = \begin{bmatrix} \mathbf{0} \\ \tilde{\mathbf{\theta}}_{GB,\,i} \\ \mathbf{0} \end{bmatrix}.$$

De même, soit $\dot{\mathbf{1}}_{G,\,i}$ le vecteur colonne de taille N^G contenant des valeurs 1 pour N_i^G éléments, et des valeurs 0 ailleurs. Si nous représentons par $\mathbf{1}_{G,\,i}$ un vecteur colonne de taille N_i^G contenant les valeurs 1, nous obtenons

$$\dot{\mathbf{1}}_{G,\,i} = \begin{bmatrix} \mathbf{0} \\ \mathbf{1}_{G,\,i} \\ \mathbf{0} \end{bmatrix}.$$

Afin que l'application de la MGPP pour passer de U^G à U^B soit sans biais, il faut que nous ayons $\tilde{\mathbf{\theta}}'_{GB,\,i}\mathbf{1}_{G,\,i} = 1$ pour toute i, ou de façon équivalente, $\tilde{\mathbf{\theta}}'_{GB,\,i}\dot{\mathbf{1}}_{G,\,i} = 1$. Ensemble, toutes ces considérations mènent au problème d'optimisation suivant :

Trouver une matrice $\tilde{\mathbf{\Theta}}_{GB,\text{opt}} = \{\tilde{\mathbf{\theta}}_{GB,\text{opt},1}, \ldots, \tilde{\mathbf{\theta}}_{GB,\text{opt},N^B}\}$ satisfaisant $\tilde{\mathbf{\theta}}'_{GB,\text{opt},i}\dot{\mathbf{1}}_{G,\,i} = 1$ pour tout $i = 1, \ldots, N^B$, et minimisant la forme quadratique $\text{Var}(\hat{Y}^B) = \mathbf{Y}'\tilde{\mathbf{\Theta}}'_{GB}\mathbf{\Delta}_G\tilde{\mathbf{\Theta}}_{GB}\mathbf{Y}$.

Ce problème n'est rien d'autre que la minimisation d'une forme quadratique positive sous des contraintes linéaires. Il s'agit d'un problème assez typique et simple à résoudre. Il est bien connu qu'il existe toujours une solution et qu'elle est unique si l'expression (6.2) est définie positive, ou que le

sous-espace nul de $\tilde{\boldsymbol{\Theta}}_{GB}$ n'est pas inclus dans l'espace nul de $\boldsymbol{\Delta}_G$.

Le problème d'optimisation susmentionné peut être réécrit sous une forme différente. Soit $\boldsymbol{\Delta}_{G,ii'}$ la sous-matrice de $\boldsymbol{\Delta}_G$ correspondant aux éléments qui occupent les positions g et g' si g possède un lien avec l'unité i et que g' possède un lien avec l'unité i'. Ces matrices constituent une partition de $\boldsymbol{\Delta}_G$. Notons que les matrices $\boldsymbol{\Delta}_{G,ii}$ sont symétriques, définies positives et que $\boldsymbol{\Delta}'_{G,ii'} = \boldsymbol{\Delta}_{G,i'i}$. Sous ces notations, le problème d'optimisation peut s'écrire sous la forme :

Minimiser

$$\sum_{i=1}^{N^B}\sum_{i'=1}^{N^B} y_i y_{i'} \tilde{\boldsymbol{\theta}}'_{GB,i} \boldsymbol{\Delta}_{G,ii'} \tilde{\boldsymbol{\theta}}_{GB,i'} \tag{6.3}$$

sous les contraintes $\tilde{\boldsymbol{\theta}}'_{GB,i}\mathbf{1}_{G,i} = 1$ pour tout $i = 1, \ldots, N^B$.

La minimisation est réalisée pour les vecteurs $\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i}$ qui satisfont

$$y_i \sum_{i'=1}^{N^B} \boldsymbol{\Delta}_{G,ii'} \tilde{\boldsymbol{\theta}}_{GB,\text{opt},i'} y_{i'} = \lambda_i \mathbf{1}_{G,i} \tag{6.4}$$

pour tout $i = 1, \ldots, N^B$ et où les λ_i représentent les multiplicateurs de Lagrange entrant dans la minimisation sous contraintes de (6.3). Comme le montre (6.4), le choix optimal $\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i}$ (et par conséquent $\tilde{\boldsymbol{\Theta}}_{GB,\text{opt}}$) dépend en général explicitement du vecteur $\mathbf{Y}$, ce qui n'est pas utile en pratique. Observons que l'ensemble des λ_i dépend aussi de la variable $\mathbf{Y}$. Ce qui apparaîtra plus explicitement à la section 6.3. Cette raison est celle pour laquelle, au lieu d'une optimalité forte, nous rechercherons une forme plus faible donnant une solution « optimale » $\tilde{\boldsymbol{\Theta}}_{GB,\text{opt}}$ (et $\tilde{\boldsymbol{\Theta}}_{AB,\text{opt}}$) indépendante de $\mathbf{Y}$.

6.3 Optimalité faible

Les équations (6.4) doivent être valides pour tout vecteur $\mathbf{Y}$. En particulier, une condition nécessaire est qu'elles doivent être vérifiées pour une variable d'intérêt particulière, telle que $y_i = 1$ pour une unité i de U^B et $y_{i'} = 0$ pour toutes les autres unités i' de U^B ($i' \neq i$). Cela nous donne les conditions nécessaires (une pour chacune de ces variables particulières) $\boldsymbol{\Delta}_{G,ii}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i} = \lambda_i \mathbf{1}_{G,i}$. Si nous supposons que $\boldsymbol{\Delta}_{G,ii}$ est inversible, nous obtenons alors $\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i} = \lambda_i \boldsymbol{\Delta}^{-1}_{G,ii}\mathbf{1}_{G,i}$. Il peut être démontré qu'il s'agit aussi d'une condition suffisante. Maintenant, comme $\tilde{\boldsymbol{\theta}}'_{GB,\text{opt},i}\mathbf{1}_{G,i} = 1$, nous avons $\lambda_i = 1/\mathbf{1}'_{G,i}\boldsymbol{\Delta}^{-1}_{G,ii}\mathbf{1}_{G,i}$. Par conséquent, une condition nécessaire et suffisante pour que l'équation (6.4) soit satisfaite est que

$$\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i} = \frac{\boldsymbol{\Delta}^{-1}_{G,ii}\mathbf{1}_{G,i}}{\mathbf{1}'_{G,i}\boldsymbol{\Delta}^{-1}_{G,ii}\mathbf{1}_{G,i}}. \tag{6.5}$$

Ce résultat correspond à une optimisation faible au sens suivant. Le poids w_i donné par (2.6) satisfait $E(w_i) = 1$ et de surcroît $E(w_i \mid i \in \Omega^B) = 1/\pi_i^B$ où π_i^B est la probabilité d'inclusion de l'unité i dans Ω^B, qu'il est généralement difficile, voire impossible, de calculer en pratique. Notons que l'estimateur d'Horvitz-Thompson est caractérisé par $\text{Var}(w_i \mid i \in \Omega^B) = 0$. L'optimisation faible obtenue ici revient à minimiser $\text{Var}(w_i \mid i \in \Omega^B)$ sur toutes les matrices de liens normalisées possibles $\tilde{\boldsymbol{\Theta}}_{GB}$, ou, de façon équivalente $\tilde{\boldsymbol{\Theta}}_{AB}$. Cette variance est strictement positive dans les cas où l'unité i de U^B peut recevoir plus qu'un seul poids pour divers échantillons s^A. En outre, si nous utilisons (6.3), le multiplicateur λ_i semble être la variance du poids w_i et est, par conséquent, toujours strictement positif (sauf, cas que nous excluons, quand l'unité i est sélectionnée avec un poids égal à 1).

6.4 Forte optimalité indépendante de Y

L'optimalité faible est une condition nécessaire à l'optimalité forte indépendante du vecteur $\mathbf{Y}$ d'une variable d'intérêt. Elle donne la forme nécessaire des vecteurs $\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i}$ dans (6.4). Pour obtenir les conditions suffisantes pour une forte optimalité indépendante de $\mathbf{Y}$, nous retournons aux équations (6.4). Ces dernières doivent être satisfaites pour tous les vecteurs $\mathbf{Y}$ et doivent par conséquent être satisfaites pour une variable d'intérêt particulière, telle que $y_i = 1$ pour une unité i de U^B, $y_{i'} = 1$, pour une autre unité i' de U^B, et $y_{i''} = 0$ pour toutes les autres unités i'' de U^B ($i'' \neq i' \neq i$). Dans ce cas, pour que les équations (6.4) soient satisfaites, il est nécessaire d'avoir les relations suivantes pour tout i et i' :

$$\boldsymbol{\Delta}_{G,ii}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i} + \boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i'} = \lambda_i^{ii'}\mathbf{1}_{G,i} \tag{6.6}$$

$$\boldsymbol{\Delta}_{G,i'i'}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i'} + \boldsymbol{\Delta}_{G,i'i}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i} = \lambda_{i'}^{ii'}\mathbf{1}_{G,i'}.$$

Comme nous devons nécessairement avoir une optimalité faible, nous avons $\boldsymbol{\Delta}_{G,ii}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i} = \lambda_i \mathbf{1}_{G,i}$. Partant de la première ligne de (6.6), nous obtenons alors

$$\begin{aligned}\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i'} &= (\lambda_i^{ii'} - \lambda_i)\mathbf{1}_{G,i} \\ &= \Phi_{ii'}\mathbf{1}_{G,i}.\end{aligned} \tag{6.7}$$

En multipliant les deux membres de (6.7) par $\tilde{\boldsymbol{\theta}}'_{GB,\text{opt},i}$, nous obtenons

$$\begin{aligned}\tilde{\boldsymbol{\theta}}'_{GB,\text{opt},i}\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i'} &= \Phi_{ii'}\tilde{\boldsymbol{\theta}}'_{GB,\text{opt},i}\mathbf{1}_{G,i} \\ &= \Phi_{ii'}\end{aligned}$$

puisque $\tilde{\boldsymbol{\theta}}'_{GB,\text{opt},i}\mathbf{1}_{G,i} = 1$. Soit $\boldsymbol{\Phi}$ la matrice contenant les éléments $\Phi_{ii'}$ hors de la diagonale et $\Phi_{ii} = \lambda_i$ sur la diagonale. En utilisant de nouveau (6.2), nous pouvons

montrer que la variance optimale (quand elle existe) a pour expression $\mathbf{Y}'\mathbf{\Phi}\mathbf{Y}$.

Démontrons que cet ensemble de conditions est également suffisant. Supposons que (6.7) est vérifiée. Notons que, pour $i = i'$, la condition (6.7) n'est rien d'autre que (6.5) qui donne les valeurs nécessaires pour les $\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i}$. Il est maintenant simple de vérifier que (6.4) tient quelle que soit la valeur de $\mathbf{Y}$ et que nous avons obtenu l'optimalité forte. Les valeurs de λ_i dépendent de $\mathbf{Y}$, ainsi que de la variance $\mathrm{Var}(\hat{Y}^B)$, mais nous savons que les équations (6.4) ont toujours la même solution (6.5) qui ne dépend pas de $\mathbf{Y}$. Par conséquent, nous avons le résultat suivant :

Résultat 7 :

Les conditions $\mathbf{\Delta}_{G,\,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'} = \Phi_{ii'}\mathbf{1}_{G,\,i}$ sont nécessaires et suffisantes pour qu'il existe une matrice de liens normalisée $\tilde{\mathbf{\Theta}}_{GB,\,\mathrm{opt}}$, ou de façon équivalente, $\tilde{\mathbf{\Theta}}_{AB,\,\mathrm{opt}}$, qui permet d'obtenir une optimalité forte indépendante du vecteur $\mathbf{Y}$ de la variable d'intérêt. Les valeurs figurant dans les colonnes de cette matrice optimale forte sont données par (6.5) qui sont les vecteurs $\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i}$ obtenus à partir de l'optimalité faible.

Il convient de souligner que, puisque $\mathbf{\Delta}_{G,\,ii}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} = \lambda_i\mathbf{1}_{G,\,i}$, l'expression (6.7) peut s'écrire de façon équivalente sous la forme

$$\Phi^{**}_{ii'}\,\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} = \mathbf{\Delta}^{-1}_{G,\,ii}\mathbf{\Delta}_{G,\,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'} \qquad (6.8a)$$

ou

$$\Phi^{*}_{ii'}\,\mathbf{1}_{G,\,i} = \mathbf{\Delta}_{G,\,ii'}\mathbf{\Delta}^{-1}_{G,\,i'i'}\mathbf{1}_{G,\,i'} \qquad (6.8b)$$

où $\Phi^{**}_{ii'} = (\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}\mathbf{\Delta}_{G,\,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'})(\mathbf{1}'_{G,\,i}\mathbf{\Delta}^{-1}_{G,\,ii}\mathbf{1}_{G,\,i})$ et $\Phi^{*}_{ii'} = (\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}\mathbf{\Delta}_{G,\,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'})(\mathbf{1}'_{G,\,i'}\mathbf{\Delta}^{-1}_{G,\,i'i'}\mathbf{1}_{G,\,i'})$. Dans certaines situations, ces expressions peuvent s'avérer plus faciles à utiliser que l'expression (6.7) énoncée dans le résultat 7.

6.5 Deux exemples

Nous présentons maintenant deux exemples qui illustrent la théorie que nous venons d'exposer sur l'optimalité faible et l'optimalité forte indépendante de $\mathbf{Y}$.

Exemple 1 : Échantillonnage de Poisson

Supposons que l'échantillon s^A soit sélectionné par échantillonnage de Bernoulli ou de Poisson. Dans ce cas, la matrice $\mathbf{\Delta}_A$ de taille $N^A \times N^A$ est donnée par $\mathbf{\Delta}_A = \mathrm{diag}(1/\pi^A_j - 1)$. Si nous considérons la factorisation de la section 6.1, nous avons $\mathbf{\Delta}_G = \tilde{\mathbf{\Theta}}'_{AG}\,\mathbf{\Delta}_A\,\tilde{\mathbf{\Theta}}_{AG} = \tilde{\mathbf{\Theta}}'_{AG}\,[\mathrm{diag}(1/\pi^A_j - 1)]\,\tilde{\mathbf{\Theta}}_{AG} = [\mathrm{diag}((1/\pi^A_j - 1)\mathbf{1}_{A,\,jj})]$, où $\mathbf{1}_{A,\,jj}$ est une matrice carrée de taille N^A_j, avec N^A_j égal au nombre de liens (ou « flèches ») ayant pour origine l'unité j de U^A. De $\mathbf{\Delta}_G$, nous extrayons les sous-matrices $\mathbf{\Delta}_{G,\,ii}$ qui sont, ici, diagonales. Chaque sous-matrice $\mathbf{\Delta}_{G,\,ii}$ est donnée par $\Delta_{G,\,ii} = \mathrm{diag}(1/\pi^A_g - 1)$, qui est de taille N^G_i. Notons que chaque valeur $(1/\pi^A_g - 1)$ correspond simplement à une unité j de U^A qui a été liée antérieurement à l'unité g de U^G, qui à son tour a été liée à l'unité i de U^B. Maintenant, partant de (6.5), nous obtenons directement les valeurs optimales $\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i}$ qui minimisent $\mathrm{Var}(\hat{Y}^B)$, au sens faible. Ces valeurs sont données par les vecteurs

$$\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i} = \left\{ \frac{\pi^A_1}{(1-\pi^A_1)\,\tau^G_i},\ \ldots,\ \frac{\pi^A_{N^G_i}}{(1-\pi^A_{N^G_i})\,\tau^G_i} \right\}$$

où

$$\tau^G_i = \sum\nolimits_{g=1}^{N^G_i} \pi^A_g/(1-\pi^A_g),\ i = 1,\ \ldots,\ N^B.$$

Les $\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}$ sont utilisés pour construire les vecteurs $\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}$, puis la matrice $\tilde{\mathbf{\Theta}}_{GB,\,\mathrm{opt}} = \{\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,1}, \ldots, \tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,N^B}\}$. Enfin, après avoir calculé la matrice optimale, $\tilde{\mathbf{\Theta}}_{AB,\,\mathrm{opt}} = \mathbf{\Theta}_{AG}\tilde{\mathbf{\Theta}}_{GB,\,\mathrm{opt}}$, nous obtenons les poids optimaux $\mathbf{W}_{\mathrm{opt}}$ en utilisant (2.6).

Il convient de souligner que, si les probabilités d'inclusion π^A_j sont égales, nous obtenons

$$\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i} = \left\{ \frac{1}{N^G_i},\ \ldots,\ \frac{1}{N^G_i} \right\} = \frac{1}{N^G_i}\mathbf{1}_{GB,\,i},$$

où N^G_i est tout simplement le nombre d'unités de U^A liées à l'unité i de U^B. Autrement dit, dans le contexte de l'échantillonnage de Bernoulli (c'est-à-dire l'échantillonnage de Poisson avec probabilités égales), pour minimiser la variance $\mathrm{Var}(\hat{Y}^B)$, le choix des valeurs de $\theta^{AB}_{\mathrm{opt},\,ji}$ devrait être 1 s'il existe un lien entre l'unité j de U^A et l'unité i de U^B, et 0 autrement. Cela correspond aux résultats obtenus par Kalton et Brick (1995), Lavallée (2002), ainsi que Lavallée et Caron (2001).

En utilisant le résultat 7, nous vérifions maintenant si les conditions (6.7), (6.8a) ou (6.8b) sont satisfaites pour la matrice optimale $\tilde{\mathbf{\Theta}}_{AB,\,\mathrm{opt}}$ que nous avons obtenue par optimisation faible. Le cas échéant, cette matrice donne aussi une optimalité forte indépendante de la variable d'intérêt y_i. Premièrement, nous avons

$$\Delta^{-1}_{G,\,ii} = \mathrm{diag}\left(\frac{\pi^A_g}{1-\pi^A_g} \right).$$

En outre, chaque sous-matrice $\Delta_{G,\,ii'}$ de taille $N^G_i \times N^G_{i'}$ a plus ou moins une structure diagonale, mais « rembourrée » de zéro. Autrement dit, un élément typique de $\Delta_{G,\,ii'}$ est donné par $(1/\pi^A_g - 1)$ sur une partie de la diagonale si i et i' sont toutes deux liées à la même unité j de U^A

(c'est-à-dire liées à l'unité g de U^G provenant de la même unité j de U^A), et 0 autrement. Par conséquent, si deux unités i et i' ne sont pas liées aux mêmes unités de U^A, alors $\Delta_{G,ii'}$ est une matrice de zéros et les conditions (6.7), (6.8a) et (6.8b) sont automatiquement satisfaites. Si nous nous référons à la figure 1, les enfants $i=2$ et $i'=3$ de U^B ne sont pas apparentés aux mêmes parents j de U^A. Si la sélection des parents est faite par échantillonnage de Poisson ou de Bernoulli, la matrice $\mathbf{\Delta}_{G,23}$ de dimension 2×2 ne contiendra alors que des zéros, c'est-à-dire

$$\mathbf{\Delta}_{G,23}=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}.$$

Donc, les relations (6.7), (6.8a) ou (6.8b) seront satisfaites avec $\Phi_{23}=0$, ce qui exprime le fait que les poids de i et de i' ne sont pas corrélés.

Si deux unités i et i' sont reliées à la même unité j de U^A, alors, si nous utilisons (6.7), le vecteur colonne $\mathbf{\Delta}_{G,ii'}\tilde{\mathbf{\theta}}_{GB,\text{opt},i'}$ contient le scalaire $(\tau_{i'}^G)^{-1}=[\sum_{g=1}^{N_{i'}^G}\pi_g^A/(1-\pi_g^A)]^{-1}$ pour ses N_i^B premières composantes, et 0 pour les $N_{i'}^B-N_i^B$ composantes restantes (en supposant que $N_{i'}^B\geq N_i^B$). Comme la quantité $\mathbf{\Delta}_{G,ii'}\tilde{\mathbf{\theta}}_{GB,\text{opt},i'}$ doit être égale à $\Phi_{ii'}\mathbf{1}_{G,i}$ pour satisfaire (6.7), elle doit contenir uniquement la valeur $\Phi_{ii'}$. Puisque $\Phi_{ii'}=\tilde{\mathbf{\theta}}'_{GB,\text{opt},i}\mathbf{\Delta}_{G,ii'}\tilde{\mathbf{\theta}}_{GB,\text{opt},i'}$, cela se produira uniquement si le vecteur $\tilde{\mathbf{\theta}}_{GB,\text{opt},i}=[1]$, ce qui signifie qu'il n'existe qu'un seul lien vers l'unité i de U^B. Comme nous le voyons, il ne s'agit pas d'une condition qui sera satisfaite en général et, par conséquent, nous pouvons dire, dans le cas de l'échantillonnage de Poisson, il n'y aura généralement pas d'optimalité forte indépendante de $\mathbf{Y}$.

Pour conclure, nous pourrions dire que dans le cas de l'échantillonnage de Poisson ou de Bernoulli, les conditions (6.7), (6.8a) ou (6.8b) seront satisfaites en pratique uniquement si les unités de U^A sont liées à une seule unité de U^B, comme dans le cas de l'échantillonnage des ménages en utilisant une liste de personnes. Dans les autres cas, la matrice optimale $\tilde{\mathbf{\Theta}}_{AB,\text{opt}}$ obtenue par optimalité faible ne donnera vraisemblablement pas lieu à une optimisation forte indépendante de $\mathbf{Y}$.

Exemple 2 : Échantillonnage aléatoire simple

Supposons que l'on sélectionne l'échantillon s^A par échantillonnage aléatoire simple. Dans ce cas, la matrice $\mathbf{\Delta}_A$ de taille $N^A\times N^A$ est donnée par

$$\mathbf{\Delta}_A=\frac{N^A}{n^A}\frac{(N^A-n^A)}{(N^A-1)}\left[\mathbf{I}_A-\frac{\mathbf{1}_A\mathbf{1}'_A}{N^A}\right].$$

Si nous considérons la factorisation de la section 6.1, nous obtenons

$$\begin{aligned}\mathbf{\Delta}_G&=\tilde{\mathbf{\Theta}}'_{AG}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AG}\\&=\frac{N^A}{n^A}\frac{(N^A-n^A)}{(N^A-1)}\times\tilde{\mathbf{\Theta}}'_{AG}\left[\mathbf{I}_A-\frac{\mathbf{1}_A\mathbf{1}'_A}{N^A}\right]\tilde{\mathbf{\Theta}}_{AG}\\&=\frac{N^A}{n^A}\frac{(N^A-n^A)}{(N^A-1)}\times\left[\text{diag}(\mathbf{1}_{A,jj})-\frac{\mathbf{1}_G\mathbf{1}'_G}{N^A}\right]\end{aligned}\qquad(6.9)$$

où $\mathbf{1}_{A,jj}$ est une matrice carrée de taille N_j^A, avec N_j^A égal au nombre de liens (ou de « flèches ») ayant pour origine l'unité j de U^A. De $\mathbf{\Delta}_G$, nous extrayons les sous-matrices $\mathbf{\Delta}_{G,ii}$. Chaque sous-matrice $\mathbf{\Delta}_{G,ii}$ est donnée par

$$\mathbf{\Delta}_{G,ii}=\frac{N^A}{n^A}\frac{(N^A-n^A)}{(N^A-1)}\times\left[\mathbf{I}_{G,i}-\frac{\mathbf{1}_{G,i}\mathbf{1}'_{G,i}}{N^A}\right],$$

qui est de taille N_i^G. Alors, en utilisant un résultat matriciel que l'on peut trouver, entre autres, dans Jazwinski (1970), nous obtenons

$$\mathbf{\Delta}_{G,ii}^{-1}=\frac{(N^A-1)}{(N^A-n^A)}\frac{n^A}{N^A}\times\left[\mathbf{I}_{G,i}+\frac{1}{(N^A-N_i^G)}\mathbf{1}_{G,i}\mathbf{1}'_{G,i}\right].$$

Ensuite, partant de (6.5), nous obtenons directement les valeurs optimales

$$\tilde{\mathbf{\theta}}_{GB,\text{opt},i}=\frac{1}{N_i^G}\mathbf{1}_{G,i}$$

qui minimisent $\text{Var}(\hat{Y}^B)$, au sens faible, $i=1,\ldots,N^B$. Nous utilisons ces valeurs pour construire les vecteurs $\tilde{\mathbf{\theta}}'_{GB,\text{opt},i}$, puis la matrice $\tilde{\mathbf{\Theta}}_{GB,\text{opt}}=\{\tilde{\mathbf{\theta}}_{GB,\text{opt},1},\ldots,\tilde{\mathbf{\theta}}_{GB,\text{opt},N^B}\}$. Enfin, après avoir calculé la matrice optimale $\tilde{\mathbf{\Theta}}_{AB,\text{opt}}=\mathbf{\Theta}_{AG}\tilde{\mathbf{\Theta}}_{GB,\text{opt}}$, nous obtenons les poids optimaux $\mathbf{W}_{\text{opt}}$ en utilisant (2.6).

De nouveau, ce résultat est important, car il va directement dans le sens des résultats de Kalton et Brick (1995), de Lavallée (2002), et de Lavallée et Caron (2001). Autrement dit, dans le cas de l'échantillonnage aléatoire simple, le choix optimal de $\theta_{\text{opt},ji}^{AB}$ devrait être 1 s'il existe un lien entre l'unité j de U^A et l'unité i de U^B, et 0 sinon.

En utilisant le résultat 7, nous vérifions maintenant si les conditions (6.7), (6.8a) ou (6.8b) pour une optimalité forte indépendante de y_i sont satisfaites pour la matrice optimale $\tilde{\mathbf{\Theta}}_{AB,\text{opt}}$ que nous obtenons par optimisation faible. D'abord, chaque sous-matrice $\Delta_{G,ii'}$ de taille $N_i^G\times N_{i'}^G$ est donnée par

$$\mathbf{\Delta}_{G,ii'}=\frac{N^A}{n^A}\frac{(N^A-n^A)}{(N^A-1)}\times\left[\mathbf{H}_{G,ii'}-\frac{\mathbf{1}_{G,i}\mathbf{1}'_{G,i'}}{N^A}\right]$$

où $\mathbf{H}_{G,ii'}$ est une matrice diagonale de taille $N_i^G \times N_{i'}^G$ de valeur 1, « rembourrée » de zéros. En suivant exactement le même scénario que l'exemple 1, un élément type de $\mathbf{H}_{G,ii'}$ est donné par 1 si i et i' sont toutes deux liées à la même unité j de U^A (c'est-à-dire liées à l'unité g de U^G), et 0 sinon. Par conséquent, nous pouvons voir facilement dans quel cas les conditions (6.7), (6.8a) ou (6.8b) peuvent être satisfaites. En fait, comme toutes les composantes de $\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i}$ sont égales, $\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i'}$ est un vecteur proportionnel à la somme des lignes de $\boldsymbol{\Delta}_{G,ii'}$, c'est-à-dire la somme des lignes de

$$\left[\mathbf{H}_{G,ii'} - \frac{\mathbf{1}_{G,i}\mathbf{1}'_{G,i'}}{N^A}\right].$$

Mais (6.7) dit que ce vecteur doit avoir les mêmes composantes. Cela n'est possible que si et seulement si la matrice $\mathbf{H}_{G,ii'}$ ne contient que des zéros, ou qu'elle est de dimension 1×1, ce qui se produit lorsque i et i' sont chacune liées uniquement à un élément de U^A. Donc, comme pour l'échantillonnage de Poisson, une optimalité fort indépendante de $\mathbf{Y}$ n'a généralement pas lieu dans le cas de l'échantillonnage aléatoire simple.

7. Conclusion

Dans le présent article, nous avons discuté de l'utilisation du sondage indirect conjugué à la méthode généralisée du partage des poids (MGPP) élaborée pour produire des poids. Puis, nous avons démontré les propriétés suivantes de la MGPP : absence de biais, calcul de la variance et transitivité. Ensuite nous avons présenté une section sur l'utilisation de la MGPP lorsque les liens entre les populations U^A et U^B sont exprimés par des valeurs 1 et 0, c'est-à-dire qu'il existe un lien ou qu'il n'en existe pas. La section suivante a été consacrée aux résultats obtenus avec diverses formes de matrices de liens. Enfin, nous avons abordé le problème de l'optimalité, c'est-à-dire le choix des valeurs optimales pour exprimer les liens entre U^A et U^B de façon à minimiser la variance des estimations obtenues en appliquant la MGPP. Nous avons fait la distinction entre deux formes d'optimisation, à savoir l'optimisation faible et l'optimisation forte.

L'optimisation faible consiste à trouver les valeurs des liens qu'il convient d'utiliser pour minimiser, pour chaque unité, la variance des poids produits par la MGPP. La solution est toujours définie de façon unique, et est facile à calculer et à appliquer en pratique. L'optimisation faible est également une condition nécessaire de l'optimisation forte. L'optimisation forte consiste à trouver les valeurs des liens permettant de minimiser la variance de l'estimation du total de toute variable d'intérêt y. Elle n'existe pas pour tous les plans d'échantillonnage et type de liens entre les populations U^A et U^B. Elle dépend aussi de relations assez compliquées.

Nous recommandons d'utiliser l'optimisation faible, parce qu'elle coule de source et qu'elle est très facile à utiliser. En outre, si notre problème d'estimation peut être optimisé également au sens fort, nous aurons obtenu ce résultat par la voie de l'optimisation faible, même si nous ne l'avons pas démontré.

Remerciements

Les auteurs remercient toutes les personnes qui ont manifesté un intérêt pour le sondage indirect et particulièrement la MGPP. Elles ont motivé la rédaction de cet article qui dépasse le cadre de ce qui avait été écrit antérieurement à ce sujet.

Bibliographie

Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. Dans *Panel Surveys* (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York : John Wiley & Sons, Inc. 135-159.

Deville, J.C., (1998). Comment attraper une population en se servant d'une autre. *Insee méthodes, n°84-85-86, Actes des Journées de methodologie statistique des 17-18 mars 1998*, 63-82.

Jazminski, A.H. (1970). *Stochastic Processes and Filtering Theory*. New York : Academic Press.

Kalton, G., et Brick, J.M. (1995). Méthodes de pondération pour les enquêtes par panel auprès des ménages. *Techniques d'enquête*, 21, 1, 37-49.

Lavallée, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 1, 27-35.

Lavallée, P. (2002). *Le Sondage Indirect, ou la Méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles, Brussels.

Lavallée, P., et Caron, P. (2001). Estimation par la méthode généralisée du partage des poids : Le cas du couplage d'enregistrements. *Techniques d'enquête*, 27, 2, 171-187.

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

Searle, S.R. (1971). *Linear Models*. New York : John Wiley & Sons, Inc.

Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 197-206
Statistique Canada, N° 12-001 au catalogue

Extensions de la méthode d'échantillonnage indirect et son application aux enquêtes dans le tourisme

Jean-Claude Deville et Myriam Maumy-Bertrand [1]

Résumé

On doit procéder à une enquête portant sur la fréquentation touristique d'origine intra ou extra-régionale en Bretagne. Pour des raisons matérielles concrètes, les « enquêtes aux frontières » ne peuvent plus s'organiser. Le problème majeur est l'absence de base de sondage permettant d'atteindre directement les touristes. Pour contourner ce problème, on applique la *méthode d'échantillonnage indirect* dont la pondération est obtenue par la *méthode généralisée de partage des poids* développée récemment par Lavallée (1995), Lavallée (2002), Deville (1999) et présentée également dans Lavallée et Caron (2001). Cet article montre comment adapter cette méthode à l'enquête. Certaines extensions s'avèrent nécessaires. On développera l'une d'elle destinée à estimer le total d'une population dont on a tiré un échantillon bernoullien.

Mots clés : Méthode généralisée de partage des poids ; base incomplète et bases multiples.

1. Introduction

Une « enquête aux frontières » portant sur la fréquentation touristique extra-régionale en Bretagne (hormis celle des Bretons) a été réalisée sur la période d'avril à septembre 1997. L'Observatoire Régional du Tourisme de Bretagne et les Comités Départementaux de Tourisme aimeraient recommencer ce type d'enquête. Malheureusement ils n'ont plus la possibilité de recueillir une certaine masse d'informations récoltées aux frontières régionales ou intra-régionales, car les forces de police ne désirent plus collaborer à la réalisation d'enquêtes au bord des routes.

C'est pourquoi l'Observatoire Régional du Tourisme de Bretagne avec l'aide d'un comité technique constitué de méthodologues et d'opérateurs de terrain ont décidé de mettre en place une nouvelle méthodologie d'enquête en remplacement de la méthodologie des « enquêtes aux frontières ». De plus, l'évaluation de la part du tourisme intra-régional (des Bretons prenant des vacances en Bretagne, par exemple) est indispensable pour définir les facteurs de développement.

Un des problèmes majeurs est l'absence d'une base de sondage permettant d'interroger directement les touristes. Pour contourner ce problème, l'idée principale, déjà utilisée par la région des Asturies en Espagne (Valdés, De La Ballina, Aza, Loredo, Torres, Estébanez, Domínguez et Del Valle (2001) et Torres Manzanera, Sustacha Melijosa, Menéndez Estébanez et Valdés Pelaáez (2002)), est d'échantillonner des services destinés principalement aux touristes et de les interroger sur les différents lieux de ces nombreuses prestations touristiques. Il est bien évident qu'un touriste peut utiliser une ou plusieurs fois un ou plusieurs services de la base de sondage pendant la période d'enquête considérée. Pour pouvoir estimer des paramètres d'intérêts relatifs aux touristes, il faut avoir la possibilité d'échantillonner de façon rigoureuse certains services puis relier le jeu de poids des services échantillonnés au jeu de poids des touristes qui ont fréquenté ces services. Le but de cet article est de présenter une méthode qui permet de faire ce calcul. Cette méthode va s'appuyer principalement sur la *méthode généralisée de partage des poids* (MGPP) mise au point par Lavallée (1995), Lavallée (2002) et Deville (1999).

2. La méthode généralisée de partage des poids

On va rappeler très brièvement le principe de la *méthode généralisée de partage des poids* (MGPP). Pour de plus amples informations, on renvoie à Lavallée (1995), Lavallée (2002) et Deville (1999).

Soient U^A une population finie contenant N^A unités, où chaque unité est désignée par j et U^B une population finie contenant N^B unités, où chaque unité est désignée par i. La correspondance entre U^A et U^B peut être représentée par une matrice de liens $\Theta_{AB} = [\theta_{ji}^{AB}]$, de taille $N^A \times N^B$ où chaque élément $\theta_{ji}^{AB} \geq 0$. Autrement dit, l'unité j de U^A est reliée à l'unité i de U^B à condition que $\theta_{ji}^{AB} > 0$; sinon, il n'existe aucun lien entre les deux unités.

Dans le cas du sondage indirect, on sélectionne l'échantillon s^A de n^A unités à partir de U^A selon un plan d'échantillonnage donné. Soit $\pi_j^A > 0$, la probabilité de sélection de l'unité j. Pour chaque unité j sélectionnée dans s^A, on identifie les unités i de U^B pour lesquelles $\theta_{ji}^{AB} > 0$. Soit s^B, l'ensemble des n^B unités de U^B identifiées au moyen des unités $j \in s^A$, c'est-à-dire

$$s^B = \{i \in U^B ; \exists j \in s^A \text{ et } \theta_{ji}^{AB} > 0\}.$$

1. Jean-Claude Deville, Laboratoire de Statistique d'Enquêtes, ENSAI/CREST, Campus de Ker-Lann, 35170 BRUZ (France). Courriel : deville@ensai.fr; Myriam Maumy-Bertrand, Laboratoire de Statistique de l'Université Louis Pasteur, 7, rue René Descartes 67084 STRASBOURG Cedex (France). Courriel : mmaumy@math.u-strasbg.fr.

Pour chaque unité i de s^B, une variable d'intérêt y_i est mesurée.

On suppose que, pour toute unité j de s^A, on peut obtenir les valeurs de θ_{ji}^{AB} pour $i = 1, ..., N^B$ par entrevue directe ou à partir d'une source administrative. Pour toute unité i identifiée de U^B (ou seulement de s^B), on suppose que l'on peut obtenir les valeurs de θ_{ji}^{AB} pour $j = 1, ..., N^A$. Par conséquent, il n'est pas nécessaire de connaître les valeurs de θ_{ji}^{AB} pour la totalité de la matrice de liens Θ_{AB}. En fait, on ne doit connaître les valeurs de θ_{ji}^{AB} que pour les lignes j de Θ_{AB}, où $j \in s^A$, ainsi que pour les colonnes i de Θ_{AB} où $i \in s^B$.

Par exemple si le but est d'estimer une variable d'intérêt Y^B de la population cible U^B, où

$$Y^B = \sum_{i=1}^{N^B} y_i, \qquad (2.1)$$

avec y_i mesurée d'après l'ensemble U^B. On utilise alors un estimateur de la forme

$$\hat{Y}^B = \sum_{i=1}^{N^B} w_i y_i, \qquad (2.1)$$

où w_i est le poids d'estimation de l'unité i de s^B, avec $w_i = 0$ pour $i \notin s^B$. Pour obtenir une estimation sans biais d'une variable d'intérêt Y^B, il suffirait d'utiliser comme poids w_i l'inverse de la probabilité de sélection π_i^B de l'unité i. Comme il est mentionné dans Lavallée (1995) et Lavallée (2002), il est généralement difficile, voire impossible, d'obtenir ces probabilités. On a alors recours à la MGPP. Dans celle-ci les poids sont donnés par

$$w_i = \sum_{j \in s^A} \frac{\tilde{\theta}_{ji}^{AB}}{\pi_j^A},$$

où $\tilde{\theta}_{ji}^{AB} = \theta_{ji}^{AB} / \sum_{j=1}^{N^A} \theta_{ji}^{AB}$. De cette construction, l'estimateur $\hat{Y}^B$ est sans biais. De même, la variance de cet estimateur peut-être calculée et estimée car elle est identique à celle de

$$\sum_{j \in s^A} \frac{z_j}{\pi_j^A},$$

avec $z_j = \sum_{i=1}^{N^B} \tilde{\theta}_{ji}^{AB} y_i$.

3. L'enquête tourisme en milieu ouvert

3.1 Objectifs de l'enquête

Le principe de l'enquête est le suivant :

« *atteindre les touristes* (*étrangers ou français habitant la Bretagne ou pas*) *par le biais de services destinés à satisfaire leurs besoins élémentaires ou spécifiques* »

comme l'hébergement, la nourriture, les activités de loisirs, les transports.

3.2 La population d'intérêt

Soit G *un champ géographique* (les quatre départements bretons) et P *une période de référence* (pour nous celle qui s'étend du mois de février 2005 au mois de décembre 2005).

Un touriste est une personne ayant passé au moins une nuit dans G hors de sa résidence principale (nuitée).

Pour un touriste, *un séjour* est un intervalle sej de P de durée le cardinal de sej noté $|sej|$, au cours duquel le touriste passe toutes ses nuits dans G hors résidence principale et, les nuits immédiatement avant ou après le séjour sej étant passées hors de G (ou à la résidence principale).

Un voyage est un ensemble de touristes (ménage touristique) partageant le même séjour et avec le même hébergement au cours du séjour. On utilisera aussi le terme de ménage touristique par un léger abus de langage (un même ménage touristique peut faire plusieurs voyages au cours d'une période, mais nous n'avons aucun moyen de les distinguer).

L'unité statistique i de l'enquête est le voyage.

Les sous unités d'enquête sont les séjours, les touristes et les nuitées. Un voyage i comporte n_i touristes pendant le séjour de durée $|sej|$ et donc $n_i \times |sej|$ nuitées. Ici la population U^B est donc l'ensemble des voyages dans G au cours de P. ($sej \cap P \neq \varnothing$).

3.3 Le plan de sondage de l'enquête

Pour utiliser la MGPP, la population théorique U^A est constituée par un ensemble de « services ». Dans cette enquête, ceux-ci sont constitués par :

- les achats en boulangerie, constituant une première strate de U^A.
- les visites d'un ensemble de sites culturels ou de loisirs ou familiaux très connus. En pratique, pour chacun d'eux, un « point de passage obligé » a été défini. C'est l'ensemble des passages par ce point qui est la seconde strate de U^A.
- les passages sortant de la Bretagne au péage autoroutier de La Gravelle qui regroupe environ 80 % des sorties des touristes de la Bretagne en voiture. Ce mode de transport caractérise lui-même 80 % des séjours de non-résidents bretons. Ce passage constitue la troisième strate de U^A.

En d'autres termes, *la base de sondage* est donc formellement constituée de trois strates :

1. les achats en boulangerie;
2. les visites d'un ensemble de sites emblématiques de la Bretagne;
3. le passage au péage autoroutier de La Gravelle.

Dans *la première strate*, on réalise un échantillon à trois degrés :

- un échantillon de boulangeries;
- un échantillon de jours d'enquête;
- un échantillon de clients dans la boulangerie à un jour donné.

Dans *la deuxième strate*, on réalise un échantillon à deux degrés :

- un échantillon de jours d'enquête;
- un échantillon de personnes qui passent sur un des 16 sites référés à un jour donné.

Enfin dans *la troisième strate*, on réalise un échantillon à deux degrés :

- un échantillon de jours d'enquête;
- un échantillon de personnes qui passent au péage autoroutier de La Gravelle à un jour donné.

On admet que tout ménage touristique consomme au moins un des « services » (achats en boulangeries, visites de sites emblématiques de la Bretagne, passage au péage autoroutier de La Gravelle), ou tout du moins, que très peu de ménages ne consomment aucun d'entre eux.

Chaque échantillonnage (boulangerie, jours, « service ») requiert des techniques particulières et il serait très long de détailler chacune d'elle. On donnera néanmoins les quelques indications techniques suivantes :

- les boulangeries sont échantillonnées selon un plan classique stratifié géographiquement (cinq strates : partie « littorale » des quatre départements bretons, intérieur de la Bretagne). Dans chaque strate les boulangeries sont échantillonnées avec des probabilités proportionnelles à leur « potentiel touristique » construit à partir de leur chiffre d'affaire et de la capacité d'hébergement touristique et du nombre de résidences principales de la commune à laquelle elles appartiennent. Théoriquement du moins, car pratiquement ce tirage a été un peu « forcé » par des circonstances fortuites (refus de boulangers, fermetures durant certaines périodes par exemple).
- Les sites ne sont pas échantillonnés mais choisis pour leur notoriété et pour la possibilité technique d'y définir un « point de passage obligé » (parfois approximativement).
- Pour chaque boulangerie, chaque site et pour le péage autoroutier de La Gravelle, on a défini des « grappes de jours » complétement homogènes de chaque période P. Une grappe a été attribuée aléatoirement à chaque boulangerie, site ainsi qu'au péage autoroutier de La Gravelle. Pratiquement cela signifie qu'un enquêteur employé à plein temps est mobilisé sur plusieurs grappes.
- Pour chacun des « services » les utilisateurs sont échantillonnés selon des techniques habituelles de sélection aléatoire à mesure des arrivées : échantillon pseudo-systématique car pendant que l'enquêteur fait accepter un questionnaire, des gens passent sans qu'il puisse compter. Le nombre total de visiteurs ne peut donc pas être estimé directement. Si un site est accessible par une billeterie (musée ou château par exemple) l'échantillonnage s'appuie sur elle. Au final, l'échantillon d'utilisateurs d'un « service » à un jour donné est considéré comme un échantillon bernoullien, c'est-à-dire un sondage aléatoire simple si on connait la taille de la population c'est-à-dire le nombre de visiteurs du jour donné.

Remarques 3.1. La définition même du *touriste* est liée à l'hébergement, et il parait naturel d'utiliser une base directement liée à ce service. La pratique montre que c'est difficilement réalisable.

On n'a, d'abord, aucune base de sondage correcte pour l'hébergement non marchand (parents, amis, résidence secondaire) ni pour les locations meublées saisonnières.

Pour l'hébergement en hôtels, campings et gîtes familiaux, les tests de l'été 2004 ont montré l'existence de biais catastrophiques liés à l'intervention des hôteliers dans le processus de sélection des enquêtés. Ceux-ci ne respectent absolument pas les consignes d'échantillon aléatoire et distribuent « essentiellement » les questionnaires à leurs bons clients. Cette partie du dispositif de l'enquête a dû être abandonnée et remplacée par le passage au péage autoroutier de La Gravelle, qui est régulièrement l'objet d'enquêtes de qualité honnête faites par divers organismes.

Par ailleurs, les questionnaires collectés dans les boulangeries et sur les sites emblématiques de la Bretagne pendant l'été 2004, rendent apparemment (qualitativement et quantitativement) bien compte des différents modes d'hébergement.

De même, l'alimentation eut sans doute mieux été capturée par des questionnaires à la sortie des supermarchés. Mais là, le problème réside dans l'hétérogénéité de ces établissements et dans la lutte au couteau que se livrent les enseignes, le groupe C ... accepte les enquêtes dans ses établisements uniquement si le groupe I ... en est exclu ! En revanche, l'adhésion des artisans boulangers au concept de l'enquête a été excellente.

Remarques 3.2. Par définition même de la méthode utilisée, on se place formellement dans le contexte de l'échantillonnage à partir de bases multiples. Le problème a donné lieu à une abondante littérature (Hartley (1962), Lund (1968) et Hartley (1974) pour le moins). La MGPP s'applique à ce problème en considérant tout simplement

chaque base de sondage comme une strate à la condition de pouvoir identifier, pour chaque unité échantillonnée, l'ensemble des bases dans laquelle elle figure. Elle fournit alors une solution rigoureuse, efficace et uniquement basée sur le plan à ce problème. Cette remarque pourrait fonder un article autonome, mais les auteurs savent que cela n'en vaut pas la peine : une idée qui s'exprime en dix lignes n'a pas besoin d'un article ou d'un livre pour sa survie.

4. Les paramètres d'intérêt

On définit l'application F, qui à tout service j durant la période de référence P dans les trois types d'établissements du champ de l'enquête, associe le voyage i utilisateur de ce service.

$$\begin{aligned} F : \text{services} &\rightarrow \text{voyage} \\ j &\rightarrow F(j) = i. \end{aligned}$$

Soit U^B, la population des voyages i de la période de référence P. Cette population d'intérêt U^B est l'image par F de l'ensemble des services durant la période de référence P dans les trois types d'établissements du champ de l'enquête. La population U^A est l'image par F^{-1} de l'ensemble des voyages durant la période de référence P. Pour tout $i \in U^B$, on définit $R_i(B) = \text{card}(F^{-1}(i))$, le nombre d'antécédents de i au cours de la période d'enquête, c'est-à-dire, le nombre de services j utilisés par le ménage touristique i donné.

Les paramètres d'intérêt peuvent être des totaux, des effectifs ou des ratios. Supposons par exemple, que l'on s'intéresse à l'estimation d'un total relatif à une variable y définie sur la population U^B,

$$Y^B = \sum_{i \in U^B} y_i. \qquad (4.1)$$

Un cas particulier de ces totaux est l'effectif de U^B, noté N^B et défini par

$$N^B = \text{card}(U^B) = \sum_{i \in U^B} 1.$$

Par exemple, Y^B peut-être le nombre de personnes ayant pratiqué une certaine activité, le budget total dépensé par le ménage touristique à l'intérieur de la Bretagne, la provenance géographique des ménages touristiques, le nombre de jours que le ménage touristique passe en Bretagne. Il faut noter que pour beaucoup de variables, le total Y^B dépend de la taille du ménage touristique, c'est-à-dire le nombre de personnes qui forment ce groupe et de la longueur du séjour (uniquement les jours passés en Bretagne).

Désormais, on peut écrire

$$Y^B = \sum_{i \in U^B} y_i = \sum_{l=1}^{3} \sum_{a_l \in A_l} \sum_{d_l \in D_l} \sum_{j \in C_{d_l}} z_j, \qquad (4.2)$$

où

$$z_j = \frac{y_i}{R_i(B)}, \quad \text{pour} \quad j \in F^{-1}(i),$$

où

- A_1 : l'ensemble des boulangeries du champ de l'enquête repéré par l'indice a_1
- A_2 : les 16 lieux de passage du champ de l'enquête repérés par l'indice a_2
- A_3 : le péage autoroutier de La Gravelle repéré par l'indice a_3
- D_l : l'ensemble des jours d'enquête, repérés par l'indice d_l dans un établissement a_l de A_l, pour l variant de 1 à 3
- C_{d_l} : l'ensemble des services dans un établissement a_l de A_l de la journée d_l de D_l repérés par l'indice j.

5. Estimation sans biais d'un total

Dans le paragraphe précédent, on a montré que le total d'intérêt s'écrit comme un total sur l'ensemble des services du champ. Supposons que l'on dispose d'un échantillon de services répondants j, auxquels on peut associer des poids de sondage δ_j. Ces poids sont supposés sans biais car l'échantillon de services suit les canons d'un échantillon à plusieurs degrés, chaque sondage élémentaire étant sans biais.

Pour alléger les notations, on ne fait pas apparaître, dans ce qui suit, tous les degrés de tirage de l'échantillon en fonction de l'établissement a_l. Soient

- s^B : l'ensemble des ménages touristiques i correspondant à l'ensemble des services échantillonnés au cours de la période d'enquête
- s_{A_l} : l'ensemble des établissements échantillonnés
- s_{D_l} : l'ensemble des jours échantillonnés dans l'établissement a_l
- s_{d_l} : le sous-échantillon de services j correspondant au jour de l'établissement a_l.

Disposant d'un jeu de poids de sondage δ_j pour les services répondants, et si on connaît les $R_i(B)$, on estime alors le total Y^B sans biais par

$$\hat{Y}^B = \sum_{i \in s^B} w_i \, y_i \qquad (5.1)$$

où

$$w_i = \frac{\sum_{l=1}^{3} \sum_{s_{A_l}} \sum_{s_{D_l}} \sum_{s_{d_l}} \delta_j}{R_i(B)}.$$

On est ramené à une estimation sur la population des ménages touristiques. Cette formule n'est autre que celle donnée par la MGPP évoquée dans la section 2. Notons que $U^A = U^{A_1} \cup U^{A_2} \cup U^{A_3} = \bigcup_{l=1}^{3} U^{A_l}$, $\theta_{ji}^{AB} = 1$ si le service j a été utilisé par le voyage i et enfin $\delta_j = 1/\pi_j^A$.

L'estimation de la variance est possible selon les mêmes principes (*cf.* Lavallée (2002)). Elle ne sera pas détaillée ici car elle n'est qu'une application assez lourde en calcul des principes généraux.

De même, l'utilisation des informations auxiliaires sous forme de totaux, que ce soit dans les populations U^{A_l} ou dans la population U^B, ne pose pas de problème particulier que ce soit pour l'estimation ponctuelle ou pour l'estimation de la variance (*cf.* Lavallée (2002)).

Remarques 5.1. La procédure qui vient d'être décrite pour partager les poids peut-être considérée comme naïve. De fait, on sait optimiser la matrice de liens Θ_{AB} comme il est montré dans Deville et Lavallée (2006). L'application de l'enquête bretonne est décrite dans Deville, Lavallée et Maumy (2005).

6. Un exemple de problème particulier : Les points de visite en rase campagne

Comme on l'a déjà signalé, la mise en place de l'enquête sur le tourisme en Bretagne a nécessité d'assez nombreuses recherches complémentaires. On a déjà signalé ce qui concerne l'optimisation du partage des poids. L'utilisation d'informations auxiliaires relatives aux diverses bases et aux divers degrés de sondage est un autre chantier. On voudrait insister ici sur celui de l'estimation de certaines de ces informations auxiliaires, en particulier pour ce qui concerne les visites des sites touristiques en rase campagne.

Dans certains cas, on ne connaît malheureusement pas le nombre total de personnes, noté $T_P^{A_2}$, venant sur le site à un jour donné. En effet, dans l'ensemble A_2, on ne connaît pas tous les services (ici le nombre de visites) de la population. On ne peut donc pas avoir directement $\pi_j^{A_2}$ et donc δ_j pour $j \in A_2$. Pour contourner ce problème, on estime alors le nombre de visiteurs journaliers afin de déduire $\tilde{\pi}_j^{A_2} = n_{A_2} / \hat{T}_P^{A_2}$.

Dans la suite, on va développer deux approches d'estimation du nombre de visiteurs journaliers pour des sites accessibles en voitures uniquement (ou presque !). La première se base sur un système d'échantillonnage de voitures destiné à estimer le nombre de visiteurs sur le site. La seconde approche utilise un échantillon de visiteurs et est destinée à estimer la même quantité à partir de l'individu interrogé qui donne le nombre de personnes qui voyagent avec lui dans la voiture. Ces deux approches sont développées dans les sections 7 et 8 suivantes.

7. Construction d'un estimateur du nombre de visiteurs à partir d'un échantillonnage de voitures

Dans ce paragraphe, on est dans le cas où un enquêteur relève en « bâtonnant » (c'est le terme utilisé par les praticiens du tourisme) le nombre d'occupants des voitures, c'est-à-dire, relève le nombre de personnes dans une voiture qui franchissent l'endroit où un oeil électronique ou un système équivalent a été placé pour compter les voitures dont le nombre total noté T_V est connu avec une erreur de mesure négligeable près.

7.1 Définition et variance de $\hat{T}_P^{A_2}$

Le nombre total de voitures vaut

$$T_V = \sum_{\kappa=1,\ \ldots} t_\kappa = \sum_{l \in U_v} 1, \qquad (7.1)$$

où t_κ représente le nombre de voitures transportant κ personnes et U_V l'univers des voitures.

Remarques 7.1. Dans un souci d'allégement des notations, on utilisera ici et jusqu'à la fin de cet article, T_P pour $T_P^{A_2}$.

Le nombre total de personnes visitant le site vaut

$$T_P = \sum_{\kappa=1,\ \ldots} \kappa t_\kappa = \sum_{k \in U_P} 1, \qquad (7.2)$$

où U_P désigne l'univers des personnes. On a aussi l'égalité

$$T_P = \sum_{l \in U_v} v_l, \qquad (7.3)$$

où v_l est le nombre de personnes dans la voiture l.

Comme on l'a mentionné dans la section précédente, le nombre total de personnes T_P est inconnu. Par conséquent construisons un estimateur de T_P. Soit $\hat{T}_P$ le $\pi-$estimateur fondé sur s_V un échantillon aléatoire simple de voitures de taille n et de probabilité d'inclusion n/T_V

$$\hat{T}_P = \frac{T_V}{n} \sum_{l \in s_V} v_l = T_V \bar{v}, \qquad (7.4)$$

en posant

$$\bar{v} = \frac{1}{n}\left(\sum_{l \in s_V} v_l \right).$$

Il est clair que $\hat{T}_P$ est un estimateur sans biais du nombre total de personnes T_P et que $\bar{v}$ estime sans biais le nombre moyen $\bar{V}$ de personnes dans une voiture.

La variance de $\hat{T}_P$ est donc égale à

$$\begin{aligned}\mathrm{Var}[\hat{T}_P] &= T_V^2\left(\frac{1}{n} - \frac{1}{T_V}\right)S_V^2 \\ &= \frac{1}{n}\, T_V^2\, S_V^2 - T_V\, S_V^2, \end{aligned} \tag{7.5}$$

où S_V^2 désigne la variance corrigée de la population U_V.

7.2 Construction d'un estimateur d'une variable d'intérêt dans le cas d'un échantillonnage de voitures

On veut estimer une variable d'intérêt Y de la population U_P qui s'écrit sous la forme

$$Y = \sum_{k\in U_P} y_k, \tag{7.6}$$

où y_k est la variable d'intérêt qu'on mesure dans le questionnaire final. Soit $\hat{Y}$ le π – estimateur défini par

$$\hat{Y} = \sum_{k\in s_P} w_k^P\, y_k, \tag{7.7}$$

où le poids w_k^P est égal à $\hat{T}_P / m$. Par conséquent l'estimateur $\hat{Y}$ peut s'écrire

$$\hat{Y} = \frac{\hat{T}_P}{m}\sum_{k\in s_P} y_k = \hat{T}_P\,\bar{y} \tag{7.8}$$

en posant

$$\bar{y} = \frac{1}{m}\left(\sum_{k\in s_P} y_k\right).$$

Par la suite, les variables $\hat{T}_P$ et $\bar{y}$ seront supposées indépendantes. L'hypothèse est réaliste, car sur le terrain nous avons recours à deux enquêteurs indépendants.

7.2.1 Calcul de la variance de l'estimateur $\hat{Y}$

D'après le théorème de Huygens (1673), en conditionnant selon l'échantillon s_V, on obtient

$$\begin{aligned} V_Y &= \mathrm{Var}[\hat{Y}] \\ &= \bar{Y}^2\,\mathrm{Var}[\hat{T}_P] + T_P^2\,\mathrm{Var}[\bar{y}] \\ &\quad + \mathrm{Var}[\hat{T}_P]\,\mathrm{Var}[\bar{y}]. \end{aligned} \tag{7.9}$$

Dans le cas présent, on assimile l'échantillon à un sondage aléatoire simple sans remise. L'égalité (7.9) devient alors

$$\begin{aligned} V_Y &= \bar{Y}^2\left(\frac{1}{n}\,T_V^2\,S_V^2 - T_V\,S_V^2\right) \\ &\quad + T_P^2\left(\frac{1}{m}S_Y^2 - \frac{S_Y^2}{T_P}\right) \\ &\quad + \left(\frac{1}{n}T_V^2\,S_V^2 - T_V\,S_V^2\right)\left(\frac{1}{m}S_Y^2 - \frac{S_Y^2}{T_P}\right), \end{aligned}$$

avec $S_Y^2 = 1/(T_P - 1)\sum_{k\in U_P}(y_k - \bar{Y})^2$. En réorganisant les termes, on obtient

$$\begin{aligned} V_Y &= \left(\bar{Y}^2 - \frac{S_Y^2}{T_P}\right)T_V^2\,S_V^2\,\frac{1}{n} \\ &\quad + (T_P^2 - T_V\,S_V^2)\,S_Y^2\,\frac{1}{m} \\ &\quad + T_V^2\,S_V^2\,S_Y^2\,\frac{1}{nm} + \frac{T_V}{T_P}\,S_V^2\,S_Y^2 \\ &\quad - \bar{Y}^2\,T_V\,S_V^2 - T_P\,S_Y^2. \end{aligned}$$

On cherche maintenant l'allocation des tailles des échantillons s_P et s_V qui minimise la variance de l'estimateur $\hat{Y}$ pour des tailles de population T_P et T_V fixées.

On doit donc minimiser l'égalité (7.10) en n, m sous la contrainte

$$C_V\, n + C_P\, m = C,$$

où C_V désigne le coût (en temps par exemple) des questionnaires posés autour des voitures, C_P le coût (en temps) des questionnaires posés aux personnes et C le coût total.

On peut écrire l'équation lagrangienne

$$\begin{aligned} L(n,\, m,\, \lambda) &= \left(\bar{Y}^2 - \frac{S_Y^2}{T_P}\right)T_V^2\,S_V^2\,\frac{1}{n} \\ &\quad + (T_P^2 - T_V\,S_V^2)\,S_Y^2\,\frac{1}{m} \\ &\quad + T_V^2\,S_V^2\,S_Y^2\,\frac{1}{nm} + \frac{T_V}{T_P}\,S_V^2\,S_Y^2 \\ &\quad - \bar{Y}^2\,T_V\,S_V^2 - T_P\,S_Y^2 \\ &\quad + \lambda(C_V\, n + C_P\, m - C). \end{aligned} \tag{7.11}$$

En annulant les dérivées partielles par rapport aux variables n, m, λ, on obtient

$$\begin{aligned} \frac{\partial L}{\partial n}(n,\, m,\, \lambda) &= \left(\bar{Y}^2 - \frac{S_Y^2}{T_P}\right)T_V^2\,S_V^2\left(-\frac{1}{n^2}\right) \\ &\quad + T_V^2\,S_V^2\,S_Y^2\left(-\frac{1}{mn^2}\right) \\ &\quad + \lambda\, C_V = 0, \end{aligned}$$

$$\frac{\partial L}{\partial m}(n,\ m,\ \lambda) = (T_P^2 - T_V\ S_V^2)\, S_Y^2 \left(-\frac{1}{m^2}\right) + T_V^2\ S_V^2\ S_Y^2 \left(-\frac{1}{nm^2}\right) + \lambda\, C_P = 0,$$

$$\frac{\partial L}{\partial \lambda}(n,\ m,\ \lambda) = C_V\ n + C_P\, m - C = 0.$$

Après calculs, on obtient une équation du troisième degré en n qui s'écrit

$$\lambda C_V^2\, n^3 - \lambda C_V C n^2 - C_V T_V^2\ S_V^2 \left(\overline{Y}^2 - \frac{S_Y^2}{T_P}\right) n + T_V^2\ S_V^2 \left(C\left(\overline{Y}^2 - \frac{S_Y^2}{T_P}\right) + C_P\, S_Y^2\right) = 0.$$

Cette équation du troisième degré en n admet une solution réelle que l'on peut déterminer avec des méthodes numériques.

En faisant le même raisonnement, on obtient une équation du troisième degré en m

$$\lambda C_P^2\, m^3 - \lambda C_P C m^2 - C_P\, S_Y^2\, (T_P^2 - T_V\ S_V^2)\, m + S_Y^2\, (C(T_P^2 + T_V\ S_V^2) + C_V\, T_V^2\ S_V^2) = 0.$$

7.2.2 Cas simplifié

Pour simplifier le calcul de la variance de l'estimateur $\hat{Y}$, nous pouvons faire une approximation dans l'égalité (7.10). En effet, nous pouvons supposer que le terme $1/nm$ est négligeable devant les termes $1/n$ et $1/m$.

On obtient alors la transformation suivante de l'égalité (7.10)

$$V_Y = \left(\overline{Y}^2 - \frac{S_Y^2}{T_P}\right) T_V^2\ S_V^2\ \frac{1}{n} + (T_P^2 - T_V\ S_V^2)\ S_Y^2\ \frac{1}{m} + \frac{T_V}{T_P} S_V^2\ S_Y^2 - \overline{Y}^2\, T_V\ S_V^2 - T_P\ S_Y^2. \qquad (7.12)$$

On cherche maintenant l'allocation des tailles des échantillons s_P et s_V qui minimise la variance de l'estimateur $\hat{Y}$ pour des tailles de population T_P et T_V fixées.

On doit donc minimiser l'égalité (7.12) en n, m sous la contrainte

$$C_V\ n + C_P\, m = C.$$

On peut écrire l'équation lagrangienne

$$L(n,\ m,\ \lambda) = \left(\overline{Y}^2 - \frac{S_Y^2}{T_P}\right) T_V^2\ S_V^2\ \frac{1}{n} + (T_P^2 - T_V\ S_V^2)\, S_Y^2\ \frac{1}{m} + \frac{T_V}{T_P} S_V^2\ S_Y^2 - \overline{Y}^2\, T_V\ S_V^2 - T_P\ S_Y^2 + \lambda\, (C_V\ n + C_P\, m - C). \qquad (7.13)$$

En annulant les dérivées partielles par rapport aux variables n, m, λ, on obtient

$$\frac{\partial L}{\partial n}(n,\ m,\ \lambda) = \left(\overline{Y}^2 - \frac{S_Y^2}{T_P}\right) T_V^2\ S_V^2 \left(-\frac{1}{n^2}\right) + \lambda\, C_V = 0,$$

$$\frac{\partial L}{\partial m}(n,\ m,\ \lambda) = (T_P^2 - T_V\ S_V^2)\, S_Y^2 \left(-\frac{1}{m^2}\right) + \lambda\, C_P = 0,$$

$$\frac{\partial L}{\partial \lambda}(n,\ m,\ \lambda) = C_V\ n + C_P\, m - C = 0.$$

Après calculs, on obtient

$$n_{\text{opt}} = \frac{C}{\left(C_V + \sqrt{C_P\, C_V\, \dfrac{T_P\, S_Y^2\, (T_P^2 - T_V\, S_V^2)}{T_V^2\, S_V^2\, (T_P\, \overline{Y}^2 - S_Y^2)}}\right)},$$

$$m_{\text{opt}} = \frac{C}{\left(C_V + \sqrt{C_P\, C_V\, \dfrac{T_V^2\, S_V^2\, (T_P\, \overline{Y}^2 - S_Y^2)}{T_P\, S_Y^2\, (T_P^2 - T_V\, S_V^2)}}\right)}.$$

8. Construction d'un estimateur du nombre de visiteurs à partir d'un échantillonnage de visiteurs

La méthode précédente peut s'avérer compliquée et coûteuse à réaliser sur certains sites. On peut obtenir une collecte plus simple en demandant à la personne k le nombre u_k de passagers de la voiture i qui l'a transportée. Ce nombre u_k est ici égal à v_l pour la voiture l qui a transporté la personne k. Cette méthode a en outre l'avantage d'obtenir avec précision le nombre de passagers au sens de l'enquête (compte-t-on les bébés ?).

8.1 Définition de $\hat{\hat{T}}_P$

Rappelons l'égalité suivante

$$T_P = \sum_{l \in U_V} v_l,$$

où v_l désigne le nombre de passagers de la voiture l. Rappelons également

$$T_P = \sum_{l \in U_P} 1.$$

Le nombre moyen de passagers dans une voiture $\overline{V}$ peut s'exprimer sous la forme

$$\overline{V} = \frac{\sum_{l \in U_V} v_l}{\sum_{l \in U_V} 1} = \frac{\sum_{\kappa=1,\ \ldots} \kappa t_\kappa}{\sum_{\kappa=1,\ \ldots} t_\kappa} = \frac{\sum_{\kappa=1,\ \ldots} m_\kappa}{\sum_{\kappa=1,\ \ldots} M_\kappa / \kappa}, \qquad (8.1)$$

où t_κ est le nombre de voitures à κ passagers et M_κ le nombre de personnes venues dans une voiture à κ passagers.

Cette dernière relation permet de donner une dernière écriture de T_P

$$T_P = T_V \overline{V}. \qquad (8.2)$$

Par conséquent un estimateur de T_P s'écrit sous la forme suivante

$$\hat{\hat{T}}_P = T_V \hat{\overline{V}}, \qquad (8.3)$$

où le nombre total de voitures T_V est parfaitement connu. En observant cette expression, on constate que pour connaître l'estimateur $\hat{\hat{T}}_P$, il suffit de déterminer la quantité $\hat{\overline{V}}$. Introduisons alors l'estimateur suivant de $\overline{V}$

$$\hat{\overline{V}} = \frac{\sum_{\kappa \in s_P} m_\kappa}{\sum_{\kappa \in s_P} m\kappa / \kappa},$$

où m_κ est le nombre de personnes de l'échantillon voyageant dans une voiture à κ passagers. L'estimateur $\hat{\overline{V}}$ peut s'écrire également de la façon suivante

$$\hat{\overline{V}} = \frac{\sum_{k \in s_P} 1}{\sum_{k \in s_P} 1/u_k}$$

ou encore

$$\hat{\overline{V}} = \frac{m}{\sum_{k \in s_P} 1/u_k}. \qquad (8.4)$$

Cette dernière égalité nous permet d'écrire l'égalité suivante

$$\frac{1}{\hat{\overline{V}}} = \frac{1}{m} \sum_{k \in s_P} \frac{1}{u_k}. \qquad (8.5)$$

Cette dernière quantité représente la moyenne empirique des $1/u_k$ et $\hat{\overline{V}}$ est la moyenne harmonique des u_k. On peut d'ailleurs calculer sa variance qui est égale à

$$\mathrm{Var}\left[\frac{1}{\hat{\overline{V}}}\right] = \left(\frac{1}{m} - \frac{1}{T_P}\right) S^2_{1/u}. \qquad (8.6)$$

8.2 Calcul de la variance de l'estimateur de $\hat{\hat{T}}_P$ sans échantillonnage de voitures

Reste à calculer la variance de l'estimateur $\hat{\overline{V}}$ sachant (8.6). Pour cela, remarquons que l'on peut écrire

$$\begin{aligned}\frac{1}{\hat{\overline{V}}} &= \frac{1}{\overline{V}\left(\frac{\hat{\overline{V}}}{\overline{V}} - 1 + 1\right)} \\ &= \frac{1}{\overline{V}} \times \frac{1}{1 + \frac{\hat{\overline{V}} - \overline{V}}{\overline{V}}} \\ &= \frac{1}{\overline{V}}\left(1 - \frac{\hat{\overline{V}} - \overline{V}}{\overline{V}} + o\left(\frac{\hat{\overline{V}} - \overline{V}}{\overline{V}}\right)\right).\end{aligned}$$

Par conséquent, on obtient

$$\mathrm{Var}\left[\frac{1}{\hat{\overline{V}}}\right] \simeq \left(\frac{1}{\overline{V}}\right)^2 \times \frac{\mathrm{Var}[\hat{\overline{V}}]}{\overline{V}^2}.$$

Finalement, on a

$$\mathrm{Var}[\hat{\overline{V}}] \simeq \overline{V}^4 \times \mathrm{Var}\left[\frac{1}{\hat{\overline{V}}}\right],$$

ou encore, avec (8.6)

$$\mathrm{Var}[\hat{\overline{V}}] \simeq \overline{V}^4 \times \left(\frac{1}{m} - \frac{1}{T_P}\right) S^2_{1/u}. \qquad (8.7)$$

Or par définition, la variance $S^2_{1/u}$ est égale à

$$S^2_{1/u} = \frac{1}{T_P - 1} \sum_{k \in U_P} \left(\frac{1}{u_k} - \frac{1}{\overline{V}}\right)^2. \qquad (8.8)$$

Comme la quantité T_P est inconnue, cette relation peut être estimée par

$$\frac{1}{m - 1} \sum_{k \in s_P} \left(\frac{1}{u_k} - \frac{1}{\overline{V}}\right)^2. \qquad (8.9)$$

Grâce à (8.7) et à (8.9) on peut donc connaître facilement la variance de l'estimateur $\hat{\overline{V}}$ et par conséquent celle de l'estimateur $\hat{\hat{T}}_P$ et finalement celle de la variable d'intérêt $\hat{Y}$.

Remarques 8.1. L'estimateur $\hat{\hat{T}}_P$ est biaisé et asymptotiquement sans biais.

Remarque 8.2. Si les variables $\hat{\hat{T}}_P$ et $\overline{y}$ ne sont pas indépendantes alors on aurait

$$\begin{aligned}\mathrm{Var}\left[\hat{\hat{T}}_P\,\bar{y}\right] &= \bar{Y}^2\,\mathrm{Var}\left[\hat{\hat{T}}_P\right] + T_P^2\,\mathrm{Var}[\bar{y}] \\ &\quad + \mathrm{Var}\left[\hat{\hat{T}}_P\,\bar{y}\right]\mathrm{Var}[\bar{y}] \\ &\quad + \text{termes liées à la non indépendance éventuelle des variables } \hat{\hat{T}}_P \text{ et } \bar{y}.\end{aligned}$$

9. Illustration numérique

Un compteur mécanique d'un site en rase campagne donne $T_V = 100$ voitures. On supose qu'il y a 20 % de voitures à une personne, 20 % de voitures à deux personnes, 20 % de voitures à trois personnes, 20 % de voitures à quatre personnes, 20 % de voitures à cinq personnes. Ainsi, on a 300 visiteurs sur ce site. La variance S_V^2 est égale à deux en négligeant les corrections de population finie. Le nombre moyen de passagers $\bar{V}$ est de trois. En effet, on a :

$$\frac{1}{V} = \frac{1}{1}\times\frac{20}{300} + \frac{1}{2}\times\frac{40}{300} + \frac{1}{3}\times\frac{60}{300} + \frac{1}{4}\times\frac{80}{300} + \frac{1}{5}\times\frac{100}{300} = \frac{1}{3}.$$

D'où $\bar{V} = 3$.

Calculons maintenant une estimation de $S_{1/u}^2$. Après simplifications de (8.8) et en supposant que T_P est suffisamment grand devant un, on a

$$S_{1/u}^2 \simeq \frac{1}{T_P}\sum_{k\in U_P}\frac{1}{u_k^2} - \left(\frac{1}{\bar{V}}\right)^2.$$

Ainsi, on a

$$\begin{aligned}S_{1/u}^2 &= \frac{1}{30}\left(2 + 1 + \frac{2}{3} + \frac{1}{2} + \frac{2}{5}\right) - \frac{1}{3^2} \\ &= \frac{1}{30}\left(\frac{60 + 30 + 20 + 15 + 12}{30}\right) - \frac{1}{3^2} \\ &= \frac{137}{30^2} - \frac{1}{3^2} = \frac{37}{30^2}.\end{aligned}$$

Pusique nous connaissons $S_{1/u}^2$, nous pouvons calculer la variance de l'estimateur $\bar{V}$. Ainsi on a

$$\mathrm{Var}[\hat{\bar{V}}] \simeq 3^4 \times \frac{37}{30^2}\times\frac{1}{m}.$$

Enfin, on peut calculer la variance de l'estimateur $\hat{\hat{T}}_P$

$$\begin{aligned}\mathrm{Var}\left[\hat{\hat{T}}_P\right] &= T_V^2\,\mathrm{Var}\left[\hat{\bar{V}}\right] \\ &\simeq 10^4 \times 3^4 \times \frac{37}{30^2}\times\frac{1}{m}.\end{aligned}$$

La première approche donne une variance de l'estimateur $\hat{T}_P$ égale à

$$\mathrm{Var}\left[\hat{T}_P\right] = 10^4 \times 2 \times \frac{1}{n}.$$

Donc, afin que l'estimateur $\hat{\hat{T}}_P$ ait la même variance que l'estimateur $\hat{T}_P$, il suffit que la taille m de l'échantillon s_P soit égale à

$$m \simeq 1{,}66\,n.$$

En première conclusion, on peut dire que la seconde approche rend les opérations de terrain plus simples et moins coûteuses en termes de personnel car elle ne nécessite qu'un seul enquêteur. Elle est plus précise qu'un comptage sans contact pour obtenir la composition du ménage touristique. Elle ne nécessite qu'un échantillon environ une fois et demie plus gros que la première approche pour apporter la même précision, ce qui est tout à fait tolérable vu la simplification de la collecte qui en résulte. En pratique donc, sur tous les sites on appliquera de préférence la seconde méthode.

Conclusion

Cet article a présenté les grandes lignes d'une nouvelle méthode applicable à la statistique du tourisme. Elle consiste à saisir les touristes à partir de la consommation de certains services sur lesquels on sait construire des échantillons probabilistes. La méthode de partage des poids permet de passer de l'exactitude statistique sur les services à l'exactitude sur les unités statistiques pertinentes en tourisme : le voyage, le séjour, le ménage touristique, le touriste ou la nuitée. Cependant la méthode requiert de nombreuses adaptations et compléments au partage des poids. On s'est attardé à l'une d'elles qui est l'estimation du nombre de visiteurs d'un site en rase campagne. Deux méthodes pouvaient être mises en concurrence. L'une, plus précise en terme de taille d'échantillon, demande en fait une organisation relativement lourde et fait courir le risque d'erreurs de mesures désagréables. Au prix d'une collecte de données un peu plus abondante, on préfère donc la seconde méthode.

D'autres études de ce genre ont été faites avant et pendant la réalisation de l'enquête, de sorte que la méthodologie complète est difficile à résumer en un seul article.

Remerciements

Les auteurs remercient chaleureusement les deux arbitres et l'éditeur associé qui ont grandement contribué à améliorer la lisibilité de ce texte.

Bibliographie

Deville, J.-C. (1999). Les enquêtes par panel : En quoi différent-elles des autres enquêtes ? suivi de : Comment attraper une population en se servant d'une autre. *Actes des journées de méthodologie statistiques*, *INSEE Méthodes*, 84-85-86, 63-82.

Deville, J.-C., et Lavallée, P. (2006). Sondage indirect : Les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête*. 32, 2, 185-196.

Deville, J.-C., Lavallée, P. et Maumy, M. (2005). Composition, factorisation et conditions d'optimalité (faible, forte) dans la méthode de partage des poids. Application à l'enquête sur le tourisme en Bretagne. *Actes des journées de méthodologie statistiques*, *INSEE Méthodes*.

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

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

Huygens, C. (1673). Horologium Oscillatorium sive de motu pendulorum.

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

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

Lavallée, P., et Caron, P. (2001). Estimation par la méthode généralisée du partage des poids : Le cas du couplage d'enregistrements. *Techniques d'enquête*, 27, 171-187.

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

Torres Manzanera, E., Sustacha Melijosa, I., Menéndez Estébanez, J.M. et Valdés Pelaáez, L. (2002). A solution to problems and disadvantages in statistical operations of surveys of visitors at accommodation establishments and at popular visitors places. (Éd. Ákos Probáld). *Proceedings Of The Sixth International Forum On Tourisme Statistics. Hungarian Central Statistical Office*, *Budapest*.

Valdés, L., De La Ballina, J., Aza, R., Loredo, E., Torres, E., Estébanez, J.M., Domínguez, J.S. et Del Valle, E. (2001). A methodology to measure tourism expenditure and total tourism production at the regional level. Dans *Tourism Statistics* : *International Perspectives and Current Issues*, (Éd. Lennon, J.J.), Continuum, Grande Bretagne, 317-334.

Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 207-216
Statistique Canada, N° 12-001 au catalogue

Combinaison de l'échantillonnage par dépistage de liens et de l'échantillonnage en grappes pour estimer la taille de populations cachées : Une approche assistée par la méthode bayésienne

Martín H. Félix-Medina et Pedro E. Monjardin [1]

Résumé

Félix-Medina et Thompson (2004) ont proposé une variante de l'échantillonnage par dépistage de liens dans laquelle on suppose qu'une part de la population (qui n'est pas nécessairement la plus grande) est couverte par une liste d'emplacements disjoints où les membres de la population peuvent être trouvés avec une probabilité élevée. Après la sélection d'un échantillon d'emplacements, on demande aux personnes se trouvant à chacun de ces emplacements de nommer d'autres membres de la population. Les deux auteurs ont proposé des estimateurs du maximum de vraisemblance des tailles de population qui donnent des résultats acceptables à condition que, pour chaque emplacement, la probabilité qu'un membre de la population soit nommé par une personne se trouvant à cet emplacement, appelée probabilité de nomination, ne soit pas faible. Dans la présente étude, nous partons de la variante de Félix-Medina et Thompson, et nous proposons trois ensembles d'estimateurs des tailles de population dérivés sous une approche bayésienne. Deux des ensembles d'estimateurs sont obtenus en utilisant des lois a priori incorrectes des tailles de population, et l'autre en utilisant des lois a priori de Poisson. Cependant, nous n'utilisons la méthode bayésienne que pour faciliter la construction des estimateurs et adoptons l'approche fréquentiste pour faire les inférences au sujet des tailles de population. Nous proposons deux types d'estimateurs de variance et d'intervalles de confiance partiellement fondés sur le plan de sondage. L'un d'eux est obtenu en utilisant un bootstrap et l'autre, en suivant la méthode delta sous l'hypothèse de normalité asymptotique. Les résultats d'une étude par simulation indiquent que i) quand les probabilités de nomination ne sont pas faibles, chacun des ensembles d'estimateurs proposés donne de bon résultats et se comporte de façon fort semblable aux estimateurs du maximum de vraisemblance, ii) quand les probabilités de nomination sont faibles, l'ensemble d'estimateurs dérivés en utilisant des lois a priori de Poisson donne encore des résultats acceptables et ne présente pas les problèmes de biais qui caractérisent les estimateurs du maximum de vraisemblance et iii) les résultats précédents ne dépendent pas de la taille de la fraction de la population couverte par la base de sondage.

Mots clés : Approche bayésienne; capture-recapture; approche fondée sur le plan de sondage; population finie; population d'accès difficile; maximum de vraisemblance; approche fondée sur un modèle; base de sondage.

1. Introduction

L'échantillonnage par dépistage de liens (EDP) s'est avéré être une méthode convenant bien à l'échantillonnage de populations humaines cachées ou d'accès difficile, comme les drogués, les sans abri ou les travailleurs clandestins. Il consiste à sélectionner un échantillon initial de personnes parmi la population cible et de demander à ces personnes de nommer d'autres membres de la population. Les personnes nommées qui ne figurent pas déjà dans l'échantillon initial sont alors incluses dans l'échantillon et il peut leur être demandé de nommer d'autres personnes. Ce processus se poursuit jusqu'à ce qu'une règle d'arrêt préétablie soit satisfaite (pour une revue de l'échantillonnage par dépistage de liens, voir Spreen 1992, ainsi que Thompson et Frank 2000).

Bien que l'échantillonnage par dépistage de liens permette à l'échantillonneur de faire des inférences fondées sur un modèle valides au sujet d'un certain nombre de paramètres de population, en pratique, les hypothèses concernant l'échantillon initial sont difficiles à satisfaire. (Voir Snijders 1992, Frank et Snijders 1994, et Heckathorn 2002). Par exemple, Frank et Snijders (1994) ont élaboré une variante de l'échantillonnage par dépistage de liens dans laquelle l'échantillon initial est un échantillon de Bernoulli, c'est-à-dire que les éléments de l'échantillon initial sont sélectionnés indépendamment et avec probabilités égales; toutefois, dans les études réelles, le recrutement initial est généralement réalisé au moyen de dossiers de personnes fournis par des centres de soins de santé et des postes de police, ce qui introduit un biais de sélection appelé biais institutionnel.

La difficulté à satisfaire les hypothèses au sujet de l'échantillon initial dans les situations pratiques ont poussé Félix-Medina et Thompson (2004) à élaborer une variante de l'échantillonnage par dépistage de liens qui ne nécessite pas d'échantillon de Bernoulli initial. Ils supposent qu'une part, qui n'est pas nécessairement la plus grande, de la population cible est couverte par une base de sondage constituée d'emplacements accessibles où les membres de la population peuvent être trouvés avec une forte probabilité (par exemple, bars, hôpitaux, îlots d'habitations ou parcs).

1. Martin H. Félix-Medina et Pedro E. Monjardin, Escuela de Ciencias Físico-Matemáticas, Universidad Autónoma de Sinaloa, Ciudad Universitaria, Culiacán Sinaloa, México.

Un échantillon aléatoire simple d'emplacements est sélectionné et les membres de la population appartenant à chaque emplacement sont identifiés. Enfin, comme dans l'échantillonnage par dépistage de liens ordinaire, il est demandé aux personnes se trouvant à chaque emplacement de nommer d'autres membres de la population.

Ces auteurs ont dérivé des estimateurs du maximum de vraisemblance (EMV) des tailles de population à partir de modèles probabilistes qui décrivent le nombre d'éléments découverts à chaque emplacement, ainsi que la probabilité qu'un membre de la population soit nommé à un emplacement, à laquelle on donne le nom de probabilité de nomination. Ils proposent aussi des estimateurs de variance fondés sur un modèle et partiellement fondés sur le plan de sondage, c'est-à-dire des estimateurs fondés à la fois sur le plan de sondage utilisé pour sélectionner l'échantillon initial et sur les modèles hypothétiques. Tout au long de l'article, nous parlerons d'estimateur « de type fondé sur le plan » pour faire référence à ce genre d'estimateur. Grâce à une étude par simulation, les auteurs ont montré que les EMV des tailles de population et leurs estimateurs de variance de type fondé sur le plan sont robustes aux écarts par rapport au modèle hypothétique, mais que les estimateurs de variance fondés sur un modèle ne sont pas robustes. En outre, ils ont constatés que les EMV ont tendance à surestimer gravement la taille de population si les probabilités de nomination sont faibles.

Comme l'ont indiqué ces auteurs, le problème de la surestimation qui se manifeste lorsque les probabilités de nomination sont faibles est dû à la petite quantité d'information que contient l'échantillon, quantité qui n'est pas suffisante pour obtenir des estimations stables des probabilités de nomination. Selon eux, un remède éventuel à ce problème consiste à suivre l'approche bayésienne pour construire des estimateurs dans lesquels sont intégrés des renseignements supplémentaires au sujet des paramètres de population.

Ici, nous utilisons la méthode bayésienne pour faciliter la construction des estimateurs des tailles de population, mais nous faisons les inférences selon une approche fréquentiste. Donc, en plus de calculer des estimateurs ponctuels, nous construisons des intervalles de confiance. Nous adoptons pour cela la stratégie proposée par Félix-Medina et Thompson (2004) en vue de construire des intervalles de confiance basés sur la loi normale et utilisons des estimateurs de variance de type fondé sur le plan obtenus par la méthode delta. En outre, nous construisons des intervalles de confiance bootstrap de type fondé sur le plan. Nous disons que cette approche inférentielle est « assistée par la méthode bayésienne ».

2. Plan d'échantillonnage et notation

La structure de la population et le plan d'échantillonnage que nous considérons dans le présent article sont les mêmes que ceux proposés par Félix-Medina et Thompson (2004). En voici une brève description. Soit $U = \{u_1, \ldots, u_\tau\}$ une population humaine cachée de taille inconnue τ. Soit U_1 un sous-ensemble de U formé par un nombre inconnu τ_1 de personnes que l'on peut trouver dans différents emplacements accessibles, comme des bars, des parcs ou des îlots d'habitations. Ce plan d'échantillonnage s'appuie sur les hypothèses qu'il est possible de construire une base de sondage de N de ces emplacements et que le chercheur a établi une règle opérationnelle qui lui permet de déterminer si une personne appartient ou non à un emplacement figurant dans la base de sondage et, dans l'affirmative, de situer cet emplacement. Soulignons que l'on ne suppose pas que le sous-ensemble U_1 couvert par la base de sondage représente la partie principale de U et que, comme dans l'échantillonnage en grappes ordinaire, on suppose qu'une personne figurant dans la base de sondage n'appartient qu'à un seul emplacement. Soit A_i le i^e emplacement ou grappe dans la base de sondage et m_i le nombre de personnes qui appartiennent à $A_i, i = 1, \ldots, N$; alors $\tau_1 = \sum_1^N m_i$. Enfin, soit $U_2 = U - U_1$ la partie de U non couverte par la base de sondage et soit $\tau_2 = \tau - \tau_1$ sa taille.

Le plan d'échantillonnage est le suivant. Un échantillon $S_0 = \{A_1, \ldots, A_n\}$ de n grappes est tiré à partir de la base de sondage par échantillonnage aléatoire simple sans remise, et les m_i personnes qui appartiennent à chaque $A_i \in S_0$ sont identifiées. Notons que nous avons utilisé les indices $1, \ldots, n$ pour dénoter les grappes dans S_0; toutefois, cela ne signifie pas que les n premières grappes dans la base de sondage sont nécessairement les grappes contenues dans l'échantillon. Puis, on demande aux personnes comprises dans la grappe échantillonnée A_i de nommer des membres de U, mais seules les personnes nommées comprises dans $U - A_i$ sont prises en considération. Cette procédure est répétée pour chaque grappe $A_i \in S_0$. Par convention, nous dirons qu'une personne est nommée par une grappe si elle est nommée par au moins un membre de cette grappe. Les nominations à partir des diverses grappes sont faites indépendamment les unes des autres, et diverses stratégies de nomination peuvent être utilisées à différents emplacements. Par exemple, à l'emplacement A_i, les nominations pourraient être faites par les m_i membres, en tant que groupe, tandis que dans un autre emplacement, A_j, chacun des m_j membres pourrait faire des nominations séparément. Enfin, pour chaque personne nommée, le chercheur doit enregistrer le ou les emplacements qui l'ont nommé, ainsi que le segment U_1 ou U_2 de la population auquel elle appartient. Il convient de souligner que ce dernier élément

d'information peut être obtenu auprès de la personne qui a fait la nomination ou, si cela est impossible, durant une interview de la personne nommée.

La nomination de personnes par les grappes sera indiquée au moyen des matrices $\mathbf{X}_1 = [x_{ij}^{(1)}]_{n\times\tau_1}$ et $\mathbf{X}_2 = [x_{ij}^{(2)}]_{n\times\tau_2}$, où $x_{ij}^{(1)} = 1$ si la personne $u_j \in U_1 - A_i$ est nommée par la grappe A_i, et $x_{ij}^{(1)} = 0$ si $u_j \in A_i$ ou u_j n'est pas nommée par A_i. De même, $x_{ij}^{(2)} = 1$ si la personne $u_j \in U_2$ est nommée par la grappe A_i, et $x_{ij}^{(2)} = 0$ autrement. Comme l'on souligné Félix-Medina et Thompson (2004), $\mathbf{X}_1$ et $\mathbf{X}_2$ ne sont connues que jusqu'aux permutations de leurs colonnes, car les personnes ne sont pas étiquetées. Par conséquent, les inférences au sujet de τ_1 et τ_2 sont basées sur l'ensemble des dénombrements $\mathbf{y} = \{y_\omega\}$, où y_ω, $\omega \subseteq \mathbf{\Omega} = \{1, \ldots, n\}, \omega \neq \varnothing$, indique le nombre de personnes dans U qui sont nommées par chaque grappe échantillonnée A_i pour laquelle i est compris dans l'ensemble ω, mais autrement non. Par exemple, si $\omega = \{4, 7, 8\}$, y_ω serait le nombre de personnes dans U qui sont nommées par A_4, A_7 et A_8 uniquement.

3. Estimateurs des tailles de population basés sur les modes a posteriori

Félix-Medina et Thompson ont constaté la ressemblance entre leur plan d'échantillonnage et celui de l'échantillonnage par capture-recapture multiple (ECRM). Cela nous permet d'appliquer à notre cas certains des modèles bayésiens qui ont été proposés pour l'analyse de l'ECRM. Voir Fienberg, Johnson et Junker (1999) pour une revue des analyses bayésienne de l'ECRM. Dans la présente étude, nous utilisons un modèle pris en considération par Castledine (1981) pour les lois a priori des logits des probabilités de nomination, ainsi que certains modèles pour les lois a priori des tailles de population.

À l'instar de Félix-Medina et Thompson (2004), nous supposons que les tailles $m_1, \ldots, m_N$ des grappes $A_1, \ldots, A_N$ sont des réalisations de variables aléatoires de Poisson indépendantes $M_1, \ldots, M_N$ de moyenne λ_1. Nous dénotons par $p_i^{(k)}$ la probabilité qu'une personne comprise dans $U_k - A_i$ soit nommée par l'emplacement $A_i \in S_0$. Les probabilités $p_i^{(k)}$ sont appelées probabilités de nomination. En outre, nous supposons que, conditionnellement aux tailles $m_1, \ldots, m_n$ des grappes dans S_0, à τ_1 et τ_2, ainsi qu'aux $p_i^{(k)}$, les variables $x_{ij}^{(k)}$ sont des réalisations de variables aléatoires de Bernoulli indépendantes $X_{ij}^{(k)}$ de moyenne $p_i^{(k)}$, $i = 1, \ldots, n$ et $k = 1, 2$.

Félix-Medina et Thompson (2004) ont utilisé le fait que la loi conditionnelle conjointe de $(M_1, \ldots, M_n, \tau_1 - \sum_1^n M_i)$, sachant que $\sum_1^N m_i = \tau_1$, est une loi multinomiale dont les paramètres sont τ_1 et $(1/N, \ldots, 1/N, 1 - n/N)$, et ont appliqué une méthode utilisée par Darroch (1958) pour montrer que la fonction de vraisemblance de $\tau_1, \tau_2, \mathbf{p}_1 = \{p_i^{(1)}\}_1^n$ et $\mathbf{p}_2 = \{p_i^{(2)}\}_1^n$ est le produit des facteurs suivants :

$$f(\mathbf{m}_s|\tau_1) = \frac{\tau_1!}{(\tau_1 - m)!\prod_1^n m_i!}(1/N)^m(1 - n/N)^{\tau_1 - m}$$

$$f(\mathbf{y}^{(1-0)}|\mathbf{m}_s, \tau_1, \mathbf{p}_1) = \frac{(\tau_1 - m)!}{(\tau_1 - m - r_1)!\prod_{\omega\neq\varnothing} y_\omega^{(1-0)}!}\prod_{i=1}^n [p_i^{(1)}]^{z_i^{(1-0)}} \times [1 - p_i^{(1)}]^{\tau_1 - m - z_i^{(1-0)}}$$

$$f(\mathbf{y}^{(A_1)}, \ldots, \mathbf{y}^{(A_n)}|\mathbf{m}_s, \mathbf{p}_1) = \prod_{i=1}^n \frac{m_i!}{(m_i - w_i)!\prod_{\omega\neq\varnothing} y_\omega^{(A_i)}!}[p_i^{(1)}]^{z_i^{(0)}} \times [1 - p_i^{(1)}]^{m - m_i - z_i^{(0)}}$$

$$f(\mathbf{y}^{(2)}|\mathbf{m}_s, \tau_2, \mathbf{p}_2) = \frac{\tau_2!}{(\tau_2 - r_2)!\prod_{\omega\neq\varnothing} y_\omega^{(2)}!}\prod_{i=1}^n [p_i^{(2)}]^{z_i^{(2)}}[1 - p_i^{(2)}]^{\tau_2 - z_i^{(2)}},$$

où $\mathbf{m}_s = \{m_i\}_1^n; m = \sum_1^n m_i$ est la valeur observée de la variable aléatoire M qui indique le nombre de personnes dans S_0; $\mathbf{y}^{(1-0)} = \{y_\omega^{(1-0)}\}_{\omega\neq\varnothing}, \mathbf{y}^{(2)} = \{y_\omega^{(2)}\}_{\omega\neq\varnothing}$, et $\mathbf{y}^{(A_i)} = \{y_\omega^{(A_i)}\}_{\omega\neq\varnothing}, A_i \in S_0$, sont les ensembles de dénombrements obtenus à partir de $\mathbf{y}$, qui correspondent aux dénombrements de personnes nommées dans $U_1 - S_0, U_2$ et $A_i \in S_0$, respectivement; $z_i^{(0)} = \sum_{j\neq i}^n \sum_{\omega\supset i} y_\omega^{(A_j)}$, $z_i^{(1-0)} = \sum_{\omega\supset i} y_\omega^{(1-0)}$ et $z_i^{(2)} = \sum_{\omega\supset i} y_\omega^{(2)}$ sont les valeurs observées des variables aléatoires $Z_i^{(0)}, Z_i^{(1-0)}$ et $Z_i^{(2)}$ qui indiquent le nombre de personnes distinctes dans $S_0, U_1 - S_0$ et U_2, respectivement, qui sont nommées par A_i; et $r_1 = \sum_{\omega\neq\varnothing} y_\omega^{(1-0)}, r_2 = \sum_{\omega\neq\varnothing} y_\omega^{(2)}$ et $w_i = \sum_{\omega\neq\varnothing} y_\omega^{(A_i)}$ sont les valeurs observées des variables aléatoires R_1, R_2 et W_i qui indiquent le nombre de personnes distinctes dans $U_1 - S_0, U_2$ et A_i, respectivement, qui sont nommées par au moins une des grappes comprises dans S_0.

Nous allons maintenant nous pencher sur le problème de la définition des lois a priori de $\tau_1, \tau_2, \mathbf{p}_1$ et $\mathbf{p}_2$. Dans le cas de τ_1 et τ_2, nous considérerons les trois modèles qui suivent pour les lois a priori :

Lois de Poisson-Gamma

$\pi(\tau_1|\lambda_1) \propto (N\lambda_1)^{\tau_1}/\tau_1!$ et $\pi(\lambda_1) \propto \lambda_1^{a_1 - 1}e^{-b_1\lambda_1}$,
$\pi(\tau_2|\lambda_2) \propto \lambda_2^{\tau_2}/\tau_2!$ et $\pi(\lambda_2) \propto \lambda_2^{a_2 - 1}e^{-b_2\lambda_2}$,
où a_1, b_1, a_2, b_2 sont des constantes connues, et (τ_1, λ_1) et (τ_2, λ_2) sont indépendants.

Lois de Jeffreys

$\pi(\tau_k) \propto 1/\tau_k$, où $k = 1, 2$, et τ_1 et τ_2 sont des variables aléatoires indépendantes.

Lois uniformes

$\pi(\tau_k) \propto 1$, où $k = 1, 2$, et τ_1 et τ_2 sont des variables aléatoires indépendantes.

La loi a priori de Poisson de τ_1 définie dans le premier cas a pour motivation le fait que $\tau_1 = \sum_1^N M_i$, et que M_i est une variable de Poisson de moyenne λ_1. Soulignons que ce cas permet aux chercheurs d'utiliser l'information au sujet de τ_1 et τ_2 qui est connue avant l'observation de l'échantillon. Par ailleurs, les lois définies dans les deux autres cas ne sont pas informatives.

Dans le cas des probabilités de nomination $p_i^{(k)}$, à l'instar de Castledine (1981), nous supposons qu'elles sont échangeables et nous utiliserons le modèle normal à deux degrés pour les logits $\alpha_i^{(k)} = \log[p_i^{(k)}/(1-p_i^{(k)})]$ des $p_i^{(k)}$:

$$\alpha_i^{(k)} | \theta_k \sim N(\theta_k, \sigma_k^2),$$

et

$$\theta_k \sim N(\mu_k, \gamma_k^2); i = 1, \ldots, n, \; k = 1, 2, \qquad (1)$$

où $N(\theta_k, \sigma_k^2)$ représente la loi normale de moyenne θ_k et de variance σ_k^2; σ_k^2, μ_k et γ_k^2 sont des constantes connues; et les $\alpha_i^{(k)}$ sont conditionnellement indépendants sachant θ_k. Sous l'hypothèse d'échangeabilité, les $\alpha_i^{(k)}$ ne sont pas indépendants, mais l'information au sujet de n'importe lequel d'entre eux est utilisée pour obtenir des renseignements sur n'importe lequel des $\alpha_i^{(k)}$. Naturellement, si nous voulions des lois a priori indépendantes pour les $\alpha_i^{(k)}$, nous pourrions obtenir un modèle normal à un degré à partir de (1) en fixant $\theta_k = \mu_k$ et $\gamma_k^2 = 0, k = 1, 2$.

Enfin, nous supposerons que tous les vecteurs aléatoires (τ_k, λ_k) et (α_k, θ_k), où $\alpha_k = (\alpha_1^{(k)}, \ldots, \alpha_n^{(k)})$, $k = 1, 2$, sont mutuellement indépendants.

Bien que nous ayons défini trois types de loi a priori pour τ_1 et τ_2, elles peuvent être traitées de manière uniforme, parce que les lois marginales a priori de τ_1 et τ_2, obtenues à partir des lois de Poisson-Gamma, sont les lois binomiales négatives :

$$\pi(\tau_1) \propto \frac{\Gamma(\tau_1 + a_1)}{\tau_1!} \left(\frac{N}{N + b_1} \right)^{\tau_1}$$

et

$$\pi(\tau_2) \propto \frac{\Gamma(\tau_2 + a_2)}{\tau_2!} \left(\frac{1}{1 + b_2} \right)^{\tau_2}, \qquad (2)$$

où $\Gamma(\cdot)$ dénote la fonction Gamma. Les lois de Jeffreys et les lois uniformes sont des cas limites de (2) obtenus en supposant que $a_k = b_k = 0, k = 1, 2$, et $a_k = 1, b_k = 0$, $k = 1, 2$, respectivement. Notons que la loi Gamma n'est pas définie pour ces valeurs de a_k et b_k; toutefois, pour la dérivation des estimateurs, nous pouvons utiliser ces valeurs dans (2).

La loi conjointe a posteriori de τ_1, τ_2, α_1 et α_2 peut être exprimée sous la forme

$$\pi(\tau_1, \tau_2, \alpha_1, \alpha_2 \,|\, \text{données})$$

$$\propto \frac{(N-n)^{\tau_1} \Gamma(\tau_1 + a_1)}{(\tau_1 - m - r_1)!(N + b_1)^{\tau_1}} \prod_{i=1}^{n} \frac{\exp[\alpha_i^{(1)} z_i^{(1)}]}{\left[1 + \exp[\alpha_i^{(1)}]\right]^{\tau_1 - m_i}}$$

$$\times \exp\left[\begin{array}{l} -\dfrac{\sum_{i=1}^{n} (\alpha_i^{(1)} - \bar{\alpha}^{(1)})^2}{2\sigma_1^2} \\ \\ -\dfrac{(\bar{\alpha}^{(1)} - \mu_1)^2}{2v_1} \end{array} \right] \frac{\Gamma(\tau_2 + a_2)}{(\tau_2 - r_2)!(b_2 + 1)^{\tau_2}}$$

$$\times \prod_{i=1}^{n} \frac{\exp[\alpha_i^{(2)} z_i^{(2)}]}{\left[1 + \exp[\alpha_i^{(2)}]\right]^{\tau_2}} \exp\left[\begin{array}{l} -\dfrac{\sum_{i=1}^{n} (\alpha_i^{(2)} - \bar{\alpha}^{(2)})^2}{2\sigma_2^2} \\ \\ -\dfrac{(\bar{\alpha}^{(2)} - \mu_2)^2}{2v_2} \end{array} \right] \qquad (3)$$

où $z_i^{(1)} = z_i^{(0)} + z_i^{(1-0)}$ est la valeur observée de la variable aléatoire $Z_i^{(1)} = Z_i^{(0)} + Z_i^{(1-0)}$ qui indique le nombre de personnes distinctes dans U_1, soit dans S_0 ou dans $U_1 - S_0$, qui sont nommées par A_i; $\bar{\alpha}^{(k)}$ est la moyenne arithmétique des $\alpha_i^{(k)}$; et $v_k = \gamma_k^2 + \sigma_k^2/n, k = 1, 2$.

Puisque nous ne pouvons pas calculer l'intégrale analytique de (3) par rapport à $\alpha_i^{(1)}$ et à $\alpha_i^{(2)}$, nous n'essayerons pas d'obtenir les expressions pour les lois a posteriori de τ_1 et τ_2, mais, comme dans Castledine (1981), nous utiliserons le mode de $\pi(\tau_1, \tau_2, \alpha_1, \alpha_2 | \text{données})$ comme estimateur de $(\tau_1, \tau_2, \alpha_1, \alpha_2)$. En adoptant cette stratégie, nous avons que l'estimateur proposé est la solution du système d'équations :

$$\hat{\tau}_1 = \frac{M + R_1 + (1 - n/N)[N(a_1 - 1)/(N + b_1)] \prod_{i=1}^{n} (1 - \hat{p}_i^{(1)})}{1 - (1 - n/N)[N/(N + b_1)] \prod_{i=1}^{n} (1 - \hat{p}_i^{(1)})};$$

$$\hat{p}_i^{(1)} = \frac{\exp\{\hat{\alpha}_i^{(1)}\}}{1 + \exp\{\hat{\alpha}_i^{(1)}\}} = \frac{Z_i^{(1)}}{\hat{\tau}_1 - M_i} - \frac{\hat{\alpha}_i^{(1)} - \hat{\bar{\alpha}}^{(1)}}{(\hat{\tau}_1 - M_i)\sigma_1^2}$$

$$- \frac{\hat{\bar{\alpha}}^{(1)} - \mu_1}{n(\hat{\tau}_1 - M_i)v_1}; i = 1, \ldots, n; \qquad (4)$$

$$\hat{\tau}_2 = \frac{R_2 + [(a_2 - 1)/(1 + b_2)] \prod_{i=1}^{n} (1 - \hat{p}_i^{(2)})}{1 - [1/(1 + b_2)] \prod_{i=1}^{n} (1 - \hat{p}_i^{(2)})};$$

$$\hat{p}_i^{(2)} = \frac{\exp\{\hat{\alpha}_i^{(2)}\}}{1 + \exp\{\hat{\alpha}_i^{(2)}\}} = \frac{Z_i^{(2)}}{\hat{\tau}_2} - \frac{\hat{\alpha}_i^{(2)} - \hat{\bar{\alpha}}^{(2)}}{\hat{\tau}_2 \sigma_2^2}$$

$$- \frac{\hat{\bar{\alpha}}^{(2)} - \mu_2}{n\hat{\tau}_2 v_2}; i = 1, \ldots, n; \qquad (5)$$

où $\hat{\bar{\alpha}}^{(k)} = \sum_1^n \hat{\alpha}_i^{(k)}/n, k = 1, 2$. Il s'ensuit qu'un estimateur de τ est $\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$.

Les formes de ces estimateurs sont fondamentalement des ajustements des formes de l'EMV proposées par Félix-Medina et Thompson (2004), de sorte qu'est intégrée dans les estimateurs proposés l'information initiale au sujet de τ_k et $\alpha_i^{(k)}, i = 1, \ldots, n;\ k = 1, 2$. En outre, comme la fait remarquer un examinateur, l'estimateur $\hat{p}_i^{(k)}$ a la forme de l'EMV de $p_i^{(k)}$ suivi par des termes de rétrécissement, l'un étant celui de $\alpha_i^{(k)}$ vers la moyenne arithmétique $\hat{\bar{\alpha}}^{(k)}$ et l'autre, celui de $\hat{\bar{\alpha}}^{(k)}$ vers la moyenne a priori μ_k.

4. Intervalles de confiance pour les tailles de population

Comme nous l'avons indiqué plus haut, nous utiliserons l'approche fréquentiste pour obtenir des intervalles de confiance de type fondé sur le plan de sondage qui sont robustes aux écarts par rapport à la loi de Poisson hypothétique des M_i. Nous examinerons des intervalles bootstrap et des intervalles de Wald basés sur une approximation normale (voir Agresti 2002, page 13 et Evans, Kim et O'Brien 1996 pour la terminologie la plus récente).

4.1 Intervalles de confiance bootstrap

Nous utiliserons une version du bootstrap obtenue en combinant la variante du bootstrap pour les populations finies proposée par Booth, Butler et Hall (1994) et la variante paramétrique du bootstrap (voir Davison et Hinkley 1997, chapitre 2).

Les étapes de la méthode que nous proposons sont les suivantes. (i) Construire une population artificielle de N valeurs des m_i en répétant N/n fois, en supposant que N/n est un nombre entier, l'échantillon sélectionné de n tailles de grappe $m_1, \ldots, m_n$. Si $N = kn + r$, où k et r sont des entiers positifs, construire la population en répétant k fois l'échantillon sélectionné de n tailles de grappe et ajouter à cet ensemble de tailles m_i un échantillon aléatoire simple sans remise (EASSR) de r valeurs des m_i sélectionnées à partir de l'échantillon observé de n tailles de grappe. ii) Sélectionner un EASSR de n tailles à partir de la population des m_i. Soit $i_1, \ldots, i_n$ les indices des m_i dans l'échantillon. (iii) Pour chaque $i = i_1, \ldots, i_n$, tirer des échantillons de tailles $\hat{\tau}_1 - m_i$ et $\hat{\tau}_2$ à partir des lois de Bernoulli de moyennes $\hat{p}_i^{(1)}$ et $\hat{p}_i^{(2)}$, respectivement, où $\hat{\tau}_1, \hat{\tau}_2, \hat{p}_i^{(1)}$ et $\hat{p}_i^{(2)}$ sont les estimations de $\tau_1, \tau_2, p_i^{(1)}$ et $p_i^{(2)}$ calculées d'après l'échantillon observé original. Ces échantillons simulent les valeurs des ensembles $\{x_{ij}^{(1)}\}$ et $\{x_{ij}^{(2)}\}$ de variables indicatrices. iv) Calculer les estimations de τ_1, τ_2 et τ à partir des échantillons tirés aux étapes (ii) et (iii) en suivant la même méthode que celle utilisée pour calculer les estimations originales $\hat{\tau}_1, \hat{\tau}_2$ et $\hat{\tau}$. (v) Obtenir les lois bootstrap de $\hat{\tau}_1, \hat{\tau}_2$ et $\hat{\tau}$ en répétant les étapes (i) à (iv) un grand nombre B de fois et en calculant les lois empiriques à partir des ensembles de B valeurs de $\hat{\tau}_1, \hat{\tau}_2$ et $\hat{\tau}$. (vi) Construire les intervalles de confiance bootstrap $100(1-\alpha)\,\%$ pour τ_1, τ_2 et τ en utilisant la méthode de base ou celle des centiles (voir Davison et Hinkley 1997, chapitre 5 pour la description de ces méthodes). Dans la méthode de base, l'intervalle pour τ est $[2\hat{\tau} - \hat{\tau}^{(1-\alpha/2)}, 2\hat{\tau} - \hat{\tau}^{(\alpha/2)}]$, et dans la méthode des centiles, il est $[\hat{\tau}^{(\alpha/2)}, \hat{\tau}^{(1-\alpha/2)}]$, où $\hat{\tau}^{(\alpha/2)}$ et $\hat{\tau}^{(1-\alpha/2)}$ sont les points $\alpha/2$ inférieur et supérieur de la distribution bootstrap de l'estimation originale $\hat{\tau}$ de τ.

Notons que cette variante du bootstrap ne repose pas sur l'utilisation de la loi de Poisson hypothétique des M_i, mais sur le plan d'échantillonnage utilisé pour sélectionner l'échantillon initial de grappes. Donc, nous pouvons considérer que les intervalles de confiance résultants sont robustes aux écarts par rapport à la loi hypothétique des M_i.

Si l'on souhaite également calculer les estimations bootstrap des variances de $\hat{\tau}_1, \hat{\tau}_2$ et $\hat{\tau}$, il est possible d'obtenir des estimations simples en calculant les variances d'échantillon des ensembles de B valeurs de ces estimateurs.

4.2 Intervalles de confiance de Wald

Bien que nous ne démontrions pas théoriquement ici que les estimateurs proposés des tailles de population suivent asymptotiquement une loi normale, nous supposerons que la loi normale est une approximation raisonnable des lois des estimateurs. Donc, nous construisons pour les tailles de population des intervalles de confiance de Wald $100(1-\alpha)\,\%$ de type fondé sur le plan de sondage de la forme $\hat{\tau}_k \pm z_{1-\alpha/2}\sqrt{\hat{\mathbf{V}}(\hat{\tau}_k)}$, où $z_{1-\alpha/2}$ est le point $\alpha/2$ supérieur de la loi normale standard, et $\hat{\mathbf{V}}(\hat{\tau}_k)$ est un estimateur de type fondé sur le plan de sondage de la variance de $\hat{\tau}_k$.

Pour construire ce genre d'intervalle, nous commençons par dériver des estimateurs de la variance de type fondé sur le plan de sondage en suivant la même stratégie que celle utilisée par Félix-Medina et Thompson (2004). Celle-ci consiste à remplacer la distribution des tailles de grappes par celle du plan d'échantillonnage utilisé pour sélectionner l'échantillon initial S_0. Nous utilisons pour cela la formule :

$$\mathbf{V}(\hat{\tau}_k) = \mathbf{V}_p[\mathbf{E}_\xi(\hat{\tau}_k|\mathbf{m}_s)] + \mathbf{E}_p[\mathbf{V}_\xi(\hat{\tau}_k|\mathbf{m}_s)], \qquad (6)$$

où $\mathbf{E}_\xi(\hat{\tau}_k|\mathbf{m}_s)$ et $\mathbf{V}_\xi(\hat{\tau}_k|\mathbf{m}_s)$ dénotent les opérateurs d'espérance et de variance conditionnelles fondées sur le modèle, sachant que $\mathbf{M}_s = \mathbf{m}_s$; et $\mathbf{E}_p(\cdot)$ et $\mathbf{V}_p(\cdot)$ dénotent les opérateurs d'espérance et de variance fondées sur le plan de sondage. Donc, nous obtenons les estimateurs de variance en appliquant (6) aux approximation de Taylor de

premier ordre $\hat{\tau}_1^*$ et $\hat{\tau}_2^*$ de $\hat{\tau}_1$ et $\hat{\tau}_2$, respectivement, autour des espérances fondées sur le modèle de $c_s^{(1)} = (\mathbf{M}_s, \mathbf{Z}_s^{(1)}, R_1)$ et $c_s^{(2)} = (\mathbf{Z}_s^{(2)}, R_2)$, où $\mathbf{Z}_s^{(k)} = (Z_1^{(k)}, \ldots Z_n^{(k)})$, $k = 1, 2$.

En utilisant la stratégie décrite antérieurement et le fait que $Z_i^{(1)}|\mathbf{m}_s \sim \text{bin}(\tau_1 - m_i, p_i^{(1)})$ et $R_1|\mathbf{m}_s \sim \text{bin}(\tau_1 - m, 1 - Q_1)$, où $Q_1 = \prod_{i=1}^n (1 - p_i^{(1)})$, nous avons qu'un estimateur de $\mathbf{V}_p[\mathbf{E}_\xi(\hat{\tau}_1^*|\mathbf{m}_s)]$ est

$$\hat{\mathbf{V}}_{11} = n(1 - n/N)\hat{K}^2 \frac{1}{n-1}\sum_{i=1}^n (m_i - \bar{m})^2, \tag{7}$$

où $\bar{m} = n^{-1}\sum_1^n m_i$; $\hat{K} = -\hat{Q}_1/[\hat{A}_1(\hat{\tau}_1 - m - r_1)]$; $\hat{Q}_1 = \prod_{i=1}^n (1 - \hat{p}_i^{(1)})$;

$$\hat{A}_1 = \sum_{i=1}^n \frac{(\hat{p}_i^{(1)})^2}{\hat{B}_i^{(1)}} - \hat{C}_1 + \frac{1}{\hat{\tau}_1 + a_1 - 1} - \frac{1}{\hat{\tau}_1 - m - r_1};$$

$$\hat{B}_i^{(1)} = (\hat{\tau}_1 - m_i)\hat{p}_i^{(1)}(1 - \hat{p}_i^{(1)}) + \sigma_1^{-2}, \; i = 1, \ldots, n;$$

et

$$\hat{C}_1 = \frac{(v_1^{-1} - n\sigma_1^{-2})\left[n^{-1}\sum_{i=1}^n \hat{p}_i^{(1)}/\hat{B}_i^{(1)}\right]^2}{1 + n^{-1}(v_1^{-1} - n\sigma_1^{-2})n^{-1}\sum_{i=1}^n 1/\hat{B}_i^{(1)}}. \tag{8}$$

En outre, puisque $\mathbf{Cov}(Z_i^{(1)}, R_1|\mathbf{m}_s) = (\tau_1 - m)Q_1 p_i^{(1)}$, un estimateur de $\mathbf{E}_p[\mathbf{V}_\xi(\hat{\tau}_1^*|\mathbf{m}_s)]$ a la forme

$$\hat{\mathbf{V}}_{12} = \hat{A}_1^{-2}\left\{\begin{array}{l} \displaystyle\sum_{i=1}^n \left(\frac{\hat{p}_i^{(1)} - \hat{D}_1}{\hat{B}_i^{(1)}}\right)^2 (\hat{\tau}_1 - m_i)\hat{p}_i^{(1)}(1 - \hat{p}_i^{(1)}) \\ \displaystyle + \frac{(\hat{\tau}_1 - m)\hat{Q}_1(1 - \hat{Q}_1)}{(\hat{\tau}_1 - m - r_1)^2} \\ \displaystyle - \frac{2(\hat{\tau}_1 - m)\hat{Q}_1}{\hat{\tau}_1 - m - r_1}\sum_{i=1}^n \left(\frac{\hat{p}_i^{(1)} - \hat{D}_1}{\hat{B}_i^{(1)}}\right)\hat{p}_i^{(1)} \end{array}\right\}, \tag{9}$$

où

$$\hat{D}_1 = \frac{n^{-1}(v_1^{-1} - n\sigma_1^{-2})n^{-1}\sum_{i=1}^n \hat{p}_i^{(1)}/\hat{B}_i^{(1)}}{1 + n^{-1}(v_1^{-1} - n\sigma_1^{-2})n^{-1}\sum_{i=1}^n 1/\hat{B}_i^{(1)}}.$$

Par conséquent, un estimateur de type fondé sur le plan de sondage de $\mathbf{V}(\hat{\tau}_1)$ est $\hat{\mathbf{V}}(\hat{\tau}_1) = \hat{\mathbf{V}}_{11} + \hat{\mathbf{V}}_{12}$.

Dans le cas de $\hat{\tau}_2^*$, puisque $Z_i^{(2)}|\mathbf{m}_s \sim \text{bin}(\tau_2, p_i^{(2)})$ et $R_2|\mathbf{m}_s \sim \text{bin}(\tau_2, 1 - Q_2)$, où $Q_2 = \prod_{i=1}^n (1 - p_i^{(2)})$, il s'ensuit que $\mathbf{E}_\xi(\hat{\tau}_2^*|\mathbf{m}_s)$ ne dépend pas de $\mathbf{m}_s$, et conséquemment que $\mathbf{V}_p[\mathbf{E}_\xi(\hat{\tau}_2^*|\mathbf{m}_s)] \approx 0$. Donc, puisque $\mathbf{Cov}(Z_i^{(2)}, R_2|\mathbf{m}_s) = \tau_2 Q_2 p_i^{(2)}$, un estimateur de $\mathbf{V}(\hat{\tau}_2)$ est

$$\hat{\mathbf{V}}(\hat{\tau}_2) = \hat{A}_2^{-2}\left\{\begin{array}{l} \displaystyle\sum_{i=1}^n \left(\frac{\hat{p}_i^{(2)} - \hat{D}_2}{\hat{B}_i^{(2)}}\right)^2 \hat{\tau}_2 \hat{p}_i^{(2)}(1 - \hat{p}_i^{(2)}) \\ \displaystyle + \frac{\hat{\tau}_2\hat{Q}_2(1 - \hat{Q}_2)}{(\hat{\tau}_2 - r_2)^2} \\ \displaystyle - \frac{2\hat{\tau}_2\hat{Q}_2}{\hat{\tau}_2 - r_2}\sum_{i=1}^n \left(\frac{\hat{p}_i^{(2)} - \hat{D}_2}{\hat{B}_i^{(2)}}\right)\hat{p}_i^{(2)} \end{array}\right\} \tag{10}$$

où $\hat{Q}_2 = \prod_{i=1}^n (1 - \hat{p}_i^{(2)})$,

$$\hat{A}_2 = \sum_{i=1}^n \frac{(\hat{p}_i^{(2)})^2}{\hat{B}_i^{(2)}} - \hat{C}_2 + \frac{1}{\hat{\tau}_2 + a_2 - 1} - \frac{1}{\hat{\tau}_2 - r_2},$$

$$\hat{B}_i^{(2)} = \hat{\tau}_2 \hat{p}_i^{(2)}(1 - \hat{p}_i^{(2)}) + \sigma_2^{-2}, \quad i = 1, \ldots, n,$$

$$\hat{C}_2 = \frac{(v_2^{-1} - n\sigma_2^{-2})\left[n^{-1}\sum_{i=1}^n \hat{p}_i^{(2)}/\hat{B}_i^{(2)}\right]^2}{1 + n^{-1}(v_2^{-1} - n\sigma_2^{-2})n^{-1}\sum_{i=1}^n 1/\hat{B}_i^{(2)}},$$

et

$$\hat{D}_2 = \frac{n^{-1}(v_2^{-1} - n\sigma_2^{-2})n^{-1}\sum_{i=1}^n \hat{p}_i^{(2)}/\hat{B}_i^{(2)}}{1 + n^{-1}(v_2^{-1} - n\sigma_2^{-2})n^{-1}\sum_{i=1}^n 1/\hat{B}_i^{(2)}}.$$

Enfin, puisque la non-dépendance de $\mathbf{E}_\xi(\hat{\tau}_2^*|\mathbf{m}_s)$ par rapport à $\mathbf{m}_s$ implique que $\mathbf{Cov}(\hat{\tau}_1^*, \hat{\tau}_2^*) \approx 0$, il s'ensuit qu'un estimateur de la variance de $\hat{\tau}$ est $\hat{\mathbf{V}}(\hat{\tau}) = \hat{\mathbf{V}}(\hat{\tau}_1) + \hat{\mathbf{V}}(\hat{\tau}_2)$.

5. Étude de Monte Carlo

Nous avons considéré quatre populations, décrites chacune au tableau 1. Dans la paire formée par les populations I et II, la base de sondage couvrait environ 45 % de la population, tandis que dans la paire formée par les populations III et IV, elle couvrait environ 70 % de la population. Les populations de chaque paire étaient fort semblables, sauf le fait que, dans l'une des populations de chaque paire, la loi des M_i était une loi de Poisson, tandis que dans l'autre il s'agissait d'une loi binomiale négative. Les probabilités de nomination $p_i^{(k)}, i = 1, \ldots, N$, $k = 1, 2$, ont été générées en utilisant le modèle $p_i^{(k)} = 1 - \exp(-\beta_k m_i)$, où les valeurs de β_k étaient fixées de façon que les valeurs suivantes de $\bar{p}^{(k)} = \sum_1^N p_i^{(k)}/N$ soient obtenues. Pour les populations I et II : $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0,05, 0,01)$ et $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0,01, 0,002)$. Pour les populations III et IV : $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0,05, 0,03)$ et $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0,01, 0,006)$. Le modèle employé pour générer les $p_i^{(k)}$ est un modèle utilisé dans les méthodes prises-effort (voir Seber 1982, chapitre 7 pour une description de ces méthodes). Comme l'a souligné un

rédacteur associé, ce modèle implique que le nombre de personnes nommées par la grappe A_i est l'espérance $(\tau_1 - m_i)\ (1-\exp(-\beta_1 m_i)) + \tau_2(1-\exp(-\beta_2 m_i))$ et que, par conséquent, le nombre de personnes nommées est approximativement proportionnel à m_i. Notons que le modèle échangeable hypothétique pour $p_i^{(k)}$ ne postule pas ce genre de relation avec m_i. Puisque l'estimation de $p_i^{(k)}$ dépend principalement de $z_i^{(k)}$, le nombre de personnes dans U_k nommées par la grappe A_i, nous nous attendons à ce que l'omission de cette relation n'ait pas d'incidence sur l'efficacité de l'estimateur de $p_i^{(k)}$. Darroch (1958) a montré, dans le cas de l'estimation du maximum de vraisemblance, que l'on ne réalise aucun gain significatif en émettant l'hypothèse d'un modèle prises-effort.

Tableau 1
Paramètres des populations simulées

Population I	Population II	Population III	Population IV
$N = 250$	$N = 250$	$N = 250$	$N = 250$
M_i Poisson	M_i Binomiale nég.	M_i Poisson	M_i Binomiale nég.
$\mathbf{E}(M_i) = 7,2$	$\mathbf{E}(M_i) = 7,2$	$\mathbf{E}(M_i) = 7,2$	$\mathbf{E}(M_i) = 7,2$
$\mathbf{V}(M_i) = 7,2$	$\mathbf{V}(M_i) = 24,48$	$\mathbf{V}(M_i) = 7,2$	$\mathbf{V}(M_i) = 24,48$
$\tau_1 = 1\,811$	$\tau_1 = 1\,872$	$\tau_1 = 1\,811$	$\tau_1 = 1\,872$
$\tau_2 = 2\,200$	$\tau_2 = 2\,200$	$\tau_2 = 700$	$\tau_2 = 700$
$\tau = 4\,011$	$\tau = 4\,072$	$\tau = 2\,511$	$\tau = 2\,572$
$\tau_1/\tau = 0,45$	$\tau_1/\tau = 0,46$	$\tau_1/\tau = 0,72$	$\tau_1/\tau = 0,73$

Pour les populations I et II, les valeurs des paramètres des lois a priori étaient $\sigma_k^2 = 25, \mu_k = -3,5, \gamma_k^2 = 25, k = 1, 2, a_1 = 1, b_1 = 0,1, a_2 = 7,84, b_2 = 0,0028,$ de sorte que $\mathbf{E}(\lambda_1) = 10, \mathbf{V}(\lambda_1) = 100, \mathbf{E}(\lambda_2) = 2\,800,$ et $\mathbf{V}(\lambda_2) = 10^6$. Pour les populations III et IV les valeurs des paramètres étaient $\sigma_k^2 = 9, \mu_k = -3,5, \gamma_k^2 = 9, k = 1,2, a_1 = 1, b_1 = 0,1, a_2 = 8, b_2 = 0,01,$ de sorte que $\mathbf{E}(\lambda_1) = 10, \mathbf{V}(\lambda_1) = 100, \mathbf{E}(\lambda_2) = 800$ et $\mathbf{V}(\lambda_2) = 80\,000$. Ces valeurs impliquent que les lois a priori sont bien dispersées sur les intervalles relativement grands qui contiennent les paramètres d'intérêt.

Nous avons réalisé l'expérience par simulation comme il suit. À partir de chaque population de $N = 250$ valeurs de m_i, nous avons sélectionné un EASSR de $n = 25$ valeurs. À partir de la grappe A_i dans l'échantillon, nous avons généré les valeurs de $X_{ij}^{(1)}$ et $X_{ij}^{(2)}$ en tirant des échantillons de taille $\tau_1 - m_i$ et τ_2 à partir de lois de Bernoulli de moyenne $p_i^{(1)}$ et $p_i^{(2)}$, respectivement. Ces données ont été utilisées pour calculer les estimateurs suivants des tailles de population : l'ensemble d'EMV $\tilde{\tau}_1, \tilde{\tau}_2,$ et $\tilde{\tau} = \tilde{\tau}_1 + \tilde{\tau}_2$ proposé par Félix-Medina et Thompson (2004); ainsi que les trois ensembles d'estimateurs bayésiens $\hat{\tau}_1^a, \hat{\tau}_2^a$ et $\hat{\tau}^a = \hat{\tau}_1^a + \hat{\tau}_2^a,\ a = U, J, P,$ obtenus en utilisant comme lois a priori les lois uniforme (U), de Jeffreys (J) et de Poisson (P) respectivement. En outre, nous avons calculé les estimateurs de variance et les intervalles de confiance. Nous avons calculé les intervalles bootstrap selon la méthode de base, sauf les intervalles fondés sur les estimateurs $\hat{\tau}_1^P, \hat{\tau}_2^P$ et $\hat{\tau}^P$, qui ont été calculés par la méthode des centiles. Tous les estimateurs bootstrap ont été obtenus en utilisant 2 000 échantillons bootstrap. Enfin, les propriétés des estimateurs ponctuels et d'intervalle ont été évaluées en utilisant $r = 10\,000$ essais de la méthode qui précède.

Nous avons évalué les propriétés d'un estimateur, disons $\hat{\tau}$, par son biais relatif et la racine carrée de son erreur quadratique moyenne relative, définis comme étant $\text{r} - \text{biais} = \sum_1^r (\hat{\tau}_i - \tau)/(r\tau)$ et $\sqrt{\text{r} - \text{eqm}} = \sqrt{\sum_1^r (\hat{\tau}_i - \tau)^2/(r\tau^2)}$, où $\hat{\tau}_i$ est la valeur de $\hat{\tau}$ obtenue lors du i^{e} essai. Nous avons également évalué les propriétés d'un estimateur de variance par son biais relatif et la racine carrée de son erreur quadratique moyenne relative, qui ont été définis de la même façon que ceux d'un estimateur de la taille de population, mais en utilisant la variance déterminée empiriquement plutôt que la variance réelle. Enfin, nous avons évalué les propriétés des intervalles de confiance à 95 % par leur probabilité de couverture et leur longueur moyenne.

6. Résultats et discussion

Faute d'espace, aux tableaux 2 à 4, nous présentons que certains résultats de l'étude numérique. Toutefois, les commentaires qui suivent ont trait à l'ensemble complet de résultats.

Malgré les limites de l'étude par simulation, nous pouvons conclure que le principal facteur qui influe sur les propriétés des estimateurs et des intervalles de confiance et la grandeur des $p_i^{(k)}$. Lorsque celles-ci sont grandes et indépendamment de la loi des M_i et de la taille de la fraction τ_1/τ couverte par la base de sondage, chacun des estimateurs des τ et des intervalles de confiance de type fondé sur le plan de sondage (Wald ou bootstrap) donne de bons résultats. Toutefois, lorsque les $p_i^{(k)}$ sont faibles et malgré tous les autres facteurs, seuls les estimateurs bayésiens $\hat{\tau}_k^P$ donnent des résultats acceptables. Il mérite d'être souligné que, si les $p_i^{(k)}$ sont faibles, les estimateurs bayésiens $\hat{\tau}_k^U$ et $\hat{\tau}_k^J$ donnent de meilleurs résultats que l'EMV $\tilde{\tau}_k$; toutefois, les propriétés de $\hat{\tau}_k^U$ et $\hat{\tau}_k^J$ ne sont pas suffisamment bonnes pour rendre les inférences fiables.

Les intervalles de confiance bootstrap pour τ_1 basés sur $\hat{\tau}_1^P$ ne sont pas aussi bons que les intervalles de Wald lorsque les $p_i^{(k)}$ sont faibles ou que les M_i ne suivent pas une loi de Poisson. L'explication de ce résultat et l'élaboration de meilleurs intervalles bootstrap sont des sujets qui devront être étudiés de façon plus approfondie.

Tableau 2
Biais relatif et racine carrée de l'erreur quadratique moyenne relative des estimateurs des tailles de population

	Population I				Population II				Population III				Population IV			
$\bar{p}_1$	0,05		0,01		0,05		0,01		0,05		0,01		0,05		0,01	
$\bar{p}_2$	0,01		0,002		0,01		0,002		0,03		0,006		0,03		0,006	
	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$
$\tilde{\tau}_1$	−0,00	0,02	−0,00	0,06	−0,00	0,02	−0,01	0,09	−0,00	0,02	−0,00	0,06	−0,00	0,02	−0,01	0,09
$\tilde{\tau}_2$	0,01	0,12	0,24[a]	0,78[a]	0,01	0,13	0,21[a]	0,76[a]	0,00	0,06	0,17[b]	0,67[b]	0,00	0,06	0,16[c]	0,63[c]
$\tilde{\tau}$	0,01	0,07	0,13[a]	0,43[a]	0,01	0,07	0,12[a]	0,42[a]	0,00	0,02	0,05[b]	0,19[b]	−0,00	0,02	0,04[c]	0,18[c]
$\hat{\tau}_1^U$	−0,00	0,02	−0,00	0,06	−0,00	0,02	−0,01	0,09	−0,00	0,02	−0,00	0,06	−0,00	0,02	−0,01	0,09
$\hat{\tau}_2^U$	0,02	0,13	0,14[a]	0,65[a]	0,01	0,12	0,14[a]	0,65[a]	0,00	0,06	0,13	0,65	0,00	0,06	0,13	0,71
$\hat{\tau}^U$	0,01	0,07	0,08[a]	0,36[a]	0,01	0,07	0,08[a]	0,36[a]	0,00	0,02	0,03	0,19	−0,00	0,02	0,03	0,20
$\hat{\tau}_1^J$	−0,00	0,02	−0,01	0,06	−0,00	0,02	−0,01	0,09	−0,00	0,02	−0,01	0,06	−0,00	0,02	−0,01	0,09
$\hat{\tau}_2^J$	−0,00	0,12	−0,14	0,48	−0,00	0,12	−0,14	0,48	−0,00	0,06	−0,04	0,37	−0,00	0,06	−0,04	0,35
$\hat{\tau}^J$	−0,00	0,07	−0,08	0,27	−0,00	0,07	−0,08	0,27	−0,00	0,02	−0,02	0,11	−0,00	0,02	−0,02	0,12
$\hat{\tau}_1^P$	−0,00	0,02	−0,01	0,06	−0,00	0,02	−0,01	0,09	−0,00	0,02	−0,01	0,06	−0,00	0,02	−0,01	0,09
$\hat{\tau}_2^P$	0,02	0,12	0,07	0,20	0,02	0,11	0,07	0,20	0,00	0,06	0,00	0,18	0,00	0,06	0,01	0,18
$\hat{\tau}^P$	0,01	0,06	0,04	0,11	0,01	0,06	0,03	0,11	0,00	0,02	−0,00	0,07	−0,00	0,02	−0,00	0,08

Nota : $r\beta$, biais relatif; $r\varepsilon^2$, erreur quadratique moyenne relative; $\tilde{\tau}_1, \tilde{\tau}_2$ et $\tilde{\tau}$, EMV. Les indices supérieurs *U*, *J* et *P* des estimateurs $\hat{\tau}_1, \hat{\tau}_2$ et $\hat{\tau}$ indiquent des estimateurs bayésiens basés sur une loi uniforme, de Jeffreys et de Poisson-Gamma à deux degrés, respectivement. Les résultats sont fondés sur 10^4 essais. Les indices supérieurs *a*, *b* et *c* indiquent des résultats obtenus en ne tenant pas compte de 8 %, de 15 % et de 21 % des essais. Les essais omis étaient ceux pour lesquels l'estimateur correspondant de τ_2 était négatif ou supérieur à 10^4.

Tableau 3
Probabilité de couverture et longueur moyenne des intervalles de confiance à 95 %

	Population I								Population II							
	$\bar{p}_1 \approx 0{,}05, \bar{p}_2 \approx 0{,}01$				$\bar{p}_1 \approx 0{,}01, \bar{p}_2 \approx 0{,}002$				$\bar{p}_1 \approx 0{,}05, \bar{p}_2 \approx 0{,}01$				$\bar{p}_1 \approx 0{,}01, \bar{p}_2 \approx 0{,}002$			
	Bootstrap		Wald		Bootstrap		Wald		Bootstrap		Wald		Bootstrap		Wald	
	cp	$\bar{l}$	cp	$\bar{l}$	cp	$\bar{l}$	cp	$\bar{l}$	cp	$\bar{l}$	cp	$\bar{l}$	cp	$\bar{l}$	cp	$\bar{l}$
$\tilde{\tau}_1^M$	NC	NC	0,95	129	NC	NC	0,94	398	NC	NC	0,93	127	NC	NC	0,76	400
$\tilde{\tau}_2^M$	NC	NC	0,95	1 044	NC	NC	0,90[a]	8 181[a]	NC	NC	0,95	1 029	NC	NC	0,90[a]	7 764[a]
$\tilde{\tau}^M$	NC	NC	0,95	1 052	NC	NC	0,90[a]	8 200[a]	NC	NC	0,95	1 037	NC	NC	0,90[a]	7 784[a]
$\tilde{\tau}_1^D$	0,95	130	0,95	129	0,92	399	0,94	404	0,97	147	0,95	137	0,96	642	0,92	657
$\tilde{\tau}_2^D$	0,94	1 110	0,95	1 044	0,74	L_1	0,90[a]	8 181[a]	0,94	1 129	0,95	1 029	0,74	L_1	0,90[a]	7 764[a]
$\tilde{\tau}^D$	0,94	1 118	0,95	1 052	0,75	L_1	0,90[a]	8 201[a]	0,95	1 139	0,95	1 038	0,78	L_1	0,90[a]	7 819[a]
$\hat{\tau}_1^U$	0,94	131	0,95	129	0,92	412	0,94	403	0,97	150	0,94	137	0,97	668	0,93	657
$\hat{\tau}_2^U$	0,94	1 116	0,95	1 049	0,72	L_2	0,89[a]	6 887[a]	0,94	1 128	0,95	1 028	0,73	L_2	0,89[a]	6 738[a]
$\hat{\tau}^U$	0,94	1 124	0,95	1 057	0,73	L_2	0,90[a]	6 908[a]	0,95	1 139	0,95	1 038	0,77	L_2	0,90[a]	6 796[a]
$\hat{\tau}_1^J$	0,95	131	0,95	128	0,93	412	0,94	402	0,96	151	0,95	137	0,96	666	0,92	652
$\hat{\tau}_2^J$	0,93	1 043	0,94	998	0,58	3 122	0,71	3 142	0,93	1 057	0,93	985	0,60	3 074	0,72	3 095
$\hat{\tau}^J$	0,93	1 052	0,94	1 007	0,60	3 199	0,72	3 178	0,94	1 072	0,93	995	0,68	3 276	0,73	3 188
$\hat{\tau}_1^P$	0,94	131	0,95	129	0,91	411	0,94	402	0,89	151	0,95	137	0,86	666	0,93	654
$\hat{\tau}_2^P$	0,97	997	0,95	957	1,00	1 506	0,92	1 573	0,97	1 000	0,95	943	1,00	1 510	0,92	1 577
$\hat{\tau}^P$	0,97	1 006	0,95	966	1,00	1 575	0,94	1 624	0,97	1 011	0,95	953	1,00	1 679	0,95	1 710

Nota : cp, probabilité de couverture; $\bar{l}$, longueur moyenne. Les indices supérieurs *M* et *D* des EMV $\tilde{\tau}_1, \tilde{\tau}_2$ et $\tilde{\tau}$ indiquent des intervalles de confiance fondés sur le modèle et fondés sur le plan, respectivement. Intervalles de confiance bootstrap calculés sur 2 000 échantillons bootstrap. NC, non calculé. Résultats fondés sur 10^4 essais. L'indice supérieur *a* indique des résultats obtenus en ne tenant pas compte de 8 % des essais. Les essais omis sont ceux pour lesquels l'estimateur correspondant de τ_2 était négatif ou supérieur à 10^4. L_1 et L_2 indiquent des longueurs supérieures à 10^9 et 10^4, respectivement.

Tableau 4
Biais relatif et racine carrée de l'erreur quadratique moyenne relative des estimateurs de variance

	Population I								Population II							
	$\bar{p}_1 \approx 0,05$, $\bar{p}_2 \approx 0,01$				$\bar{p}_1 \approx 0,01$, $\bar{p}_2 \approx 0,002$				$\bar{p}_1 \approx 0,05$, $\bar{p}_2 \approx 0,01$				$\bar{p}_1 \approx 0,01$, $\bar{p}_2 \approx 0,002$			
	Bootstrap		Taylor		Bootstrap		Taylor		Bootstrap		Taylor		Bootstrap		Taylor	
	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$	$r\beta$	$\sqrt{r\varepsilon^2}$
$\tilde{\tau}_1^M$	NC	NC	0,01	0,17	NC	NC	−0,04	0,08	NC	NC	−0,20	0,31	NC	NC	−0,64	0,65
$\tilde{\tau}_2^M$	NC	NC	0,01	0,49	NC	NC	1,9[a]	5,3[a]	NC	NC	−0,02	0,64	NC	NC	1,8[a]	5,4[a]
$\tilde{\tau}^M$	NC	NC	0,01	0,48	NC	NC	1,9[a]	5,3[a]	NC	NC	−0,02	0,64	NC	NC	1,7[a]	5,3[a]
$\tilde{\tau}_1^D$	0,03	0,19	0,01	0,17	−0,02	0,17	−0,00	0,17	0,08	0,46	−0,07	0,28	−0,05	0,40	−0,01	0,37
$\tilde{\tau}_2^D$	0,16	0,62	0,01	0,49	L_1	L_2	1,9[a]	5,3[a]	0,20	1,10	−0,02	0,64	L_2	L_2	1,7[a]	5,3[a]
$\tilde{\tau}^D$	0,15	0,61	0,01	0,48	L_1	L_2	1,9[a]	5,3[a]	0,20	1,10	−0,02	0,64	L_2	L_2	1,7[a]	5,3[a]
$\hat{\tau}_1^U$	0,02	0,20	−0,01	0,17	0,03	0,19	−0,01	0,17	0,14	0,51	−0,06	0,28	0,05	0,37	0,01	0,37
$\hat{\tau}_2^U$	0,13	0,62	−0,01	0,49	0,24	1,20	1,7[a]	4,6[a]	0,22	0,92	−0,00	0,62	0,30	1,40	1,6[a]	6,4[a]
$\hat{\tau}^U$	0,13	0,61	−0,01	0,48	0,24	1,20	1,6[a]	4,5[a]	0,23	0,91	0,01	0,61	0,30	1,40	1,6[a]	6,2[a]
$\hat{\tau}_1^J$	0,06	0,21	0,02	0,17	0,05	0,19	−0,01	0,17	0,12	0,50	−0,08	0,28	0,00	0,35	−0,04	0,36
$\hat{\tau}_2^J$	0,07	0,51	−0,03	0,44	−0,25	0,66	−0,11	1,40	0,13	0,69	−0,03	0,55	−0,25	0,74	−0,13	1,50
$\hat{\tau}^J$	0,06	0,50	−0,03	0,43	−0,25	0,66	−0,12	1,40	0,12	0,68	−0,03	0,53	−0,24	0,72	−0,15	1,40
$\hat{\tau}_1^P$	0,03	0,20	−0,01	0,17	0,03	0,18	−0,02	0,17	0,16	0,52	−0,05	0,28	0,05	0,37	0,01	0,37
$\hat{\tau}_2^P$	0,07	0,34	−0,02	0,35	−0,07	0,16	−0,03	0,12	0,10	0,42	−0,01	0,41	−0,06	0,17	−0,01	0,16
$\hat{\tau}^P$	0,06	0,34	−0,02	0,34	−0,05	0,14	−0,02	0,11	0,10	0,42	−0,01	0,41	−0,03	0,15	0,01	0,16

Nota : $r\beta$, biais relatif; $r\varepsilon^2$, erreur quadratique moyenne relative. Les indices supérieurs M et D des EMV $\tilde{\tau}_1, \tilde{\tau}_2$ et $\tilde{\tau}$ indiquent des estimateurs de variance fondés sur le modèle et fondés sur le plan, respectivement. Intervalles de confiance bootstrap calculés sur 2 000 échantillons bootstrap. NC, non calculé. Résultats basés sur 10^4 essais. L'indice supérieur *a* indique des résultats obtenus en ne tenant pas compte de 8 % des essais. Les essais omis étaient ceux pour lesquels l'estimateur correspondant de τ_2 était négatif ou supérieur à 10^4. L_1 et L_2 indiquent des valeurs supérieures à 10^2 et 10^4, respectivement.

Enfin, les meilleures propriétés de l'ensemble d'estimateurs $\hat{\tau}_k^P$ sont une conséquence de la plus grande quantité d'information qu'ils utilisent. Bien que nous nous soyons servis de lois a priori relativement uniformes pour les τ_k, l'information qu'ils fournissent est suffisante pour éviter les problèmes de biais et de forte variabilité observés pour les autres estimateurs. Nous avons réalisé certains essais par simulations supplémentaires et les résultats (qui ne sont pas présentés dans les tableaux) indiquent qu'à condition que les lois a priori soient maintenues relativement uniformes, les estimations ne sont pas affectées par les valeurs de leurs paramètres. Manifestement, une information initiale erronée combinée à de faibles valeurs des $p_i^{(k)}$ aura une incidence sur les estimations. À titre d'exemple, mentionnons une loi a priori de τ_2 dont la densité de probabilité est fortement concentrée autour d'une valeur très éloignée de la valeur réelle de τ_2. Cependant, nous pensons que si le chercheur dispose d'information correcte, même si elle est vague, il vaut la peine d'utiliser l'ensemble d'estimateurs $\hat{\tau}_k^P$.

Remerciements

Cette étude a été financée par la subvention UASIN-EXB-01-01 du PROMEP et par la subvention PAFI-UAS-2002-I-MHFM-0 de l'UAS. Nous remercions Eduardo Gutierrez, le rédacteur associé et les examinateurs de leurs suggestions et commentaires constructifs.

Bibliographie

Agresti, A. (2002). *Categorical Data Analysis.* 2ième édition. New York : John Wiley & Sons, Inc.

Booth, J.G., Butler, R.W. et Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89, 1282-1289.

Castledine, B.J. (1981). A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika*, 67, 197-210.

Darroch, J.N. (1958). The multiple-recapture census I: Estimation of a closed population. *Biometrika*, 45, 343-359.

Davison, A.C., et Hinkley, D.V. (1997). *Bootstrap Methods and their Applications*. New York : Cambridge University Press.

Evans, M.A., Kim, H.-M. et O'Brien, T.E. (1996). An application of profile-likelihood based confidence interval to capture-recapture estimators. *Journal of Agricultural, Biological, and Environmental Statistics*, 1, 131-140.

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

Fienberg, S.E., Johnson, M.S. et Junker, B.W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society*, A, 162, 383-405.

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

Heckathorn, D.D. (1994). Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49, 11-34.

Seber, G.A.F. (1982). *The Estimation of Animal Abundance*. 2ième édition. London : Griffin.

Snijders, T.A.B. (1992). Estimation on the basis of snowball samples: How to weight? *Bulletin de Méthodologie Sociologique*, 36, 59-70.

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

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

Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 217-239
Statistique Canada, N° 12-001 au catalogue

Plans de sondage pour les indices des prix à la consommation

Alan H. Dorfman, Janice Lent, Sylvia G. Leaver et Edward Wegman [1]

Résumé

L'échantillonnage en vue d'estimer un indice des prix à la consommation (IPC) est assez compliqué et requiert généralement la combinaison de données provenant d'au moins deux enquêtes, l'une donnant les prix et l'autre, la pondération par les dépenses. Deux approches fondamentalement différentes du processus d'échantillonnage – l'échantillonnage probabiliste et l'échantillonnage par choix raisonné – ont été vivement recommandées et sont utilisées par divers pays en vue de recueillir les données sur les prix. En construisant un petit « univers » d'achats et de prix à partir de données scannées sur les céréales, puis en simulant diverses méthodes d'échantillonnage et d'estimation, nous comparons les résultats de deux approches du plan de sondage et de l'estimation, à savoir l'approche probabiliste adoptée aux États-Unis et l'approche par choix raisonné adoptée au Royaume-Uni. Pour la même quantité d'information recueillie, mais avec l'utilisation d'estimateurs différents, les méthodes du Royaume-Uni semblent offrir une meilleure exactitude globale du ciblage d'un indice superlatif des prix à la consommation basé sur la population.

Mots clés : Indice élémentaire; échantillonnage avec probabilité proportionnelle à la taille; échantillonnage par choix raisonné; données scannées; indice superlatif.

1. Contexte

Du début à la fin, l'échantillonnage en vue d'établir un indice des prix à la consommation (IPC) représente l'une des entreprises d'échantillonnage les plus compliquées. La population cible est difficile à définir, le domaine approprié des articles est débattu, les définitions des ingrédients bruts, c'est-à-dire les prix, les quantités et les articles, sont ambiguës et mises en doute. L'estimateur ultime – l'estimateur de l'IPC d'ensemble – repose sur des données provenant d'au moins deux enquêtes, l'une donnant les prix et l'autre, les « pondérations ». Sous le niveau des « articles composites » (ou « strates d'articles »), c'est-à-dire des groupes d'articles dont il est supposé que le mouvement des prix est homogène, il n'existe habituellement pas de moyen rentable de tenir à jour les poids d'échantillonnage. Le débat se poursuit donc quant au choix approprié parmi divers estimateurs simples de la variation du prix pour les catégories d'articles, c'est-à-dire les « indices élémentaires ». La méthode appropriée d'agrégation de ces variations des prix, au moyen des pondérations, fait également l'objet de débats.

On distingue deux grandes approches de l'échantillonnage en vue de relever les prix : l'échantillonnage probabiliste et l'échantillonnage par choix raisonné, ou échantillonnage au jugé. L'approche d'échantillonnage la plus généralement reconnue consiste habituellement à injecter un élément aléatoire dans le processus de sondage et à se fonder sur cet élément aléatoire pour faire des inférences au sujet des caractéristiques de population que l'on veut étudier, c'est-à-dire un échantillonnage probabiliste ou « basé sur un plan de sondage »; voir, par exemple, Särndal, Swensson et Wretman (1992). Cette approche n'a pas toujours été tenue pour acquise. Au début du XX^{e} siècle, l'échantillonnage « par choix raisonné » ou « représentatif » était considéré comme une option viable, voire préférable. Plus récemment, l'école de pensée « prédictionniste » de Royall a contesté les hypothèses basées sur le plan de sondage; voir, par exemple, Valliant, Dorfman et Royall (2000).

Aux États-Unis, toutes les enquêtes reliées à l'IPC sont réalisées en utilisant des méthodes d'échantillonnage probabilistes complexes. Ailleurs dans le monde, la plupart des IPC sont construits à partir d'échantillons au jugé, dans lesquels les spécialistes des différentes strates d'articles choisissent des catégories plus ou moins générales ou sélectives d'articles pour lesquelles des agents de terrain relèvent les prix. La raison fondamentale de cette approche est la difficulté qu'il y a à obtenir toutes les données sur la pléthore d'articles vendus et sur les endroits où ils sont vendus nécessaires pour que l'échantillonnage probabiliste soit faisable.

L'aspect intéressant est que fort peu d'évaluations de l'exactitude relative des diverses approches d'échantillonnage ont été faites. En réalité, il n'est pas certain que ce genre d'évaluation soit réalisable. Même pour les pays les plus petits, l'indice des prix sous-jacent calculé en se basant sur la population totale d'articles, ou indice des prix de population, englobe un si grand nombre de transactions portant sur un si grand nombre d'articles dans un si grand nombre d'emplacements qu'il est inaccessible. De surcroît, la population d'articles sur le marché varie constamment, ce qui complique l'application des formules d'indice de population classiques. Alors, comment peut-on juger de la

1. Alan H. Dorfman, Office of Survey Methods Research et Sylvia G. Leaver, Office of Prices and Living Conditions, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE, Washington, D.C., États-Unis, 20212; Janice Lent, U.S. Bureau of Transportation Statistics, 400 7th Street, SW, Washington, D.C., États-Unis, 20590; Edward Wegman, Center for Computational Statistics, George Mason University, Fairfax, VA, États-Unis, 22030.

justesse relative de diverses estimations basées sur un échantillon par rapport à la « valeur réelle »? De surcroît, dans la plupart des cas, on ne dispose même pas d'information échantillonnale pour l'un des éléments clés de l'indice de population, à savoir les *quantités* d'articles vendus, de sorte que même la construction d'une population artificielle à partir de données d'échantillon en vue d'une évaluation n'a pas été possible.

La disponibilité relativement récente de données *scannées*, aux États-Unis et ailleurs, offre une occasion sans précédent d'évaluer les approches d'échantillonnage et les estimateurs. Ces données comprennent les prix *et* les quantités, habituellement sur une base hebdomadaire, de *tous* les articles vendus dans une catégorie donnée dans un grand échantillon de points de vente dotés de scanneurs. Ce genre de données peuvent être utilisées pour construire des populations réalistes de transactions pour lesquelles l'indice des prix réel est *connu*. Nous pouvons alors utiliser diverses méthodes pour échantillonner cette population, construire différentes estimations d'indice d'intérêt et comparer les résultats aux paramètres de population connus. Une étude de ce type, décrite par de Haan, Opperdoes et Schut (1999), semble indiquer que l'« échantillonnage avec seuil d'inclusion » (*cutoff sampling*), c'est-à-dire l'échantillonnage des quelques articles les plus importants (en ce qui a trait aux recettes générées) dans la population, donne de meilleurs résultats que deux grandes approches probabilistes, à savoir l'échantillonnage aléatoire simple (*eas*) et l'échantillonnage avec probabilité proportionnelle à la taille (*ppt*) (où la mesure de taille est, de nouveau, les recettes).

L'une des difficultés que pose toute étude faisant ce genre de comparaison est la nécessité de maintenir des « règles du jeu équitables ». Si l'une des méthodes d'échantillonnage s'appuie, par exemple, sur des données (de population) susceptibles, en réalité, ne pas être disponibles en pratique, tandis qu'une autre ne le fait pas, la comparaison des méthodes est sérieusement minée. De même, si une méthode ne fournit qu'un seul échantillon ou quelques échantillons, et qu'une autre en fournit des milliers, des précautions particulières doivent être prises pour les comparer; en effet, ce genre de comparaison pourrait nécessiter d'importantes restrictions. Étant donné la complexité des méthodes d'échantillonnage et d'estimation utilisées pour le calcul des indices de prix, il n'est pas étonnant que ces difficultés et de nombreuses autres compliquent les expériences conçues en vue de comparer diverses méthodes.

Idéalement, pour comparer les approches, disons, de deux pays, nous imiterions entièrement le processus complexe d'échantillonnage et d'estimation de chacun et nous évaluerions les coûts. Le même budget serait affecté aux deux processus et nous serions capables, au moyen d'une mesure préétablie et équitable, d'évaluer la justesse de chaque estimation par rapport à un indice cible connu.

Le présent article porte sur deux études, une grande étude principale et une étude de suivi secondaire plus petite.

L'étude principale est décrite aux sections 2 à 4. La section 2 donne la construction de la population cible. La section 3 expose les méthodologies des « États-Unis » et du « Royaume-Uni », et fournit les renseignements sur les simulations. Aucun effort n'est fait pour évaluer les coûts relatifs (ce qui nous écarte de la situation idéale), mais les approches concurrentes sont rendues aussi égales que possible en ce qui a trait à l'information utilisée. Les résultats, qui donnent l'avantage à l'approche britannique, sont présentés à la section 4.

L'étude de suivi, décrite à la section 5, a pour but d'essayer de dégager les effets des diverses composantes des deux approches, en particulier la méthode d'échantillonnage et la formule de l'indice élémentaire. La section 6 comprend un résumé final et une discussion.

Note sur les indices cibles. La littérature sur les indices de prix contient des myriades de formules pour calculer la variation des prix d'une période à une autre. Divers indices sont compatibles avec différentes hypothèses quant au comportement d'achat du consommateur « moyen » en réponse à la variation des prix. Les indices à « panier de consommation fixe », les formules fréquemment employées de Laspeyres et celles, moins fréquemment utilisées, de Paasche, sont compatibles avec l'hypothèse selon laquelle les consommateurs continuent d'acheter les mêmes articles en même quantité quelle que soit la variation des prix relatifs. L'indice de Laspeyres projette les quantités de la période 1 (« période de base ») sur la période 2 (« période courante »), tandis que celui de Paasche applique les quantités de la période 2 à la période 1. L'indice géométrique (ou « de Jevons » ou « *moyenne géométrique* »), habituellement pondéré par les parts des dépenses à la période de base, suppose que le consommateur ajuste les quantités qu'il achète de telle façon que la part des dépenses pour chaque article demeure constante au cours du temps. Les formules des indices « superlatifs » de Fisher, de Törnqvist et de Walsh, qui reposent sur des données sur les quantités (ou parts des dépenses) pour les deux périodes, ne nécessitent pas ces hypothèses. Les formules de ces indices, avec les exposants y représentant la période de base, $y+1$, la période courante et i, l'article acheté, sont données à l'annexe A.

Le débat concernant l'indice d'ensemble cible se résume habituellement à choisir entre l'indice de Laspeyres et l'un des indices superlatifs. La plupart des pays optent pour un indice cible de Laspeyres, mais de solides arguments peuvent être présentés (Diewert 1997) en faveur du choix d'un indice superlatif comme cible (habituellement celui de

Fisher ou de Törnqvist), même si la formule de l'estimateur ne ressemble pas à l'une des formules d'indice superlatif de population. Compte tenu de la forme des agrégats élémentaires utilisés aux États-Unis – *moyenne géométrique* – et du fait que des travaux de recherche antérieurs (Dorfman, Leaver et Lent 1999) ont indiqué que le niveau le plus faible d'estimation peut avoir un effet important, la *moyenne géométrique* pondérée sera incluse parmi les indices cibles possibles. Les cibles pour un domaine particulier sont calculées en se basant sur les prix et sur les quantités de tous les articles compris dans le domaine, conformément aux formules qui figurent à l'annexe A (agrégation à un degré des prix et des quantités).

Nota : Ces formules donnent l'illusion d'être simples, mais nécessitent la notation de la section 3 pour leur développement complet. Donc, dans une formule telle que celle de l'indice de Fisher F (que nous choisirons comme cible dans le corps de l'étude aux sections 2 à 4), « i » représente un article i appartenant à une petite catégorie c (un « ANE » (article de niveau d'entrée) ou « article représentatif » – voir la section 3), où c est elle-même un sous-ensemble d'une catégorie plus grande. En outre, l'article i est vendu dans un point de vente particulier j, classifié comme faisant partie d'une chaîne particulière k, et situé dans une région géographique échantillonnée particulière, l'unité primaire d'échantillonnage (*upe*) l. Donc, dans le cas de l'indice de population global, l'expression pour une somme $\sum_i$ est en fait une abréviation de $\sum_{l=1}^{3} \sum_{k=1}^{8} \sum_{j\in(k,l)} \sum_{C=1}^{4} \sum_{h\in C} \sum_{c\in h} \sum_{i\in(j,c)}$; une remarque semblable s'applique à $\prod_i$. Brièvement, il existe des sommes et des produits sur *la totalité* des articles dans la population. Les contractions de ce développement complet donneront les indices de population pour les diverses catégories C, *etc*.

2. La population pour l'étude principale

La source des données sur lesquelles porte l'étude est un ensemble de données scannées sur les céréales pour petit déjeuner couvrant la période de 1995 à 2000 dans trois zones séparées, mais contiguës, d'une grande région métropolitaine. Le U.S. Bureau of Labor Statistics a acheté l'ensemble de données à la société A.C. Nielsen en vue de déterminer s'il est possible d'intégrer des données scannées dans l'IPC des États-Unis; voir Richardson (2000).

Des « populations » artificielles ont été tirées à partir de ces données par la méthode décrite plus loin. Donc, l'étude a pour champ d'observation un univers en apparence limité, celui des céréales, dans un domaine géographique relativement restreint. Toutefois, même cet univers limité permet d'observer des tendances des prix assez divergentes au cours de la période de six ans. Donc, bien que nous ne puissions pas généraliser nos résultats, de façon simple, aux indices de prix globaux couvrant une vaste gamme hétérogène de produits, nous arriverons peut-être à dégager d'importants éclaircissements quant aux effets de diverses méthodes d'échantillonnage et au comportement d'estimateurs particuliers.

L'ensemble de données portant sur une période de six années nous a permis d'établir des tendances des prix d'assez long terme. Afin que le volume de données demeure raisonnable et pour éviter les complications de la saisonnalité, nous nous sommes limités aux données de février. Pour février de l'année y, pour chaque article (c'est-à-dire chaque combinaison particulière k de marque, type, format) dans un point de vente particulier, nous avons combiné les données sur les prix et les quantités pour une période de quatre semaines t en un prix et une quantité uniques pour un mois, en utilisant la somme des quantités $q_k^{\text{Feb},\,y} = \sum_{t\in\text{Feb},\,y} q_k^t$ vendues durant le mois comme valeur de la quantité et la *valeur unitaire* $p_k^{\text{Feb},\,y} = \sum_{t\in\text{Feb},\,y} q_k^t\ p_k^t / \sum_{t\in\text{Feb},\,y} q_k^t$ comme prix. Les valeurs unitaires calculées sur de courtes périodes (par exemple, un mois) correspondant peut-être au sens le plus significatif du prix « moyen » d'un article particulier. L'utilisation des valeurs unitaires lisse les données et les réduit à un volume plus raisonnable; pour une discussion des circonstances sous lesquelles l'utilisation des valeurs unitaires est ou non appropriée, voir Balk (1999).

Pour la présente étude, la population de céréales pour petit déjeuner a été subdivisée en quatre groupes :

1. céréales chaudes (C)
2. céréales « sucrées » (S)
3. céréales « fruitées » (F)
4. céréales « ordinaires » (O), c'est-à-dire les céréales froides ne rentrant pas dans les catégories (2) et (3).

Pour chaque groupe, pour chaque paire d'années successives, nous avons calculé les indices superlatifs et non superlatifs en utilisant les combinaisons article-point de vente disponibles les deux années. En pratique, il est généralement difficile d'obtenir des appariements parfaits de période en période, et il est important de trouver des moyens de faire face à ce problème en découvrant des substituts pour les produits originaux ou d'autres façons; la présente étude ne tient pas compte de ce problème particulier.

Nous avons calculé les indices de long terme (1995 à 2000) directement, ainsi que par enchaînement des indices annuels. En outre, nous avons calculé les indices fondés sur les articles « de base », c'est-à-dire ceux disponibles chacune des six années. D'une année sur l'autre, les articles de base représentaient entre 53 % et 61 % des articles disponibles durant une année particulière pour la comparaison d'une année à l'autre; les dépenses de base représentaient entre 83 % et 91 % des dépenses totales au titre de tous les

articles (de base et autres). Au cours de la période de 1995 à 2000, il y avait 326 articles de base et, en tout, 848 articles distincts.

Les valeurs des indices annuels de population sont représentées aux figures 1 à 5. La figure 1 donne les valeurs de l'indice $\hat{I}^{y,y+1}$ pour les céréales sucrées basé sur l'ensemble des articles vendus dans les magasins durant les années y et $y+1$, pour (février de) $y = 1995, ..., 1999$ (l'indice « d'ensemble »). Les valeurs sont présentées pour cinq indices, y compris l'indice de Paasche P et, en raison de son intérêt théorique, un indice de valeur unitaire U, le ratio des prix moyens pondérés par les quantités, la moyenne étant calculée sur tous les types d'articles et tous les points de vente. La figure 2 donne les résultats des mêmes calculs, mais en se limitant aux articles « de base ». Les figures 1 et 2 sont presque identiques et la ressemblance entre les indices calculés en utilisant tous les articles (indice d'ensemble) et ceux obtenus en utilisant uniquement les articles de base est vérifiée pour les autres catégories de céréales également. Les figures 3 à 5 donnent les résultats relatifs aux indices calculés pour les articles de base dans le cas des céréales chaudes, fruitées et ordinaires. Pour tout indice, les figures révèlent d'importants écarts entre les catégories de céréales. Les tendances des prix des quatre grands groupes sont assez différentes : celle de C est à la hausse, celle de S est fortement à la baise, celle de F est moyennement à la baisse et celle de O est moyennement à la hausse.

Le tableau 1 donne les indices directs de long terme (1995 à 2000) et les indices enchaînés pour « l'ensemble des articles » et pour les « articles de base ». (« L'ensemble des articles » pour la construction d'un indice entre deux années données comprend les articles/points de vente pour lesquels les quantités vendues sont positives les deux années.) De nouveau, l'écart est très faible entre les valeurs obtenues pour les « articles de base » et « l'ensemble des articles », mais prononcé entre les catégories de céréales. Les résultats enchaînés et directs sont proches des indices superlatifs, mais ont tendance à diverger dans le cas de la *moyenne géométrique*, et des indices de Laspeyres et de Paasche. Les indices de valeur unitaire enchaînés et directs sont proches et, en fait, le dernier serait identique à l'indice enchaîné fondé sur les articles de base, si, par souci de commodité, les indices annuels n'avaient pas été basés uniquement sur les combinaisons article-point de vente disponibles pour les deux années.

À part une certaine oscillation de la position de l'indice de valeur unitaire, nous observons un classement manifeste des indices en fonction de la formule, le même pour toutes les catégories de céréales, qui peut se résumer comme suit : 1) Les indices superlatifs diffèrent assez peu les uns des autres, résultat qui mérite d'être souligné étant donné le degré de variabilité due aux « ventes » des prix relatifs et des quantités liés aux combinaisons article-point de vente. 2) Les indices non superlatifs classiques diffèrent fortement les uns des autres et des indices superlatifs, la *moyenne géométrique*, pondérée par les dépenses à la première période, étant beaucoup plus *élevée* que les indices superlatifs, l'indice de Laspeyres étant encore plus élevé et l'indice de Paasche étant (sans surprise) nettement plus faible. Ces résultats donnent à penser que, dans l'univers des céréales, non seulement la quantité, mais aussi la part des dépenses, a tendance à diminuer à la période 2 pour un article dont le prix augmente fortement durant cette période. 3) L'indice de valeur unitaire est faible également, mais, sauf dans le cas des céréales chaudes, s'approche davantage des indices superlatifs que les indices non superlatifs classiques. 4) À la lumière de faits présentés plus loin dans l'article, et à la recommandation d'un examinateur, nous avons inclus deux indices non fondés sur les quantités dans ce tableau (mais pas dans les figures) : l'indice de *Dutot*, qui est un simple ratio des prix moyens (RM) – voir l'annexe A, et une *moyenne géométrique non pondérée* (c'est-à-dire pondérée par une constante); tous deux sont habituellement réservés au calcul des indices au niveau élémentaire. Les résultats sont surprenants : en ce qui a trait à l'approximation des indices superlatifs, ils donnent d'aussi bons, voire de meilleurs, résultats que les indices non superlatifs basés sur les quantités traditionnelles, à peu près à égalité avec l'indice de valeur unitaire.

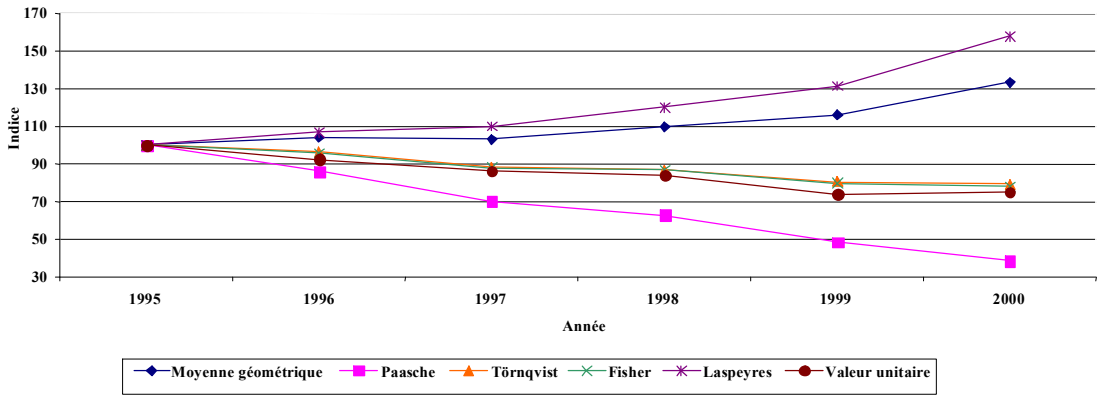

Figure 1. Indices de population cibles enchaînés annuellement pour les indices de février à février pour l'ensemble des céréales sucrées, 1995 = 100.

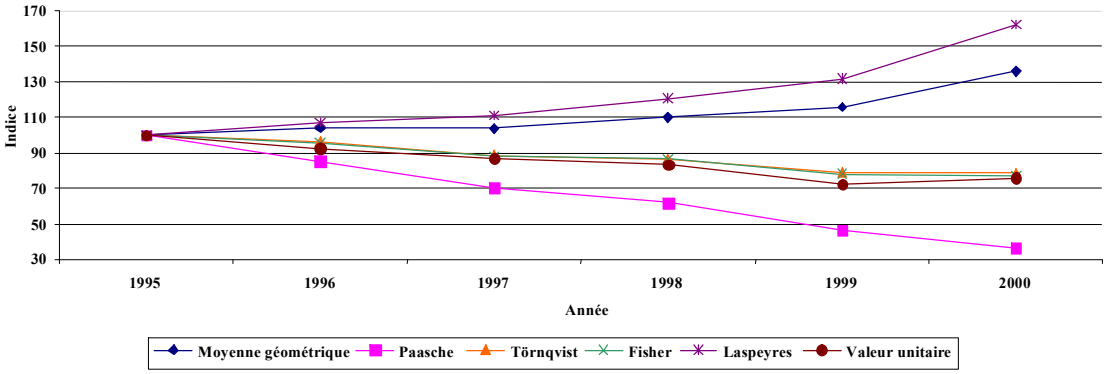

Figure 2. Indices de population cibles enchaînés annuellement pour les indices de février à février pour les céréales sucrées de base, 1995 = 100.

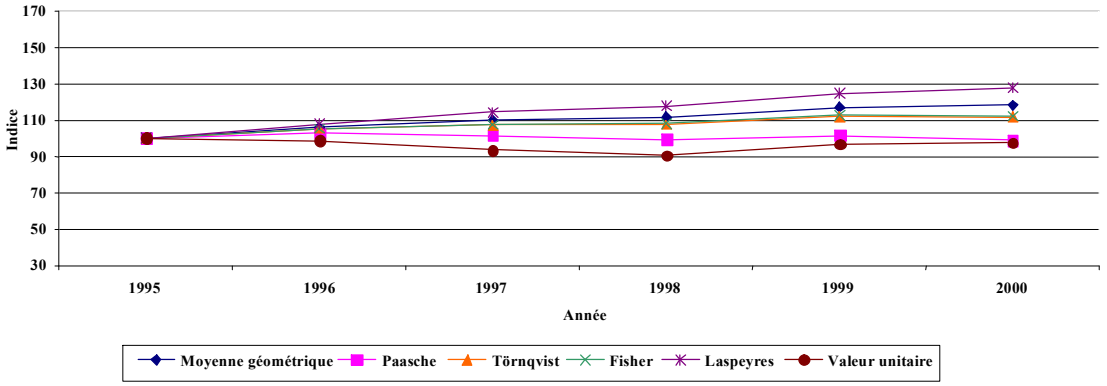

Figure 3. Indices de population cibles enchaînés annuellement pour les indices de février à février pour l'ensemble des céréales chaudes, 1995 = 100.

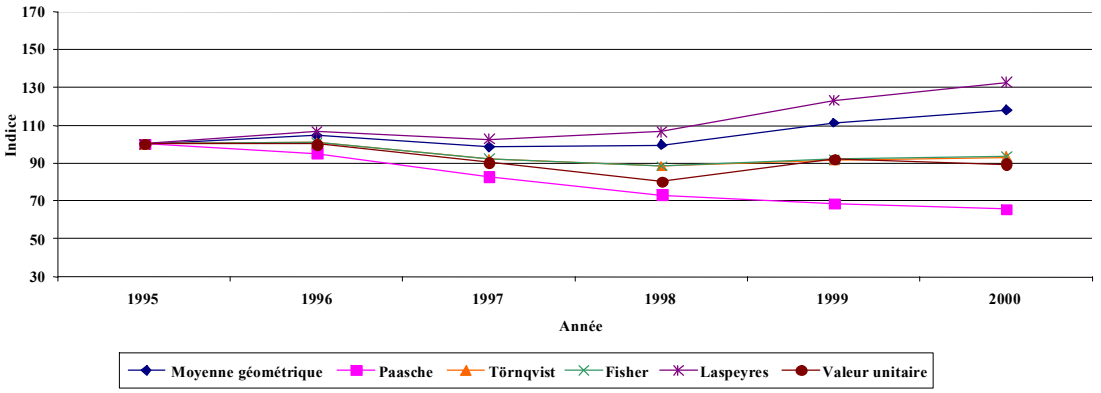

Figure 4. Indices de population cibles enchaînés annuellement pour les indices de février à février pour l'ensemble des céréales fruitées, 1995 = 100.

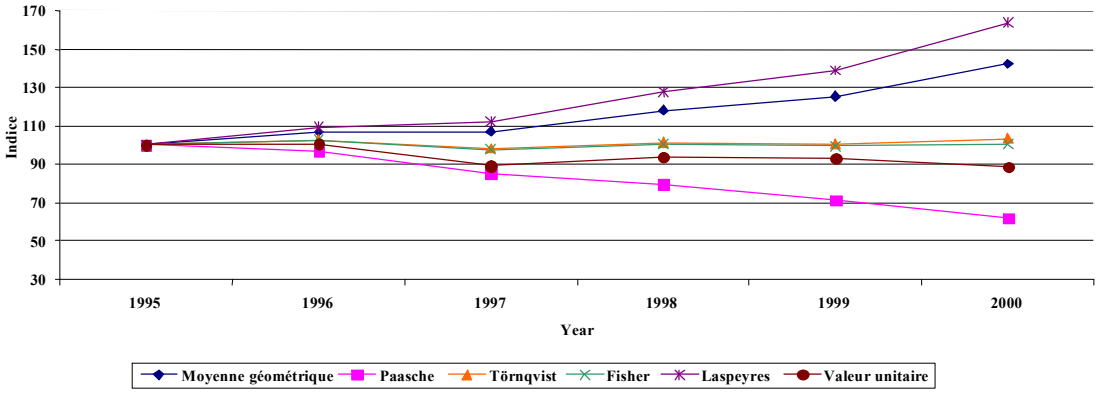

Figure 5. Indices de population cibles enchaînés annuellement pour les indices de février à février pour l'ensemble des céréales ordinaires, 1995 = 100.

Tableau 1
Indices directs et enchaînés pour 1995 à 2000

		Moyenne géométrique	Paasche	Törnqvist	Fisher	Laspeyres	Valeur unitaire
Chaudes	Direct	1,1176	1,0253	1,0847	1,0891	1,1569	0,9576
	Enchaîné, ensemble des articles	1,1801	0,9874	1,1159	1,1216	1,2742	0,9453
	Enchaîné, articles de base	1,1804	0,9865	1,1160	1,1221	1,2763	0,9759
Sucrées	Direct	0,8855	0,6739	0,7913	0,7898	0,9257	0,7417
	Enchaîné, ensemble des articles	1,3341	0,3825	0,7925	0,7771	1,5786	0,7506
	Enchaîné, articles de base	1,3591	0,3661	0,7849	0,7704	1,6212	0,7585
Fruitées	Direct	0,9716	0,8676	0,9319	0,9296	0,9960	0,8932
	Enchaîné, ensemble des articles	1,2202	0,6849	0,9661	0,9696	1,3728	0,9308
	Enchaîné, articles de base	1,1808	0,6557	0,9320	0,9328	1,3269	0,8950
Ordinaires	Direct	1,0811	0,8641	1,0045	0,9816	1,1150	0,8554
	Enchaîné, ensemble des articles	1,3969	0,6330	1,0333	1,0053	1,5965	0,8935
	Enchaîné, articles de base	1,4234	0,6175	1,0353	1,0054	1,6370	0,8879

Tableau 1
Indices directs et enchaînés pour 1995 à 2000

		Moyenne géométrique*	Paasche	Törnqvist	Fisher	Laspeyres	Valeur unitaire	RM	*Moyenne géométrique* +
Chaudes	Direct	1,1176	1,0253	1,0847	1,0891	1,1569	0,9576	1,1192	1,0949
	Enchaîné, ensemble des articles	1,1801	0,9874	1,1159	1,1216	1,2742	0,9453	1,1395	1,1128
	Enchaîné, articles de base	1,1804	0,9865	1,1160	1,1221	1,2763	0,9759	1,1374	1,1151
Sucrées	Direct	0,8855	0,6739	0,7913	0,7898	0,9257	0,7417	0,8817	0,8702
	Enchaîné, ensemble des articles	1,3341	0,3825	0,7925	0,7771	1,5786	0,7506	0,9124	0,9010
	Enchaîné, articles de base	1,3591	0,3661	0,7849	0,7704	1,6212	0,7585	0,8984	0,8894
Fruitées	Direct	0,9716	0,8676	0,9319	0,9296	0,9960	0,8932	0,9815	0,9726
	Enchaîné, ensemble des articles	1,2202	0,6849	0,9661	0,9696	1,3728	0,9308	1,0263	1,0165
	Enchaîné, articles de base	1,1808	0,6557	0,9320	0,9328	1,3269	0,8950	0,9935	0,9820
Ordinaires	Direct	1,0811	0,8641	1,0045	0,9816	1,1150	0,8554	1,0620	1,0511
	Enchaîné, ensemble des articles	1,3969	0,6330	1,0333	1,0053	1,5965	0,8935	1,0642	1,0572
	Enchaîné, articles de base	1,4234	0,6175	1,0353	1,0054	1,6370	0,8879	1,0653	1,0571

* Pondérée par les dépenses à la période de base.
\+ Non pondérée.

D'après les examens préliminaires, et pour simplifier, nous limitons nos investigations plus approfondies aux données sur les articles de base. Afin d'étudier l'exactitude relative des échantillonnages probabiliste et par choix raisonné, tels qu'ils sont appliqués en pratique pour établir les IPC, nous nous sommes efforcés d'approximer les plans d'échantillonnage utilisés aux États-Unis et au Royaume-Uni, qui représentent l'échantillonnage probabiliste et l'échantillonnage par choix raisonné, respectivement. Dans les deux cas, nous avons eu la chance de disposer d'information détaillée sur les processus d'enquête complexes grâce à des manuels et à des contacts avec les organismes respectifs. L'idée fondamentale consistait à procéder à l'échantillonnage répété d'une population donnée, par exemple, les transactions de base effectuées en 1995 et en 1996. Chaque « passage-machine » était une combinaison d'activités d'échantillonnage et d'estimation exécutées conformément aux méthodes d'un pays ou de l'autre. Il ne faut pas oublier que nous voulions comparer les mérites des *méthodologies* et non évaluer le succès avec lequel les États-Unis et le Royaume-Uni estiment les paramètres de leur population cible.

Les quatre groupes de population « naturels » décrits plus haut, qui sont appelés « grands groupes » au Royaume-Uni et « catégories de dépenses » aux États-Unis, ont été divisés en sous-groupes moins agrégés. En pratique, les sous-groupes seraient définis en fonction de types de produits. L'une des raisons, outre tout intérêt intrinsèque que l'on pourrait avoir pour ces produits, est que les sous-groupes ainsi formés ont tendance à être homogènes en ce qui a trait aux tendances des prix. Aux fins de la présente étude par simulation, nous définissons par conséquent les sous-groupes de la façon suivante :

1) Nous avons calculé la variation de prix de long terme pour chacun des 326 articles compris dans les données de base, en utilisant les indices de valeur unitaire pour les articles (sur l'ensemble des points de vente) pour 2000 comparativement à 1995.
2) Nous avons ajouté un bruit à ces indices, trié les articles dans chaque grand groupe en fonction de leur valeur de l'indice perturbé, et regroupé les articles adjacents. Le regroupement des articles dont les indices de long terme étaient proches avait pour but de rendre les sous-groupes homogènes, et l'ajout d'un bruit a été fait de sorte que l'homogénéité soit raisonnablement imparfaite.

Le tableau 2 donne la structure des articles de la population qui a été construite, y compris la nomenclature utilisée dans les deux pays, le nombre de groupes à chaque niveau d'agrégation et le symbole correspondant à chaque niveau de classification utilisé dans le présent article. L'« article représentatif » est le niveau d'agrégation le plus faible auquel un indice est produit aux Royaume-Uni. Il correspond à l'article de niveau d'entrée ou ANE (en anglais *Entry Level Item* ou ELI) des États-Unis, qui est en fait un ensemble d'articles similaires ou connexes. Aux États-Unis, les indices sont produits pour les catégories obtenues au niveau directement supérieur d'agrégation, c'est-à-dire le niveau de la « strate d'articles », mais ces catégories sont encore subdivisées en fonction des régions géographiques dans lesquelles les articles sont vendus. Notons qu'il existe 2 ou 3 strates d'articles/sections h dans une catégorie/un grand groupe C, 3 ANE/articles représentatifs c par strate d'articles/section h (sauf dans un cas où il y en a 2), et 10 ou 11 articles/variétés i dans chaque ANE/article représentatif c. (*Nota* : une catégorie réelle du Royaume-Uni pourrait être plus grande ou plus petite que la catégorie correspondante des États-Unis; par exemple, en règle générale, l'ANE comprend probablement plus de sortes d'articles spécifiques que l'article représentatif. Nous avons donc dû forcer l'équivalence pour assurer que la même quantité d'information soit utilisée dans chaque approche. Cet ajustement n'aura pas d'incidence sur nos conclusions en ce qui concerne les mérites relatifs des méthodes de base utilisées dans les deux pays.

Tableau 2
Structure de population de l'« univers des céréales » : articles

R.-U.	É.-U.	Nombre de groupes	Symbole
Grand groupe	Catégorie de dépenses	4	C
Section	Strate d'articles	10	h
Article représentatif	Article de niveau d'entrée (ANE)	29	c
Variété	Article	326	i

En plus de la structure fondée sur les articles, chaque population de transactions possède une structure « spatiale » caractéristique de l'endroit où a été vendu un article. Cette structure est résumée au tableau 3. Les points de vente appartiennent à des chaînes (par exemple, Safeway, Kroger), qui recoupent les trois unités primaires d'échantillonnage géographique des États-Unis à partir desquelles les données sur les céréales ont été recueillies. (Dans la terminologie du Royaume-Uni, les chaînes sont appelées « *magasins multiples* ».) Les points de vente appartenant à une chaîne donnée ont un propriétaire commun, à l'exception de la « chaîne 8 », qui est un groupe « fourre-tout » composé des points de vente *n'*appartenant *pas* à une grande chaîne (il pourrait y avoir certaines « mini-chaînes »). Lors de l'appariement de cette « structure basée sur les chaînes » à la classification des points de vente utilisée pour l'échantillonnage au Royaume-Uni, la chaîne 8 a été considérée comme un ensemble de « magasins indépendants » (le terme utilisé pour désigner les magasins appartenant à un propriétaire indépendant au Royaume-Uni). La chaîne 4, qui semble présenter la plus grande homogénéité des prix sur l'ensemble des points de vente, a été considérée comme un « magasin multiple pour lequel le relevé des prix est centralisé » (centrally collected multiple), expression utilisée au Royaume-Uni pour les groupes de points de vente dont les prix sont contrôlés centralement. Chaque chaîne restante a été considérée comme une chaîne dont le relevé des prix n'est pas centralisé. Les méthodes de collecte et d'estimation pour ces trois types de chaînes sont données plus loin dans la description des méthodes appliquées au Royaume-Uni.

Donc, la population est constituée de $N^{95} \approx 20\,000$ enregistrements pour les indices de 1995 – 1996, chaque enregistrement représentant l'achat d'un article i dans un point de vente j. Sont reliés à chaque article/point de vente son UPE/région l, sa chaîne/son type de magasin k, le point de vente/magasin j, l'article/la variété i, l'ANE/article représentatif c, la strate d'articles/section h, la catégorie de dépenses/le grand groupe C, et p^y, q^y, p^{y+1}, et q^{y+1}, les prix et quantités (en onces) des articles vendus (en février des) deux années en question. Nous avons utilisé ce fichier de population (appelé simplement « le fichier »

dans la suite) pour simuler toutes les phases des opérations aux États-Unis et au Royaume-Uni.

3. Méthodes d'échantillonnage simulées

Les méthodes d'échantillonnage compliquées que nous avons utilisées pour simuler les approches adoptées aux États-Unis et au Royaume-Uni sont modélisées d'après les pratiques respectives de ces deux pays. Ces pratiques évoluent au cours du temps et présentent même des variantes à un point particulier dans le temps. Notre objectif n'était pas de déterminer quel pays utilise la meilleure méthode, ni d'englober toutes les variantes. Nous cherchions plutôt à comparer deux modes distincts d'échantillonnage, en tenant compte de la gamme de complexités que postulent ces modes. Le lecteur que cela intéresse trouvera une description de la construction de l'IPC des États-Unis dans le *BLS Handbook of Methods* (2005), chapitre 17. Pour l'indice des prix de détail du Royaume-Uni, nous nous sommes fondés sur le document intitulé *The Retails Prices Index Technical Manual* (1998). Une description des pratiques plus récentes adoptées au Royaume-Uni peut être consultée dans le *Consumer Price Indexes Technical Manual* (2005).

3.1 Méthodes d'échantillonnage aux États-Unis

Nous commencerons par décrire les méthodes d'échantillonnage appliquées aux États-Unis, qui nécessitent trois enquêtes avec échantillonnage probabiliste, à savoir 1) une enquête-ménages, appelée Consumer Expenditure Survey (CEX), pour estimer la répartition des dépenses des ménages entre diverses catégories de biens, 2) une deuxième enquête-ménages, appelée Point of Purchase Survey (POPS) pour estimer, dans chaque groupe d'articles, les montants relatifs dépensés dans divers points de vente, et 3) une enquête auprès des points de vente, grâce à laquelle sont sélectionnés des articles individuels dont le prix est relevé. Dans les trois cas, l'échantillonnage pour la simulation est aléatoire avec remise (quoique l'échantillonnage employé en pratique soit nettement plus compliqué). Les deux premières enquêtes sont fondées sur des échantillons aléatoires simples et la dernière, sur un échantillon sélectionné avec probabilité proportionnelle à la taille (*ppt*), où les mesures de taille sont fonction des dépenses estimées d'après la CEX et la POPS. L'échantillon de la troisième enquête est un ensemble d'articles compris dans les combinaisons point de vente/ANE.

3.1.1 CEX (enquête-ménages)

L'objectif est d'estimer E_{lc}, c'est-à-dire les dépenses brutes des ménages au titre de l'ANE c dans l'UPE l. Nous avons procédé à un échantillonnage aléatoire simple avec remise (*easar*) à partir du fichier décrit plus haut, dans les UPE, de manière à obtenir des estimations à facteur d'extension sans biais

$$\hat{E}_{lc}^{95} = \frac{N_l^{95}}{n_{xl}} \sum_{j \in l \cap s(xl)} \sum_{i \in c \cap s(xl)} E_{ljci}^{95},$$

où $E_{lji}^{y} = q_{lji}^{y}\, p_{lji}^{y}$, N_l^{95} est la taille de population (nombre d'enregistrements pour l'*upe l* en 1995 – 1996) et n_{xl} est la taille de l'échantillon CEX $s(xl)$ dans l'UPE l, choisies de façon à concorder avec les tailles réelles d'échantillon de la CEX américaine et d'obtenir des coefficients de variation des estimations s'approchant de ceux obtenus dans le cas de la CEX américaine réelle; le x dans $s(xl)$ et dans n_{xl} sert simplement à faire la distinction entre l'enquête CEX et l'enquête POPS (dont la notation correspondante est « p »; voir plus loin) ou l'enquête sur les prix. Cette « imitation de la CEX » est une version simplifiée de l'enquête réelle. Notre méthodologie reposait sur l'hypothèse tacite que tous les clients d'un point de vente donné achètent les articles dans les mêmes proportions; elle ne tenait pas compte de l'erreur de mesure inévitable dans toute enquête réelle sur les dépenses, et (pour 1995 – 1996) elle était trop récente : les données réelles de la CEX sont souvent antérieures de plusieurs années aux enquêtes auprès des points de vente pour lesquelles elles sont utilisées. Toutefois, puisque les « données des ménages » recueillies ont aussi été utilisées dans les méthodes correspondantes du Royaume-Uni (voir plus loin), la version simplifiée suffit pour la comparaison souhaitée des méthodologies.

Tableau 3
Structure de population de l'« univers des céréales » : points de vente

R.-U.		É.-U.		N^bre	Symbole
Région		Unité d'échantillonnage	primaire	3	l
Type de magasin :	Indépendant	Chaîne 8			k
	Multiple : Relevé central	Chaîne 4			
	Multiple : Non central	Chaînes 1 à 3; 5 à 7			
Magasin		Point de vente		~300	j

Les dépenses de plus haut niveau ont été estimées par simple sommation. Par exemple, le total, étendu à l'ensemble des UPE, dans un ANE donné c est estimé par $\hat{E}^{95}_{.c} = \sum_l \hat{E}^{95}_{lc}$, *etc*. En tout, nous avons tiré 500 échantillons CEX, chacun produisant un ensemble correspondant d'estimations des dépenses.

3.1.2 POPS (enquête-ménages)

L'objectif de cette enquête est d'estimer la distribution des dépenses dans différents points de vente pour des catégories particulières de biens. Ces catégories pourraient être des ANE ou des groupes d'ANE; ici, nous supposons qu'il s'agit d'ANE. La TPOPS (Telephone Point of Purchase Survey) réelle des États-Unis est, comme son nom l'indique, réalisée par téléphone, selon un plan avec renouvellement de l'échantillon tous les quatre ans. Nous nous sommes efforcés, comme nous l'avons fait dans le cas de la CEX, de faire concorder les propriétés statistiques de notre procédure avec celles de la TPOPS réelle, mais il s'est avéré que faire correspondre les tailles d'échantillon dans notre fichier de 20 000 enregistrements nous aurait donné des fractions d'échantillonnage dans les UPE plus grandes qu'il n'était souhaitable. Par conséquent, nous avons réduit les tailles d'échantillon de moitié, de sorte que notre « imitation de POPS » devrait avoir une précision d'environ $1/\sqrt{2}$ celle de la TPOPS réelle. De nouveau, cette modification n'aura pas d'incidence sur les conclusions de l'étude, parce que nous avons utilisé des données identiques pour la construction de l'enquête britannique. Nous avons tiré des échantillons $s(pl)$ de taille n_{pl} par *easar* et procédé à l'estimation au moyen de l'estimateur à facteur d'extension :

$$\tilde{E}^{y}_{lcj} = \frac{\mathrm{N}^{y}_{l}}{\mathrm{n}_{pl}} \sum_{i \in c \cap s(pl)} E^{y}_{ljci}.$$

Puisque les données de la POPS ont tendance à être plus à jour que celles de la CEX, nous choisissons y comme année de base de l'indice, 1995 pour 1995–1996, mais 1996 pour 1996–1997, *etc*. Nous avons exécuté 500 passages machines et obtenu 500 ensembles d'estimations, qui ont chacun été apparié à une réalisation de la CEX.

3.1.3 Échantillonnage des points de vente

Pour chaque année y, la sélection des articles pour lesquels les prix doivent être relevés comprend les étapes suivantes :

a) Pour chaque UPE l, et chacune des dix strates d'articles h, nous sélectionnons deux ANE c par échantillonnage avec probabilité proportionnelle à la taille avec remise (*pptar*), avec la mesure de taille $\hat{E}^{95}_{lc}$ dérivée de la CEX.

b) Pour chaque ANE c sélectionné, nous tirons huit points de vente j par échantillonnage *pptar*, en utilisant comme mesure de taille les estimations des dépenses d'après la POPS $\tilde{E}^{\mathrm{y}}_{ljc}$. Donc, en tout, nous obtenons 160 paires ANE-point de vente par UPE, et un nombre total de 480, avec un certain degré de répétition éventuel.

c) Dans chaque groupe point de vente-ANE (j, c), nous « allons » (comme l'agent de terrain irait littéralement) dans le point de vente et « dressons la liste » de tous les articles appartenant à l'ANE et des dépenses correspondantes à la première période E^{y}_{ljci}, et, au moyen de cette base de sondage dans le point de vente, nous sélectionnons un article par échantillonnage *ppt*.

Pour chaque article ainsi sélectionné, nous enregistrons les prix p^{y}_{ljhci}, $y = 1, 2$. Donc, nous notons que tous les aspects de l'échantillonnage des points de vente sont *ppt* avec remise, en fonction des estimations des dépenses provenant de l'une ou l'autre des deux enquêtes-ménages ou provenant directement du magasin sélectionné. De nouveau, nous avons exécuté 500 passages machines, chacun correspondant à une réalisation CEX/POPS unique.

3.2 Estimation aux États-Unis

Les « agrégats élémentaires » $\hat{I}^{y,\,y+1}_{lh}$, c'est-à-dire des estimations d'indice au niveau de l'UPE × strate d'articles, sont les éléments à partir desquels est construit l'IPC. Dans la plupart des IPC partout dans le monde, les indices de niveau le plus faible sont des moyennes non pondérées d'une sorte ou l'autre, comme l'estimateur RM du Royaume-Uni décrit plus loin, et les données sur les dépenses ne sont utilisées que pour agréger ces indices à des niveaux plus élevés. Aux États-Unis, les indices élémentaires sont essentiellement des estimateurs d'Horvitz-Thomson fondés explicitement ou implicitement sur les estimations des dépenses provenant de la CEX ainsi que de la POPS. Ces dernières années, pour la plupart des strates d'articles, les États-Unis ont adopté la formule de la *moyenne géométrique* (voir l'annexe A), de sorte que les estimations à ce niveau prennent la forme

$$\hat{I}^{y,\,y+1}_{lh} = \prod_{\substack{j \in l,\\ i \in c \in h \\ (i,\,j) \in s}} \left(\frac{p^{y+1}_{ljhci}}{p^{y}_{ljhci}} \right)^{s_{ljhci}},$$

où

$$\mathrm{s}_{ljhci} = \frac{w_{ljhci}}{\sum\limits_{\substack{j \in l,\, c \in h,\, i \in c \\ (i,\,j) \in s}} w_{ljhci}},$$

avec

$$w_{ljhci} = \frac{\tilde{E}_{lc}\,\hat{E}_{lh}}{\hat{E}_{lc}}\, w_{ljhci},$$

$j \in l$, $i \in c \in h$ et $(i,\ j) \in s$. Notons que les poids ne sont pas particuliers au i^{e} article; nous omettons les indices supérieurs de période par souci de brièveté. Les poids ne sont pas simplement égaux à l'inverse du nombre n_{lh} d'articles échantillonnés dans lh, comme des considérations d'absence de biais d'échantillon pourraient porter à le croire (Balk 2003), parce que les probabilités d'échantillonnage ne reflètent pas les dépenses exactes en articles à la période de base; voir le *BLS Handbook of Methods* (2005).

Puis, les indices élémentaires sont agrégés en utilisant les dépenses estimées d'après la CEX conformément à la formule de Laspeyres, par exemple

$$\hat{I}_{.h}^{y,\,y+1} = \frac{\sum_{l} \hat{E}_{lh}\,\hat{I}_{lh}^{y,\,y+1}}{\sum_{l} \hat{E}_{lh}}$$

pour obtenir l'indice pour une strate d'articles donnée h, sur l'ensemble des UPE.

3.3 Méthodes d'échantillonnage du Royaume-Uni

Au Royaume-Uni, comme aux États-Unis, la méthode d'estimation comporte la combinaison de trois composante : 1) une enquête-ménages, appelée Family Expenditure Survey (FES), pour estimer les montants consacrés à l'achat de divers groupes d'articles, 2) une enquête-magasins, appelée Annual Retailing Inquiry (ARI), pour obtenir des renseignements sur les dépenses par section et type de magasin et 3) une enquête auprès des points de vente des magasins, pour sélectionner des articles pour l'établissement des prix.

3.3.1 FES (enquête-ménages)

L'objectif est d'estimer les dépenses $E_{..c}$ pour des articles représentatifs c, et les dépenses $E_{l..h}$ pour des combinaisons région-section. Aux fins de la présente étude, nous supposons qu'il existe une concordance passage-machine par passage-machine entre les données pour la CEX des États-Unis et pour la FES du Royaume-Uni, de sorte que nous avons, de nouveau, 500 ensembles de données FES. Notons que le Royaume-Uni ne cherche pas à obtenir les estimations plus détaillées $E_{l..c}$ que visent les États-Unis.

3.3.2 Annual Retailing Inquiry (enquête-magasins)

L'objectif est d'obtenir des estimations des dépenses $\tilde{E}_{kh}$, par section et type de magasin. Cet objectif est considérablement plus général que l'obtention d'estimations point de vente (magasin) par ANE (article représentatif) visées par la POPS des États-Unis. Nous utilisons, pour construire les estimations ARI, pour chacun des 500 passages, les mêmes données que celles utilisées pour construire les estimations POPS pour la simulation de l'IPC des États-Unis.

3.3.3 Échantillonnage des points de vente

La sélection des articles pour lesquels les prix doivent être relevés comprend les étapes suivantes :

a) Un « échantillon au jugé » d'articles représentatifs c est sélectionné dans chaque section h. Dans la présente étude (uniquement pour permettre la simulation), dans chaque section, nous sélectionnons les deux articles représentatifs ayant les valeurs les plus grandes de $\hat{E}_{hc}$. Il convient de souligner deux différences par rapport à l'étape (a) correspondante de la méthode américaine : i) la sélection est uniforme sur l'ensemble des régions l; ii) la sélection n'est pas aléatoire et, en particulier, ne permet pas la sélection répétée des articles représentatifs. (La sélection répétée peut avoir lieu dans la méthode américaine simulée, à cause de l'échantillonnage avec remise des ANE dans les strates d'articles.)

b) Les économistes de terrain choisissent les magasins dans une localité particulière dans laquelle le prix d'un article représentatif doit être établi. Traditionnellement, cela se faisait par *eassr*, après que l'économiste de terrain ait construit une base de sondage des magasins appropriés. Plus récemment, la sélection a été faite par échantillonnage *ppt*, où la mesure de taille est la superficie consacrée dans le magasin au type de biens que l'article représentatif représente. Les économistes de terrain ne tirent pas d'échantillons d'articles « dont le relevé des prix est centralisé » : dans le cas d'un très grand magasin multiple, le prix d'un article est relevé auprès du bureau central de ce magasin et est considéré comme représentatif du prix de l'article dans tous les magasins faisant partie du multiple. Dans la présente étude, nous avons procédé de la façon suivante : pour chaque région l et chaque article représentatif c, nous avons sélectionné huit magasins comme il suit :
   - 4 parmi les magasins multiples à relevé de prix non centralisé (chaînes 1, 2, 3, 5, 6, 7)

1 provenant d'un magasin multiple à relevé de prix centralisé (chaîne 4)
3 parmi les magasins indépendants (chaîne 8)

Dans chaque cas, pour simplifier, nous avons procédé à l'échantillonnage aléatoire simple (*eas*) sans remise des magasins pour lesquels les dépenses étaient positives pour l'article représentatif. Le nombre de magasins dans la simulation pour le Royaume-Uni (8 par article représentatif de chaque région) coïncide avec le nombre de « points de vente » dans la simulation pour les États-Unis; il existe 160 paires magasin-article représentatif par région, soit 480 en tout. Soulignons les différences qui suivent par rapport à la méthode appliquée aux États-Unis :

1. L'information sur le type de magasin est utilisée pour la stratification (et jouera un rôle dans l'estimation décrite plus loin). Cette information est disponible dans l'échantillon des États-Unis, mais est passée outre au profit de la méthode d'échantillonnage *ppt*.
2. Pour le Royaume-Uni, nous permettons de l'information sur la présence ou l'absence de l'article représentatif spécifique c (équivalent à l'ANE) dans la liste de magasins avant l'échantillonnage, tandis que pour les États-Unis, on ne connaît effectivement que l'existence *d'un certain* ANE dans la strate d'articles donnés. (Cela sous-entend des mises en correspondance multiplies de catégories ANE à POPS, ce qui était typiquement le cas jusqu'à récemment aux États-Unis; la version courante des appariements de catégories ANE à TPOPS (telephone point of purchase survey) est 1 à 1; autrement dit, une base de sondage de points de vente est construite pour chaque ANE.)

c) Traditionnellement, pour chaque élément représentatif c, dans un magasin particulier, l'économiste de terrain sélectionne la variété i qu'il ou elle considère comme dominant les ventes, c'est-à-dire un échantillonnage au jugé de la variété achetée la plus systématiquement. Nous formalisons cela de la façon suivante :

1. Pour une paire magasin-article représentatif donnée (j, c), nous dressons la liste de toutes les variétés i.
2. Pour chaque variété, nous trouvons la quantité minimale $q_i^* = \text{Min}(q_i^y, q_i^{y+1})$ sur deux années.
3. Nous échantillonnons la variété i avec $\text{Max}\{q_i^*\}$.

Naturellement, ce processus requiert plus d'information que n'en posséderait un économiste de terrain à la première période (et, de nouveau, n'est pas utilisé dans la méthode d'échantillonnage américaine décrite plus haut), et peut être considéré comme offrant un substitut pour l'évaluation par l'économiste de terrain de la continuité relative des biens vendus.

Nota : Il est commode d'utiliser l'expression échantillonnage *maxminq* pour désigner la combinaison de la sélection d'un point de vente par *eassr* comme en (b) et d'un article dans le magasin comme en (c).

3.4 Estimation au Royaume-Uni

Pour le Royaume-Uni, les agrégats élémentaires ont été calculés au moyen d'une formule de ratio des moyennes (RM) dans chaque cellule de classification croisée définie par la région, le type de magasin et l'article représentatif. Il s'agit fondamentalement d'une estimation non pondérée donnée, pour les magasins indépendants, par

$$\hat{I}_{lkhc}^{y,\,y+1} = \frac{\frac{1}{n}\sum_{i\in c,\ j\in k} p_{ljhci}^{y+1}}{\frac{1}{n}\sum_{i\in c,\ j\in k} p_{ljhci}^{y}}.$$

Dans le cas des magasins multiples, une version pondérée de la formule susmentionnée est utilisée avec les dépenses par type de magasin, estimées d'après l'ARI, qui fournissent les poids relatifs des magasins multiples à relevé des prix centralisé par opposition à non centralisé.

Un indice à l'échelle du pays pour les articles représentatifs c dans l'échantillon (agrégés sur les types de magasin k et les régions l) est alors calculé à l'aide d'un estimateur de type Laspeyres :

$$\hat{I}_c^{y,\,y+1} = \sum_l \sum_k \tilde{w}_{lkhc}\, \hat{I}_{lkhc}^{y,\,y+1},$$

où $c \in h$, et $\tilde{w}_{lkhc}$ est fondé sur $\tilde{E}_{.kh}^{y}$ provenant de l'ARI et $\hat{E}_{lh}^{95}$ provenant de la FES (l'utilisation de ces périodes fait en sorte que l'information utilisée est la même pour les États-Unis et le Royaume-Uni). Une agrégation supplémentaire (sur les articles représentatifs c) est faite en utilisant $\hat{E}_{hc}^{y}$, *etc*. provenant de la FES.

3.5 Comparaison

Le tableau 4 donne une comparaison sommaire des méthodes appliquées aux États-Unis et au Royaume-Uni que nous avons considérées. La caractéristique prédominante de

la méthode américaine est l'échantillonnage et l'estimation probabiliste, typiquement *pptar*, et celle de la méthode britannique est l'échantillonnage sélectif, en choisissant l'article ou la catégorie que l'on estime avoir le plus d'importance selon le montant des dépenses ou la quantité vendue. Les méthodes de formation des agrégats élémentaires diffèrent et les poids utilisés pour l'agrégation au Royaume-Uni sont estimés à un niveau un peu moins fin de détail aux étapes inférieures.

Le tableau 5 résume ce que l'on pourrait considérer comme étant les points forts et les points faibles des méthodes américaine et britannique. Par avantage de « force brute », que nous attribuons à l'approche britannique, nous entendons l'exploitation d'une combinaison de deux facteurs qui jouent souvent un rôle dans l'établissement des prix et la construction des indices des prix. En premier lieu, les leaders du marché ont tendance à dicter le prix; par exemple, s'ils augmentent ou réduisent fortement les prix, leurs concurrents moins importants vendant des biens similaires peuvent penser qu'il est nécessaire ou justifié de suivre leur exemple. En deuxième lieu, même si la tendance des prix varie entre biens semblables, les principaux vendeurs domineront vraisemblablement l'indice des prix en raison du montant élevé des dépenses, autrement dit, à cause de leur pondération conséquemment élevée.

Tableau 4
Comparaison sommaire des méthodes appliquées aux États-Unis et au Royaume-Uni

	É.-U.	R.-U.
Enquête sur les dépenses des ménages	$\hat{E}_{lc}^{95}$	$\hat{E}_{.c}^{95}$, $\hat{E}_{lh}^{95}$
Dépenses par point de vente/catégorie	Ménages (POPS) $\tilde{E}_{ljc}^{y}$	Enquête auprès des magasins (ARI) $\tilde{E}_{.kh}^{y}$
Choix des catégories d'articles	2 ANE *c*/strate d'articles *h*/UPE *l* *pptar* $(\hat{E}_{lc}^{95} / \hat{E}_{lh}^{95})$	2 articles représentatifs *c*/section *h*/région *l* *Les plus importants* $(\hat{E}_{.c}^{95} / \hat{E}_{.h}^{95})$
Choix des points de vente	8 points de vente *j*/ANE $c \times$ UPE *l* *pptar* $(\tilde{E}_{ljc}^{y} / \tilde{E}_{l.c}^{y})$	8 points de vente *j*/article représentatif $c \times$ région l – *eas* avec type de magasin *k*, $E_{ljc}^{y} > 0$
Article dans un point de vente/une catégorie	1 article *i*/*jc ppt* $(E_{ljci}^{y} / E_{ljc.}^{y})$	1 variété *i*/*jc* $\max[\text{Min}(q_{ji}^{y}, q_{ji}^{y+1})]$
Indice élémentaire	$\hat{I}_{lh}^{y,\,y+1} = \prod_{\substack{j\in l,\\ i\in c\in h\\ (i,\,j)\in s}} \left(\frac{p_{ljhci}^{y+1}}{p_{ljhci}^{y}}\right)^{s_{ljhci}}$	$\hat{I}_{lkhc}^{y,\,y+1} = \frac{\frac{1}{n}\sum_{i\in c,\, j\in k} p_{ljhci}^{y+1}}{\frac{1}{n}\sum_{i\in c,\, j\in k} p_{ljhci}^{y}}$
Niveau plus élevé d'agrégation	$\hat{I}_{.h}^{y,\,y+1} = \frac{\sum_{l} \hat{E}_{lh}\, \hat{I}_{lh}^{y,\,y+1}}{\sum_{l} \hat{E}_{lh}}$	$\hat{I}_{c} = \sum_{l}\sum_{k} \hat{\tilde{w}}_{lkhc} \hat{I}_{lkhc}$, $\hat{\tilde{w}}_{lkhc} = f(\tilde{E}_{.kh}, \hat{E}_{l.h})$

Tableau 5
Comparaison des approches américaine et britannique

	Points forts	Points faibles
É.-U.	Recueille plus d'information Plus grande utilisation de l'information Satisfait la théorie classique de l'échantillonnage Donne des estimateurs pondérés des estimations régionales (UPE) au niveau le plus faible Procédure opérationnelle plus normalisée	Répétition éventuelle lors de la sélection Ne tient pas compte de la stratification des magasins (c'est-à-dire, de la classification en chaînes)
R.-U.	S'appuie sur le principe de « force brute » Stratification des points de vente L'enquête sur le terrain auprès des magasins s'appuie sur diverses sources	Ensemble disparate de coefficients de pondération Inconsistante dans le cas de l'agrégation des prix relevés de façon centralisée? Estimateur non pondéré et en apparence arbitraire au niveau le plus faible.

4. Résultats de l'étude principale

Le tableau 6 donne les indices comparant 1995 à 1996 pour la population 1) dans son ensemble (les trois domaines combinés), 2) ventilée en classes/grands groupes et 3) ventilée encore davantage en strates d'articles/sections. Quatre indices, qui pourraient être considérés comme les cibles de l'estimation, sont donnés. À cet égard, rappelons la discussion sur les cibles qui figurent à la section 1.

Le tableau 7 donne les moyennes, variances et erreurs quadratiques moyennes correspondantes pour les estimations américaines et britanniques, où l'erreur quadratique moyenne est calculée par rapport aux indices de Fisher. Nous faisons les constatations suivantes :

1) Pour l'ensemble des articles, les catégories et les strates d'articles, les estimations américaines semblent s'approcher de la *moyenne géométrique* G. Cela confirme ce que d'autres travaux nous avaient fait soupçonner (Dorfman et coll., 1999), à savoir que le niveau le plus faible d'agrégation est dominant (nous avons utilisé une formule de Laspeyres pour les niveaux d'agrégation plus élevés). Le fait que G se situe entre l'indice de Laspeyres et l'indice cible superlatif donne certaines preuves que le passage des États-Unis à cette méthode d'agrégation élémentaire représente un pas dans la bonne direction.

2) Il ne semble exister aucune relation d'ordre claire entre les estimations britanniques au niveau de la *section* et les cibles correspondantes; par exemple, l'indice pour la section 11 est plus élevé que la cible L, tandis que l'indice pour la section 12 est plus faible que les indices superlatifs, *etc*. Toutefois, lorsque nous poursuivons l'agrégation jusqu'aux niveaux du grand groupe et de l'ensemble des articles, les estimations commencent manifestement à s'approcher des indices superlatifs F ou T. (Dalén (1998) a noté un résultat similaire lors de l'agrégation d'échantillons avec seuil d'inclusion.)

3) Si nous choisissons l'indice de Fisher comme cible, *même au* niveau de la section, la racine carrée de l'erreur quadratique moyenne de l'estimateur du Royaume-Uni est nettement plus faible que celle de l'estimateur des États-Unis. Étant donné la nature relativement restreinte du plan d'échantillonnage du Royaume-Uni, il n'est pas étonnant que l'estimateur de ce pays présente une variance plus faible, mais sa forme ne porterait pas à penser qu'il donne une approximation sans biais d'un indice de Fisher. Néanmoins, nos résultats laissent entendre que, du moins pour une population d'achats tels que ceux utilisés dans l'étude, les méthodes par choix raisonné, à « force brute », du Royaume-Uni (et de nombreux autres pays) donnent de bons résultats.

Des résultats similaires ont été obtenus pour les paires successives d'années jusqu'à 1999 – 2000. La figure 6 donne la *moyenne géométrique* et l'indice de Fisher d'une année sur l'autre pour l'ensemble des articles pour cinq paires d'années, ainsi que les moyennes sur l'ensemble des échantillons des estimateurs américain et britannique correspondant. (Il convient de souligner la différence d'échelle entre la figure 6 et les figures 1 à 5.) Il est facile de voir que l'estimateur américain a tendance à suivre la *moyenne géométrique* de population. L'estimateur britannique, qui suit l'indice de Fisher, a tendance à surestimer les prix dans les années les plus récentes, bien qu'il s'approche nettement plus de l'indice de Fisher que de la *moyenne géométrique* de population. Il convient de souligner que nous avons utilisé des données sur les dépenses de plus en plus périmées, à savoir les données pour 1995, pour l'échantillonnage et l'estimation. Il se peut que les données périmées sur les dépenses aient une incidence plus grande sur les estimations britanniques que sur les estimations américaines, peut-être en nous menant à suréchantillonner les articles représentatifs coûteux ou à nous concentrer sur certains groupes de magasins qui pratiquent des prix de plus en plus élevés.

Les résultats pour les catégories (« chaudes », *etc*.) étaient forts semblables pour les États-Unis relativement à la *moyenne géométrique* et ne sont pas présentés. La figure 7 donne la différence entre les estimations britanniques moyennes d'une année sur l'autre et l'indice de Fisher, pour chacune des quatre catégories. Son examen révèle que la tendance à la surestimation dans les années les plus récentes affecte les quatre catégories.

Dans l'ensemble, les estimateurs du Royaume-Uni fournissent de meilleures estimations de l'indice superlatif cible de Fisher que ceux des États-Unis. Le tableau 8 donne le ratio de la racine carrée de l'erreur quadratique moyenne du Royaume-Uni à celle des États-Unis, pour les cinq paires d'années, pour l'ensemble des articles, pour les groupes et pour les sections. Il contient quelques valeurs anormales, notamment dans les indices 1998 – 1999 où, pour la section 2 de « chaudes » et, par conséquent, pour la catégorie « chaudes » complète, les estimations britanniques sont appréciablement moins bonnes. Cependant, en général, les méthodes du Royaume-Uni produisent de nettement meilleures estimations. Cela est attribuable, en partie, à une structure d'échantillonnage plus stricte (principalement parce que l'échantillonnage par choix raisonné/avec seuil d'inclusion est sensiblement plus contraignant que la sélection aléatoire de l'ensemble d'articles qui peuvent entrer dans l'échantillon), qui produit, ce qui n'est pas

étonnant, moins de variance. Néanmoins, cela est également dû, en partie, à une tendance étonnante des estimateurs britanniques à cibler les indices de Fisher correspondants, ce qui réduit le biais. Puisque les estimateurs britanniques ne ressemblent pas formellement à l'indice de Fisher, les raisons de leur tendance à approximer cet indice méritent d'être étudiées plus en profondeur. Nous nous penchons sur cette question à la section suivante.

Tableau 6
Indices 1995-1996 cibles possibles

Description	*Moyenne géométrique*	Törnqvist	Fisher	Laspeyres
Tous les articles	1,053	1,002	0,997	1,079
Catégories/Grands groupes				
1 – Chaudes	1,058	1,052	1,052	1,078
2 – Sucrées	1,042	0,964	0,956	1,072
3 – Fruitées	1,044	1,007	1,007	1,067
4 – Ordinaires	1,069	1,027	1,027	1,092
Strates d'articles/Sections				
Chaudes – 11	1,043	1,044	1,044	1,057
Chaudes – 12	1,073	1,059	1,058	1,097
Sucrées – 21	1,003	0,917	0,910	1,034
Sucrées – 22	1,063	0,982	0,972	1,093
Sucrées – 23	1,093	1,052	1,054	1,119
Fruitées – 31	0,977	0,955	0,950	0,985
Fruitées – 32	1,165	1,110	1,116	1,204
Ordinaires – 41	1,067	1,021	1,021	1,094
Ordinaires – 42	1,030	0,996	0,996	1,050
Ordinaires – 43	1,104	1,063	1,062	1,125

Tableau 7
Résultats des simulations pour les indices 1995-1996

		É.-U.			R.-U.		
Description	Indice cible	Moyenne	Écart-type	REQM	Moyenne	Écart-type	REQM
Tous les articles	0,997	1,057	0,016	0,062	1,002	0,011	0,012
Catégories/Grands groupes							
1 – Chaudes	1,052	1,059	0,031	0,032	1,045	0,022	0,023
2 – Sucrées	0,956	1,046	0,030	0,095	0,971	0,023	0,027
3 – Fruitées	1,007	1,053	0,035	0,058	0,986	0,027	0,034
4 – Ordinaires	1,027	1,072	0,025	0,051	1,025	0,016	0,016
Strates d'articles/Sections							
Chaudes – 11	1,044	1,045	0,035	0,035	1,064	0,025	0,032
Chaudes – 12	1,058	1,072	0,049	0,051	1,027	0,035	0,047
Sucrées – 21	0,910	1,004	0,050	0,106	0,850	0,045	0,074
Sucrées – 22	0,972	1,070	0,051	0,111	1,089	0,030	0,121
Sucrées – 23	1,054	1,095	0,044	0,060	1,026	0,027	0,039
Fruitées – 31	0,950	0,979	0,020	0,035	0,932	0,020	0,027
Fruitées – 32	1,116	1,178	0,084	0,104	1,077	0,059	0,071
Ordinaires –41	1,021	1,069	0,050	0,070	1,060	0,030	0,049
Ordinaires –42	0,996	1,033	0,035	0,051	0,987	0,031	0,032
Ordinaires - 43	1,062	1,107	0,042	0,061	1,028	0,023	0,041

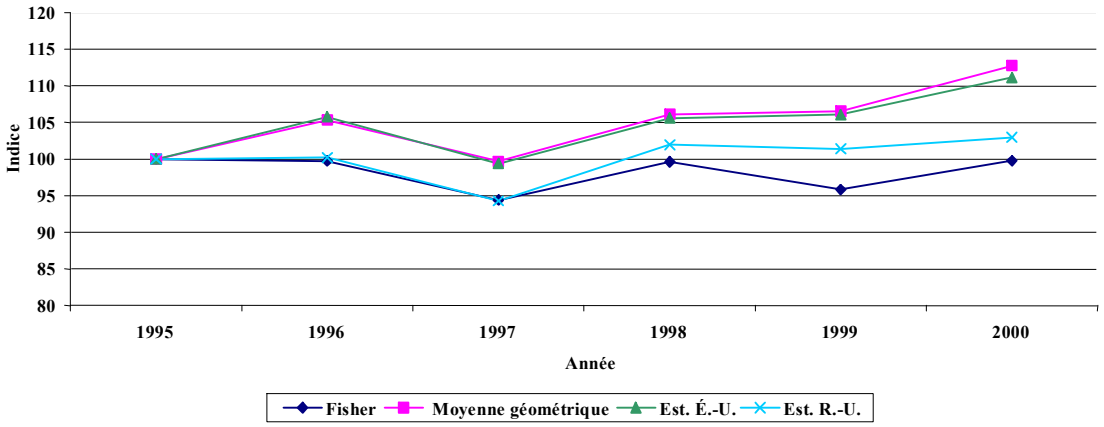

Figure 6. Indices cibles et estimations pour l'ensemble des céréales, indices et estimations de l'indice de février à février, 1995 = 100.

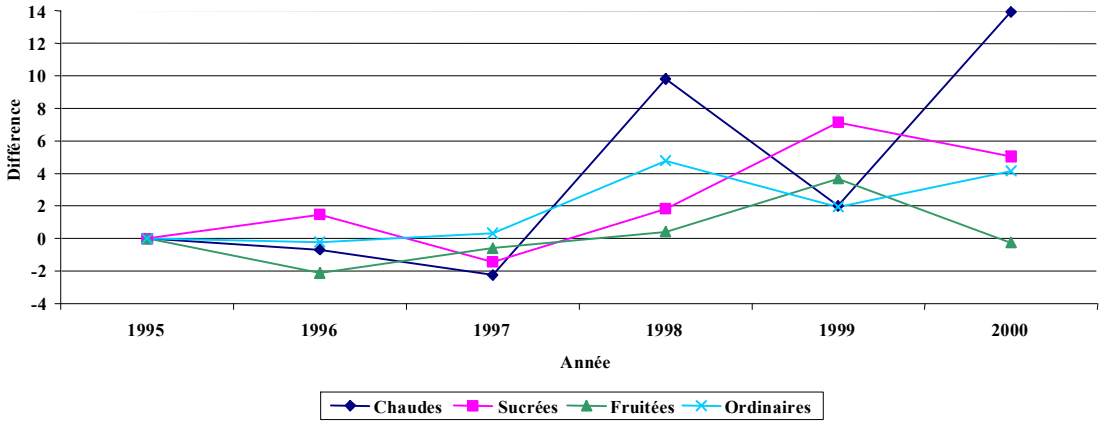

Figure 7. Différences entre les estimations britanniques et les indices de population de Fisher, indices et estimations de l'indice de février à février, 1995 = 100.

Tableau 8
Ratios de la REQM du R.-U. à la REQM des É.-U.

Description	1995–1996	1996–1997	1997–1998	1998–1999	1999–2000
Tous les articles	0,196	0,192	0,419	0,548	0,288
Catégories/Grands groupes					
1 – Chaudes	0,713	0,517	0,483	1,437	0,589
2 – Sucrées	0,286	0,336	0,314	0,522	0,282
3 – Fruitées	0,595	0,508	0,308	0,501	0,405
4 – Ordinaires	0,310	0,297	0,777	0,319	0,404
Strates d'articles/Sections					
Chaudes – 11	0,923	1,066	0,682	0,529	0,508
Chaudes – 12	0,920	0,850	1,169	1,860	0,842
Sucrées – 21	0,702	0,392	0,421	0,595	0,330
Sucrées – 22	1,092	0,426	0,380	0,341	0,365
Sucrées – 23	0,650	0,455	0,448	0,925	0,851
Fruitées – 31	0,778	1,059	0,637	0,581	0,618
Fruitées – 32	0,683	0,809	0,314	0,457	0,356
Ordinaires –41	0,709	0,623	0,494	0,567	0,317
Ordinaires –42	0,642	0,511	1,117	1,092	1,005
Ordinaires - 43	0,678	0,839	0,641	0,815	0,701

5. Étude de suivi

Les approches du Royaume-Uni et des États-Unis diffèrent à quatre égards : 1) la structure de stratification, en particulier l'utilisation au Royaume-Uni de différentes strates de magasins et, dans une certaine mesure, de l'échantillonnage centralisé, 2) la structure d'agrégation et de pondération, 3) le mode d'échantillonnage à divers degrés et 4) la formule des agrégats élémentaires. Il est donc difficile de déterminer dans quelle mesure chaque aspect contribue au mérite relatif des méthodes américaine et britannique de construction de l'indice. En particulier, comme nous l'avons mentionné à la dernière section, la raison pour laquelle l'estimateur de l'indice du Royaume-Uni a tendance à cibler les indices superlatifs, surtout au niveau plus élevé d'agrégation, demeure un peu mystérieuse.

Dans l'étude de suivi, nous nous concentrons sur le niveau le plus faible de construction de l'indice, c'est-à-dire sur (3), le niveau magasin-article représentatif (ANE) d'échantillonnage et sur (4), les formules des indices élémentaires. Nous comparons les avantages relatifs des diverses options, en prenant comme cibles les indices élémentaires à l'intérieur des régions. L'agrégation en vue d'obtenir des indices de niveau plus élevé sera exécutée uniformément pour toutes les options de niveau plus faible considérées, en utilisant les parts réelles des dépenses de population. L'importance de la méthode de construction des indices élémentaires est généralement reconnue; voir Diewert (2004) et les références, ainsi que Dorfman et coll. (1999). L'exemple exposé à l'annexe B, conjugué aux résultats présentés au tableau 9, illustre l'effet décisif que le niveau le plus faible de construction de l'indice a sur l'indice dans son ensemble.

Donc, une source vraisemblablement importante de la différence entre les résultats donnés par les méthodes américaine et britannique tient à l'estimation sur échantillon des indices élémentaires de population. Mais cela laisse ouverte la question de savoir si les écarts sont dus à des différences entre les méthodes d'échantillonnage ou entre les formules utilisées pour l'estimation, ou les deux. Donc, nous cherchons à déterminer 1) comment l'échantillonnage au jugé (ici, échantillonnage avec seuil d'inclusion basé sur *maxminq*) se compare à l'échantillonnage probabiliste représenté par *pptar*, en maintenant fixe l'estimateur des indices élémentaires, et 2) comment les estimateurs des indices élémentaires se comparent lorsque nous maintenons fixe la méthode d'échantillonnage. Il sera également intéressant de déterminer ce qui se passe lorsque l'échantillonnage *maxminq* est fondé sur des données provenant de la période de base et de la période *précédente*, plutôt que de la période de base et la période courante.

5.1 Méthodes d'échantillonnage et estimateurs au niveau élémentaire

Pour explorer ces questions, nous avons exécuté d'autres études par simulation. Nous nous sommes servis des mêmes données sur les céréales que celles utilisées pour l'étude principale (mois de février successifs), mais nous nous limitons aux magasins indépendants, *chaîne 8*. Nous avons procédé ainsi pour que l'étude soit plus facile à gérer, mais aussi parce que, pour les autres chaînes, les estimateurs des indices élémentaires utilisés au Royaume-Uni étaient plus compliqués que le simple indice de *Dutot*. En outre, il est raisonnable de s'attendre à ce que le comportement des prix soit le plus hétérogène dans cette chaîne, de sorte que les différences intrinsèques seront plus évidentes. La chaîne 8 est la plus grande des chaînes étudiées, comprenant chaque année environ 30 % de l'ensemble de la population, soit environ 6 000 enregistrements.

La structure de base est restée la même : 3 *upe*, 4 grands groupes/catégories de dépenses (chaudes, sucrées, fruitées et ordinaires), 10 sections/strates d'articles, et 29 articles représentatifs/*ANE*. Pour chaque *ANE/article représentatif*, 3 points de vente (un article par point de vente) ont été sélectionnés, par opposition à 10 dans le cas de l'étude principale. Pour étudier l'approche *maxminq* basée sur des périodes antérieures, nous avons réduit les cinq ensembles de données originaux, contenant chacun les données sur les prix et les quantités pour une paire d'années (1995 – 1996, 1996 – 1997, *etc*.) afin de n'inclure que les articles permettant un « rétro-appariement »; c'est-à-dire l'appariement sur trois années pour comparer les prix des articles dans les points de vente pour 1995/1996/1997, 1996/1997/1998, *etc*. Environ 90 % des enregistrements de la chaîne 8 ont permis un rétro-appariement (en ce qui concerne les résultats qui suivent, il vaut probablement la peine de souligner que la réduction de l'échantillon pourrait influencer de façon disproportionnée le *maxminq*). Nous déplaçons notre attention de l'indice de Fisher vers l'indice superlatif de Walsh, grâce à une suggestion astucieuse d'un examinateur, dont nous discutons à l'annexe C.

Nous avons utilisé trois estimateurs pour les indices élémentaires : le ratio des moyennes arithmétiques (RM) (le *Dutot*), la *moyenne géométrique* non pondérée (aussi appelée indice de Jevons) et la moyenne des ratios (*MR*). Dans l'échantillonnage *ppt* des points de vente, puis dans l'échantillonnage des articles dans les points de vente, nous avons supposé que la variable de taille (dépenses) était connue (au lieu d'être estimée, comme dans l'étude principale). Outre l'échantillonnage *ppt* avec remise (comme dans l'approche américaine), et *maxminq*, nous avons également étudié l'échantillonnage *ppt* sans remise, parce que nous soupçonnions qu'il serait moins variable que la version avec remise.

Pour chaque mode d'échantillonnage, dans chaque combinaison *upe/ANE*, nous avons tiré 500 échantillons. Nous avons calculé l'erreur quadratique moyenne des estimations par rapport à un indice cible de Walsh au niveau de l'*ANE*. Les moyennes des *eqm* sur l'ensemble des *ANE* ont été calculées pour chaque mode d'échantillonnage/ estimation, dans chaque *upe*.

Le tableau 10 donne le ratio de ces moyennes par rapport à l'*eqm* moyenne pour la combinaison *maxminq/Dutot*. Pour chaque estimateur, pour chaque *upe*, à une exception près (*upe* 3, 1999 – 2000), *maxminq* donne l'*eqm* la plus faible, souvent avec une marge appréciable. L'échantillonnage *ppt* sans remise est la deuxième des solutions les meilleures. Si nous maintenons la méthode d'échantillonnage fixe (en comparant les lignes 1, 4 et 7, puis 2, 5 et 8, *etc*. au tableau 10), nous constatons qu'à quelques exceptions près, les résultats sont meilleurs pour le *Dutot* que pour la *moyenne géométrique*, qui donne de meilleurs résultats que le *RM*. Ces résultats donnent à penser que 1) *maxminq* est meilleur que *ppt*(dép.), et que *ppt*(dép.) est meilleur que *pptar*(dép.) et que 2) le *Dutot* est plus efficace que la *moyenne géométrique*, et que la *moyenne géométrique* est plus efficace que la moyenne des ratios. Il existe une synergie favorable entre l'échantillonnage *maxminq* et l'indice de *Dutot.* Nous avons également étudié les biais et les variances, et les résultats (non présentés) avaient tendance à suivre le même profil.

Tableau 9
Indices de population 1995-1996, chaîne 8

Description	Laspeyres	*Moyenne géométrique**	Fisher	Walsh	Laspeyres de l'indice élémentaire de Walsh
Tous les articles	1,129	1,091	1,028	1,030	1,040
Catégories/Grands groupes					
1 – Chaudes	1,161	1,115	1,080	1,082	1,084
2 – Sucrées	1,129	1,088	1,007	1,012	1,025
3 – Fruitées	1,084	1,054	0,997	1,005	1,015
4 – Ordinaires	1,135	1,101	1,046	1,042	1,050
Strates d'articles/Sections					
Chaudes – 11	1,157	1,117	1,088	1,089	1,090
Chaudes – 12	1,164	1,113	1,072	1,075	1,079
Sucrées – 21	1,086	1,045	0,962	0,970	0,992
Sucrées – 22	1,187	1,142	1,055	1,056	1,058
Sucrées – 23	1,117	1,091	1,034	1,039	1,043
Fruitées – 31	1,003	0,992	0,949	0,965	0,966
Fruitées – 32	1,228	1,172	1,100	1,091	1,102
Ordinaires – 41	1,212	1,161	1,091	1,080	1,090
Ordinaires – 42	1,048	1,030	0,997	0,997	0,998
Ordinaires – 43	1,136	1,107	1,048	1,046	1,056

* Pondérée par les dépenses à la période de base.

Tableau 10
Moyenne standardisée de l'erreur quadratique moyenne relative sur l'ensemble des ANE, populations réduites, chaîne 8

	upe 2				*upe* 3				*upe* 4			
Estimateur/méthode d'échantillonnage	96 – 97	97 – 98	98 – 99	99 – 00	96 – 97	97 – 98	98 – 99	99 – 00	96 – 97	97 – 98	98 – 99	99 – 00
Dutot/*maxminq*(R.-U.)	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Dutot/*pptsr*	1,73	1,70	1,68	1,91	1,23	1,82	1,35	2,24	1,22	1,06	1,12	0,93
Dutot/*pptar*	2,13	2,10	1,91	2,14	1,42	2,10	1,46	2,67	1,45	1,23	1,36	1,07
Moyenne géométrique/ maxminq	1,20	1,16	1,16	1,06	1,06	1,14	1,08	1,05	1,10	1,11	1,12	0,96
Moyenne géométrique/*pptsr*	2,08	1,88	1,98	2,27	1,33	1,94	1,47	2,59	1,33	1,09	1,28	0,97
Moyenne géométrique/*pptar* (É.-U.)	2,49	2,29	2,18	2,53	1,58	2,23	1,58	3,09	1,59	1,30	1,52	1,12
RM/*maxminq*	1,42	1,32	1,31	1,14	1,24	1,03	1,30	1,05	1,11	1,20	1,21	1,07
RM/*pptsr*	2,81	2,35	2,49	2,85	1,70	2,31	1,77	3,43	1,57	1,30	1,42	1,17
RM/*pptar*	3,23	2,77	2,66	3,08	2,03	2,58	1,87	3,96	1,83	1,49	1,66	1,30
Dutot/*maxminq*, *q* antérieures	1,12	1,19	1,19	1,41	1,56	1,42	1,69	1,51	1,20	1,02	0,85	1,48

Le fait que l'indice d'échantillon de *Dutot* puisse cibler l'indice de population de Walsh (et donc, indirectement, tout indice superlatif) lorsque les vendeurs les plus importants sont systématiquement échantillonnés est, selon nous, le résultat d'un mécanisme très simple, de « force brute » : dans la mesure où l'indice de Walsh peut être représenté par un petit échantillon d'articles, il est représenté le mieux par ceux pour lesquels les quantités sont régulièrement les plus grandes, et ce sont ces articles que le scénario d'échantillonnage *maxminq* fournit presque toujours. À l'annexe C, nous discutons d'une autre explication des bonnes propriétés de la combinaison *maxminq/Dutot*.

La moyenne des erreurs quadratiques moyennes a également été calculée pour la combinaison *maxminq/Dutot* en se fondant sur les valeurs *antérieures* de q, c'est-à-dire q_i^{y-1}, q_i^y. Les résultats sont présentés à la dernière ligne du tableau 10. Nous observons un affaiblissement attendu comparativement à la combinaison *maxminq/Dutot* mise à jour, mais la comparaison des résultats à ceux d'autres options demeure favorable. Nous étudions cet aspect plus en profondeur à la sous-section 5.2.

5.2 Effet des quantités décalées sur l'échantillonnage *maxminq*

Afin de placer les résultats de la section 4 dans leur contexte, nous devons déterminer quel est l'effet de l'utilisation de valeurs décalées de q dans *maxminq*. La raison en est simple : si, à première vue, l'utilisation des quantités des périodes de base et courante semble le moyen évident de refléter la notion d'articles persistants utilisée au Royaume-Uni, cela implique néanmoins l'utilisation d'information (les quantités de la période courante) qui n'a pas été utilisée pour simuler l'échantillonnage utilisé aux États-Unis, ce qui pourrait donner un avantage injuste à la méthode du Royaume-Uni.

Par conséquent, nous comparons l'approche des États-Unis, c'est-à-dire *pptar* (avec taille variable basée sur les dépenses de la période) et la *moyenne géométrique* au niveau élémentaire à l'approche du Royaume-Uni représentée par *maxminq – Dutot*, mais en fondant ici *maxminq* sur les quantités q_{y-1} et q_y. Nous avons réduit légèrement les ensembles de données pour être certains d'obtenir des données concordantes pour les trois années consécutives. L'agrégation pour produire des indices de niveau plus élevé a été faite d'après les dépenses réelles de population pour les États-Unis, ainsi que le Royaume-Uni.

Le tableau 11 donne les résultats pour les indices calculés pour l'ensemble des céréales pour la chaîne 8, en comparant les biais, les écarts-types et les racines de l'erreur quadratique moyenne par rapport à l'indice de population de Walsh. Comme prévu, les résultats ne sont pas aussi bons que ceux obtenus en utilisant les valeurs courantes de q. Néanmoins, en ce qui a trait aux trois mesures d'exactitude (biais, écart-type et racine de l'erreur quadratique moyenne), la combinaison *maxminq/Dutot* du Royaume-Uni continue de donner de meilleurs résultats que l'approche des États-Unis représentant l'échantillonnage probabiliste.

Pour les catégories plus fines, le tableau 12 donne les ratios des erreurs quadratiques moyennes obtenues sous la méthode du Royaume-Uni avec les valeurs décalées de q à celles obtenues sous la méthode des États-Unis. Bien qu'ils soient généralement plus élevés que ceux du tableau 8, ils donnent encore à penser que l'approche de l'échantillonnage par choix raisonné du Royaume-Uni est meilleure.

6. Discussion

Nous avons présenté une comparaison de deux approches fondamentalement différentes de l'échantillonnage et de l'inférence pour l'établissement d'un indice des prix à la consommation. La conclusion inévitable est que, dans la population que nous avons étudiée, l'approche « R.-U. », qui comporte une stratification plus stricte et, par-dessus tout, un échantillonnage au jugé dans les strates plus restrictif que l'échantillonnage probabiliste de l'approche « É.-U. », produit de meilleurs estimations d'un indice superlatif cible.

Nous montrons qu'il en est ainsi, quel que soit l'estimateur de l'indice de prix de faible niveau (*Dutot*, ou *moyenne géométrique*, ou la moyenne des ratios) employé, bien que le *Dutot* (rapport des moyennes) donne les meilleurs résultats.

L'approche du Royaume-Uni est supérieure pour deux raisons : 1) son échantillonnage plus strict, limité aux articles sélectionnés (par exemple, voir le tableau 13 décrit à l'annexe C), mène, sans surprise, à une variance plus faible, constatation qui avait déjà été faite par de Haan et coll. (1999) et 2) les indices d'échantillon de *Dutot* ciblent les indices superlatifs sous échantillonnage du marché dominant, ce qui nous a surpris et a suscité l'étude décrite à la section 5. Par ailleurs, l'approche des États-Unis a donné un estimateur de l'indice pouvant être décrit comme étant sans biais, mais il était sans biais pour le « mauvais » indice de population basé sur la *moyenne géométrique* pondérée par les dépenses à la première période. Donc, il avait tendance à être considérablement plus élevé que l'indice superlatif cible, qu'il s'agisse de celui de Fisher, de Walsh ou de Törnqvist).

Si nous permettions aux tailles d'échantillon d'augmenter, nous pourrions nous attendre à ce que les variances des approches américaine et britannique diminuent l'une et l'autre, mais la variance du Royaume-Uni demeurerait plus faible. Le biais de l'estimateur des États-Unis pour l'indice superlatif cible ne serait pas affecté par l'accroissement de la taille d'échantillon, de sorte que l'erreur

quadratique moyenne relative de l'approche du Royaume-Uni deviendrait de plus en plus faible.

En pratique, évidemment, les quantités de la période 2 ne sont pas disponibles au moment de la sélection de l'échantillon (la période 1) et, dans le cadre de notre étude de suivi, nous donnons une certaine idée de la dégradation partielle qui résulte de l'utilisation des quantités antérieures : elle n'est pas suffisamment importante pour empêcher de conclure que l'approche du Royaume-Uni donne de meilleurs résultats. De surcroît, le jugement de l'économiste de terrain quant au meilleur vendeur pourrait être fondé sur des données plus récentes que celles d'il y a un an. Donc, l'effet réel pourrait être compris entre ceux des versions décalées et non décalées de *maxminq* que nous avons utilisées. Toutefois, en pratique, les économistes de terrain des États-Unis pourraient échantillonner fréquemment des articles dans les points de vente d'après une estimation de la part des dépenses qui est réellement une moyenne lissée des parts des dépenses de la période de base *et* de la période récente. Cela pourrait atténuer le biais que nous avons observé dans nos simulations, où seules les dépenses de la période de base ont été utilisées pour l'échantillonnage dans les magasins.

Tableau 11
Biais, écart-type et racine de l'erreur quadratique moyenne (tous multipliés par 1 000), dans l'estimation de l'indice de population de Walsh pour l'ensemble des céréales, chaîne 8, fondée sur trois approches d'échantillonnage/estimation des indices élémentaires*

	a) Biais				
	1995 – 1996	1996 – 1997	1997 – 1998	1998 – 1999	1999 – 2000
Dutot/maxminq	29	15	–13	33	2
Dutot/maxminq, q antérieures	–	46	32	82	36
Moyenne géométrique/pptar	78	62	66	82	66
	b) Écart-type				
	1995 – 1996	1996 – 1997	1997 – 1998	1998 – 1999	1999 – 2000
Dutot/maxminq	16	13	11	14	12
Dutot/maxminq, q antérieures	–	14	12	15	14
Moyenne géométrique/pptar	22	18	17	18	20
	c) Racine de l'erreur quadratique moyenne				
	1995 – 1996	1996 – 1997	1997 – 1998	1998 – 1999	1999 – 2000
Dutot/maxminq	33	20	17	36	12
Dutot/maxminq, q antérieures	–	48	34	83	39
Moyenne géométrique/pptar	80	65	68	84	68

* Au niveau de l'ANE/article représentatif. Pour obtenir des estimations de l'indice global, nous avons agrégé les estimations des indices élémentaires en utilisant les dépenses de population connues.

Tableau 12
Ratios de la REQM du R.-U. à la REQM des É.-U., chaîne 8, indices cibles de Walsh : *maxminq* en utilisant les valeurs décalées de *q* et *Dutot* versus *pptar*(dépenses) et *moyenne géométrique*

Description	1996 – 1997	1997 – 1998	1998 – 1999	1999 – 2000
Tous les articles	0,748	0,498	0,993	0,567
Catégories/Grands groupes				
1 – Chaudes	1,539	0,495	1,280	0,765
2 – Sucrées	0,563	0,676	0,941	0,797
3 – Fruitées	0,409	0,323	0,463	0,852
4 – Ordinaires	0,915	0,560	1,164	0,359
Strates d'articles/Sections				
Chaudes – 11	0,748	0,607	0,660	0,657
Chaudes – 12	1,695	0,599	1,333	0,843
Sucrées – 21	0,757	0,593	1,136	0,924
Sucrées – 22	0,370	0,776	0,751	0,671
Sucrées – 23	0,479	0,785	0,796	0,508
Fruitées – 31	0,570	0,443	0,678	1,008
Fruitées – 32	0,526	0,350	0,277	0,674
Ordinaires – 41	1,167	0,509	1,395	0,397
Ordinaires – 42	0,623	0,411	0,918	0,624
Ordinaires – 43	0,919	1,171	0,668	0,560

Tableau 13
Articles sélectionnés par *maxminq* et *ppt* $\sqrt{q_y q_{y+1}}$ dans 500 échantillons

1995 – 1996, chaîne 8, *upe* 2 ANE 105											
ppt	articles sélectionnés	2889	2803	1564	2763	1558	2242	2344	2776	760	2850
	% d'échantillons dans lesquels sélectionnés	43,2	32,2	10,4	5,4	3,87	1,53	1,33	0,87	0,8	0,4
maxminq	articles sélectionnés	2889	2803								
	% d'échantillons dans lesquels sélectionnés	80,87	19,13								
1995 – 1996, chaîne 8, *upe* 3 ANE 401											
ppt	articles sélectionnés	1731	2378	2866	1742	2922	2375	2528	403	871	
	% d'échantillons dans lesquels sélectionnés	33,27	18,8	12,8	12,73	9,47	4,6	4,27	2,8	1,27	
maxminq	articles sélectionnés	2378	1731	2866	1742						
	% d'échantillons dans lesquels sélectionnés	46,27	24,47	15	14,27						
1999 – 2000, chaîne 8, *upe* 4 ANE 401											
ppt	articles sélectionnés	1731	2866	1742	2378	2922	2528	403			
	% d'échantillons dans lesquels sélectionnés	30,07	21,93	14,3	11,07	9,53	6,8	6,27			
maxminq	articles sélectionnés	1742	2866	2922	1731						
	% d'échantillons dans lesquels sélectionnés	34,27	30,87	18	16,87						

Il est généralement reconnu que les approches sans randomisation sont intrinsèquement moins coûteuses. Ainsi, le nombre de points de vente à visiter est habituellement plus faible et le relevé des prix dans les points de vente requiert moins de main-d'œuvre. Donc, pour un budget donné, nous pouvons nous attendre à ce que la supériorité de l'approche du Royaume-Uni par rapport à l'échantillonnage probabiliste appliqué aux États-Unis soit encore plus grande que ne le laisse entendre la présente étude.

Il serait utile d'étendre l'étude aux données scannées recueillies pour d'autres produits que les céréales. Les articles dont les fluctuations des prix sont plus importantes seraient particulièrement intéressants. Dans une certaine mesure, le bon comportement de l'approche *maxminq/Dutot* peut être relié à la justesse étonnante de l'indice de population de *Dutot* par rapport aux indices superlatifs (comme l'illustre le tableau 1). Dans quelle mesure cette justesse est-elle typique et, en son absence, le bon comportement d'échantillonnage persisterait-il?

Un dernier *avertissement*. En pratique, il serait souhaitable d'injecter une dose de randomisation à certains stades du processus d'échantillonnage et, en particulier, de se montrer légèrement prudent à l'égard de l'échantillonnage centralisé – non pas pour des raisons statistiques, mais pour assurer l'impartialité et l'apparence d'impartialité (Reinsdorf et Triplett 2005, section II; Royall 1976).

Remerciements

Les opinions exprimées dans le présent article sont celles des auteurs et ne représentent pas la politique du Bureau of Labor Statistics ni du Bureau of Transportation Statistics des États-Unis. Les auteurs remercient David Richardson et Lyuba Rozental de leur avoir fourni les données sur les céréales et de leur appui à point nommé, Sonja Mapes et Scott Pinkerton, pour leurs travaux sur la classification des céréales en catégories, et Mick Silver, Adrian Ball et Dawn Camus, de leur avoir fourni des explications et de la documentation sur les méthodes d'établissement de l'IPC du Royaume-Uni. Les auteurs remercient également trois examinateurs et un rédacteur adjoint de leurs commentaires constructifs et de leurs encouragements à étendre l'étude, ainsi que J. De Haan, M. Reinsdorf et B. Moulton, pour leurs suggestions utiles. Nous tenons tout spécialement à souligner les encouragements du regretté M.P. Singh dont les suggestions en tant que rédacteur en chef ont orienté l'évolution finale du présent article.

Annexe A
Cibles – Indices de population

Laspeyres
$$L = \frac{\sum_i q_i^y p_i^{y+1}}{\sum_i q_i^y p_i^y}$$

Paasche
$$P = \frac{\sum_i q_i^{y+1} p_i^{y+1}}{\sum_i q_i^{y+1} p_i^y}$$

Walsh
$$W = \frac{\sum_i \sqrt{q_i^y q_i^{y+1}}\, p_i^{y+1}}{\sum_i \sqrt{q_i^y q_i^{y+1}}\, p_i^y}$$

Fisher
$$F = \left\{ \frac{\sum_i q_i^y p_i^{y+1}}{\sum_i q_i^y p_i^y} \frac{\sum_i q_i^{y+1} p_i^{y+1}}{\sum_i q_i^{y+1} p_i^y} \right\}^{1/2} = \sqrt{LP}$$

Törnqvist $$T = \prod_i \left(\frac{p_i^{y+1}}{p_i^y} \right)^{s_i^{y,\,y+1}},$$

où

$$s_i^{y,\,y+1} = \frac{1}{2}\left(\frac{p_i^y q_i^y}{\sum_i p_i^y q_i^y} + \frac{p_i^{y+1} q_i^{y+1}}{\sum_i p_i^{y+1} q_i^{y+1}} \right)$$

Moyenne géométrique $$G = \prod_i \left(\frac{p_i^{y+1}}{p_i^y} \right),$$

où

$$w_i = s_i^y = \left(\frac{p_i^y q_i^y}{\sum_i p_i^y q_i^y} \right)$$

ou

$$w_i = 1/N$$

Valeur unitaire $$U = \frac{\sum_i q_i^{y+1} p_i^{y+1} \Big/ \sum_i q_i^{y+1}}{\sum_i q_i^y p_i^y \Big/ \sum_i q_i^y}$$

Dutot $$RM = \frac{\sum_i p_i^{y+1} \Big/ N}{\sum_i p_i^y \Big/ N}$$

(« rapport des moyennes arithmétiques »)

Moyenne des ratios $$MR = \frac{\sum_i p_i^{y+1} \Big/ p_i^y}{N}$$

Annexe B
Exemple illustrant l'importance du niveau le plus faible d'agrégation

Nous présentons ici un exemple simple en vue d'illustrer l'importance de la méthode utilisée pour construire les indices élémentaires. Nous comparons les indices de population de Walsh aux indices résultant de l'agrégation des indices élémentaires de Walsh selon une formule de Laspeyres. La raison pour laquelle nous nous concentrons sur l'indice de Walsh est donnée à l'annexe C. L'indice de Walsh « pur » est

$$W = \frac{\sum_i \sqrt{q_i^y\, q_i^{y+1}\, p_i^{y+1}}}{\sum_i \sqrt{q_i^y\, q_i^{y+1}\, p_i^y}} = \sum \tilde{s}_h\, W_h^{y,\,y+1},$$

où $W_h^{y,\,y+1}$ est le h^{e} indice élémentaire de Walsh et

$$\tilde{s}_h = \frac{\sum_{i \in h} \sqrt{q_i^y\, q_i^{y+1}\, p_i^y}}{\sum_i \sqrt{q_i^y\, q_i^{y+1}\, p_i^y}}$$

sont les poids d'agrégation de Walsh appropriés. Nous comparons à cela une agrégation selon Laspeyres d'indices élémentaires de Walsh (« ersatz de Walsh »), $L_W^{y,\,y+1} = \sum\sum\sum s_h W_h^{y,\,y+1}$, où les s_h sont les poids standard à la période de base.

Les résultats sont donnés au tableau 9. Nous observons un écart perceptible entre l'indice de population réelle de Walsh et l'agrégat selon Laspeyres des indices élémentaires de Walsh, celui-ci ayant tendance à être un peu plus élevé. Cependant, ces différences sont du même ordre que celles entre ces indices et l'indice de Fisher. Elles sont faibles comparativement à l'écart entre la *moyenne géométrique* ou l'indice de Laspeyres et les indices superlatifs. Ce genre de résultat confirme qu'une procédure valable au niveau le plus faible est un élément essentiel de la construction d'un indice.

Annexe C
La combinaison *maxminq/Dutot*

Pourquoi la combinaison *maxminq/Dutot* donne-t-elle d'aussi bons résultats, paraissant donner lieu à une absence de biais pour les indices superlatifs?

Un examinateur nous a fait remarquer que l'échantillonnage *maxminq* ressemble considérablement à l'échantillonnage *ppt* avec taille variable $\sqrt{q_i^y\, q_i^{y+1}}$; pour l'échantillonnage *pptsr* ($\sqrt{q^y q^{y+1}}$), l'indice de *Dutot* est approximativement sans biais pour un indice cible de Walsh et, par conséquent, indirectement, pour tout autre indice superlatif.

En effet, pour l'espérance du numérateur de l'indice de *Dutot*, sous le scénario d'échantillonnage probabiliste, nous avons

$$E_\pi\left(\sum_{i \in s} p_i^{y+1} \right) = E_\pi\left(\sum_{i' \in U} I_{i'}\, p_{i'}^{y+1} \right) = \frac{n}{\sum_{i'} \sqrt{q_{i'}^y\, q_{i'}^{y+1}}} \sum_{i'} \sqrt{q_{i'}^y\, q_{i'}^{y+1}}\, p_{i'}^{y+1},$$

où $E_\pi(\,)$ signifie l'espérance par rapport au plan d'échantillonnage et $I_{i'}$ est un indicateur aléatoire prenant la valeur de 1 ou 0, si i' est dans l'échantillon ou non. Nous obtenons une expression similaire pour le dénominateur. Le ratio de ces deux valeurs espérées est l'indice de Walsh. Par conséquent, à part le (léger) biais de ratio habituel, qui, comme nous pouvons le montrer, est généralement positif, l'indice de *Dutot* cible en effet l'indice de Walsh sous le scénario d'échantillonnage *ppt*.

Nous devons nous demander si les deux modes d'échantillonnage ont effectivement tendance à présenter un chevauchement appréciable en ce qui concerne les articles sélectionnés. Pour chaque passage-machine, pour chaque *upe l*, *ANE c*, trois articles ont été sélectionnés par échantillonnage *maxminq* ou par échantillonnage *pptsr* ($\sqrt{q^y q^{y+1}}$) des articles compris dans *lc*. Le tableau 13 donne le pourcentage de fois (sur 500 passages) que des articles différents ont été sélectionnés dans l'échantillon, pour certains cas représentatifs sélectionnés arbitrairement. Nous concluons, pas entièrement sans nous étonner, que a) l'échantillonnage *ppt* produit une plus grande dispersion des articles sélectionnés, b) les articles sélectionnés par échantillonnage *maxminq* constituent un sous-ensemble de ceux sélectionnés par échantillonnage *ppt*, c) il existe une certaine corrélation des « articles dominants », c'est-à-dire de ceux qui ont le plus tendance à être sélectionnés par l'une ou l'autre méthode. Brièvement, les échantillonnages *maxminq* et *ppt* ($\sqrt{q^y\, q^{y+1}}$) semblent être reliés, mais lâchement.

Afin de mieux comprendre la relation entre les deux méthodes d'échantillonnage, nous avons estimé le biais et l'erreur quadratique moyenne, par rapport à l'indice de population de Walsh, de l'indice de *Dutot* pour chaque ANE, pour l'échantillonnage *maxminq* ainsi que *ppt* ($\sqrt{q_y\, q_{y+1}}$). Les estimations du biais et de l'EQM étaient fondées sur 500 passages pour chaque méthode d'échantillonnage. Les statistiques sommaires ont été calculées sur l'ensemble des ANE pour chaque paire d'années et chaque *upe*. Le tableau 14 donne le pourcentage d'ANE pour lequel les indices élémentaires de *Dutot* présentent un biais positif pour chaque mode d'échantillonnage. Comme prévu, l'échantillonnage *ppt* a tendance à produire un biais positif; nous constatons que l'échantillonnage *maxminq* donne des résultats tout aussi biaisés positivement que négativement.

Le tableau 15 (a) donne le pourcentage d'ANE pour lequel le biais absolu dû à l'utilisation de *maxminq* est plus important que celui du *ppt* ($\sqrt{q_y\, q_{y+1}}$). À cet égard, l'échantillonnage *ppt* est meilleur. Cependant le tableau 15(b) donne le pourcentage d'ANE pour lequel *maxminq* a produit une plus grande erreur quadratique moyenne et, ici, *maxminq* donne de meilleurs résultats pour toutes les combinaisons période/*upe*, sauf deux. Nous considérons l'erreur quadratique moyenne comme étant un critère plus décisif, surtout sachant la bidirectionnalité des biais produits par *maxminq*.

Tableau 14
Pourcentage d'ANE pour lequel l'indice de *Dutot* présente un biais positif pour un indice cible de Walsh pour deux scénarios d'échantillonnage

	ppt($\sqrt{q_y q_{y+1}}$)			*maxminq*		
	upe 2	*upe* 3	*upe* 4	*upe* 2	*upe* 3	*upe* 4
1995 – 1996	75,0	86,2	75,9	64,3	61,1	61,1
1996 – 1997	60,7	72,4	65,5	53,6	65,5	51,8
1997 – 1998	65,5	75,9	78,6	41,4	27,6	42,9
1998 – 1999	72,4	75,9	70,4	48,3	75,9	40,8
1999 – 1900	89,7	72,4	75,9	48,3	20,7	44,9

Tableau 15
Pourcentage d'ANE pour lequel le biais et l'erreur quadratique moyenne de l'indice de *Dutot* pour un indice cible de Walsh sont plus faibles pour l'échantillonnage avec probabilité proportionnelle à la taille (taille variable = $\sqrt{q_y q_{y+1}}$) que pour l'échantillonnage *maxminq*

	a) *Biais de ppt plus faible*			b) *EQM de ppt plus faible*		
	upe 2	*upe* 3	*upe* 4	*upe* 2	*upe* 3	*upe* 4
1995 – 1996	82,1	93,1	86,2	32,1	58,6	41,4
1996 – 1997	89,2	96,6	100,0	35,7	37,9	27,6
1997 – 1998	89,7	86,2	100,0	41,4	24,1	64,3
1998 – 1999	89,7	82,8	92,6	41,4	37,9	40,7
1999 – 1900	89,7	96,6	41,4	34,5	31,0	37,9

Nous concluons que les bons effets de l'échantillonnage *maxminq* combinés à l'estimateur de *Dutot* ne peuvent pas être expliqués par l'imitation approximative de l'échantillonnage *ppt*. Les comportements sont différents; et, dans l'ensemble, *maxminq* semble être un peu *meilleur* que *ppt* ($\sqrt{q_y\, q_{y+1}}$).

Nous ne voyons aucune autre raison expliquant pourquoi l'indice d'échantillon de *Dutot* devrait cibler l'indice de population de Walsh lorsque les produits les mieux vendus sont échantillonnés systématiquement, à part le mécanisme de « force brute » : dans la mesure où l'indice de Walsh peut être représenté par un petit échantillon d'articles, il est mieux représenté par ceux pour lesquels les quantités sont systématiquement les plus grandes, et ces articles sont ceux que fournit le scénario d'échantillonnage *maxminq*.

Bibliographie

Balk, B. (1999). On the use of unit values as consumer price subindices. *Proceedings of the Fourth Meeting of the International Working Group on Price Indices*, BLS, Washington, D.C.

Balk, B. (2003). Price indexes for elementary aggregates: The sampling approach. *Proceedings of the Seventh Meeting of the International Working Group on Price Indices* (*Ottawa Group*), Paris.

BLS Handbook of Methods (2005). http://stats.bls.gov/bls/descriptions.htm.

Consumer Price Indexes Technical Manual (2005). Office for National Statistics, London, http://www.statistics.gov.uk/downloads/theme_economy/CPI_Technical_Manual_2005.pdf.

De Haan, J., Opperdoes, E. et Schut, C. (1999). Le choix des produits pour l'indice des prix à la consommation : Le seuil d'inclusion par opposition au sondage probabiliste. *Techniques d'enquête*, 25, 1, 33-45.

Dalén, J. (1998). Studies on the comparability of consumer price indices. *Revue Internationale de Statistique*, 66, 1, 83-113.

Diewert, E. (1997). "Commentary" [sur 'Alternative Strategies for Aggregating Prices in the CPI' par M.D. Shapiro et D.W. Wilcox]. Federal Reserve Bank of St. Louis Review, 79, 3, 27-37.

Diewert, E. (2004). Index number theory: Past progress and future challenges. Presented at the SSHRC Conference on Price Index Concepts and Measurement, Vancouver, Canada, à http://www.econ.ubc.ca/diewert/concepts.pdf.

Dorfman, A.H., Leaver, S.G. et Lent, J. (1999). Some observations on price index estimators. *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, *Monday B Sessions*, 56-65.

Reinsdorf, M., et Triplett, J.E. (2005). A review of reviews: Ninety years of professional thinking about the consumer price index. À paraître, *Proceedings of the June 2004 NBER-CRIW Conference on Price Indexes*, Vancouver.

The Retail Prices Index Technical Manual (1998). (Éd. M. Baxter, The Stationary Office, London, à http://www.statistics.gov.uk/downloads/theme_economy/RPI_TECHNICAL_MANUAL.pdf.

Richardson, D.H. (2000). Scanner indexes for the CPI. *Proceedings of the Conference on Scanner Data and Price Indexes*, NBER, Cambridge, http://www.nber.org/books/.

Royall, R.M. (1976). Current advances in sampling theory: Implications for human observational studies. *American Journal of Epidemiology*, 104, 463-473.

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

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


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 241-257
Statistique Canada, N° 12-001 au catalogue

Une évaluation des méthodes d'échantillonnage matriciel à l'aide de données provenant de la National Health and Nutrition Examination Survey

Neal Thomas, Trivellore E. Raghunathan, Nathaniel Schenker, Myron J. Katzoff et Clifford L. Johnson [1]

Résumé

Les chercheurs et les responsables des politiques utilisent souvent des données provenant d'enquêtes par échantillonnage probabiliste représentatives de la population nationale. Le nombre de sujets couverts par ces enquêtes, et par conséquent la durée des entrevues, a généralement augmenté au fil des ans, ce qui a accru les coûts et le fardeau de réponse. Un remède éventuel à ce problème consiste à regrouper prudemment les questions d'une enquête en sous-ensembles et à demander à chaque répondant de ne répondre qu'à l'un de ces sous-ensembles. Les plans de sondage de ce type sont appelés plans à « questionnaire scindé » ou plans d'« échantillonnage matriciel ». Le fait de ne poser qu'un sous-ensemble des questions d'une enquête à chaque répondant selon un plan d'échantillonnage matriciel crée ce que l'on peut considérer comme des données manquantes. Le recours à l'imputation multiple (Rubin 1987), une approche polyvalente mise au point pour traiter les données pour lesquelles des valeurs manquent, est tentant pour analyser les données provenant d'un échantillon matriciel, parce qu'après la création des imputations multiples, l'analyste peut appliquer les méthodes standard d'analyse de données complètes provenant d'une enquête par sondage. Le présent article décrit l'élaboration et l'évaluation d'une méthode permettant de créer des questionnaires d'échantillonnage matriciel contenant chacun un sous-ensemble de questions devant être administrées à des répondants sélectionnés aléatoirement. La méthode peut être appliquée dans des conditions complexes, y compris les situations comportant des enchaînements de questions. Les questionnaires sont créés de telle façon que chacun comprenne des questions qui sont prédictives des questions exclues, afin qu'il soit possible, lors des analyses subséquentes fondées sur l'imputation multiple, de recouvrer une partie de l'information relative aux questions exclues qui aurait été recueillie si l'on n'avait pas recouru à l'échantillonnage matriciel. Ce dernier et les méthodes d'imputation multiple sont évalués au moyen de données provenant de la National Health and Nutrition Examination Survey, l'une des nombreuses enquêtes par échantillonnage probabiliste représentatives de la population nationale réalisées par le National Center for Health Statistics des Centers for Disease Control and Prevention. L'étude démontre que l'approche peut être appliquée à une grande enquête nationale sur la santé à structure complexe et permet de faire des recommandations pratiques quant aux questions qu'il serait approprié d'inclure dans des plans d'échantillonnage matriciel lors de futures enquêtes.

Mots clés : Données manquantes; imputation multiple; fardeau de réponse; questionnaire scindé; enquête par sondage.

1. Introduction

Les données provenant d'enquêtes par sondage sont utilisées par les chercheurs et les responsables des politiques dans de nombreux domaines. Souvent, ces enquêtes mettent en jeu des échantillons probabilistes représentatifs de la population nationale et une collecte à grande échelle de données au moyen de questionnaires, et doivent concilier deux objectifs concurrents, c'est-à-dire être d'une longueur et d'une complétude raisonnables tout en fournissant l'information pertinente. Le nombre de sujets couverts par ce genre d'enquêtes, et par conséquent la longueur des entrevues, ont généralement augmenté au fil des ans. L'accroissement résultant du fardeau de réponse pourrait être l'un des facteurs qui contribuent à la diminution observée des taux de réponse. Cette baisse des taux de réponse risque de réduire la précision des estimations fondées sur les données d'enquête. Elle peut aussi accroître le biais, si les différences systématiques entre les non-répondants et les répondants ne sont pas prises en compte dans les analyses des données incomplètes. En outre, l'élargissement de la gamme de sujets couverts conjugué aux efforts en vue de maintenir les taux de réponse élevés ont accru le coût de la réalisation des enquêtes.

Un moyen éventuel d'obtenir l'information nécessaire tout en limitant le fardeau de réponse consiste à regrouper prudemment les questions d'une enquête en sous-ensembles et à demander à chaque répondant de ne répondre qu'à l'un

1. Neal Thomas, Datametrics Research, Inc., 61 Dream Lake Drive, Madison, CT 06443, É.-U. Courriel : snthomas99@yahoo.com; Trivellore E. Raghunathan, Department of Biostatistics and Institute for Social Research, University of Michigan, Ann Arbor, MI 48106, É.-U. Courriel : teraghu@umich.edu; Nathaniel Schenker, Office of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782, É.-U. Courriel : nschenker@cdc.gov; Myron J. Katzoff, Office of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782, É.-U. Courriel : mkatzoff@cdc.gov; Clifford L. Johnson, Division of Health and Nutrition Examination Surveys, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782, É.-U. Courriel : cljohnson@cdc.gov.

de ces sous-ensembles. Les différents sous-ensembles de questions (items) sont administrés à différents sous-ensembles de répondants, afin que chaque question soit posée au moins à certains des répondants. Les plans d'enquête basés sur des questionnaires de ce type sont appelés plans à « questionnaire scindé » ou plans d'« échantillonnage matriciel », cette dernière expression reflétant l'idée que les répondants (lignes) et les questions (colonnes) sont les uns et les autres « échantillonnés » à partir d'une matrice conceptuelle de données sur la population complète. Dans de nombreux plans d'échantillonnage matriciel, certaines questions (appelées ici questions « communes ») sont posées à tous les répondants, tandis que d'autres (appelées ici questions « échantillonnées ») ne sont posées qu'à un sous-ensemble de répondants. Habituellement, les questions choisies comme questions communes sont soit particulièrement importantes, soit prédictives d'un grand nombre de questions échantillonnées.

Le fait de ne poser qu'un sous-ensemble des questions de l'enquête à chaque répondant dans le cas d'un plan d'échantillonnage matriciel crée ce que l'on peut considérer comme des données manquantes, de type manquant au hasard, ou même manquant entièrement au hasard (Rubin 1976), puisqu'elles sont le résultat d'un mécanisme probabiliste connu fondé éventuellement sur des variables du plan de sondage. Le recours à l'imputation multiple (Rubin 1987), une approche polyvalente mise au point pour traiter les données présentant des valeurs manquantes, est tentant pour analyser les données provenant d'un échantillon matriciel, parce qu'après la création des imputations multiples, l'analyste peut appliquer les méthodes standard d'analyse des données complètes d'enquête par sondage. De surcroît, si l'échantillon matriciel a été conçu de telle façon que les questions posées à chaque répondant soient prédictives des questions qui ne lui sont pas posées, alors, dans l'imputation multiple, il est possible d'utiliser les questions incluses pour recouvrer l'information au sujet de celles qui sont exclues. Nous nous concentrons sur l'imputation multiple, parce qu'elle est bien adaptée à cette situation : 1) l'application de méthodes multivariées complexes peut être exécutée une seule fois par l'organisme d'enquête qui connaît le mieux le plan de sondage; 2) l'imputation peut être implémentée au moyen de logiciels existants et 3) elle ne nécessite pas de nouvelles méthodes pour chacun des nombreux paramètres que ciblent la plupart des études. Néanmoins, d'autres méthodes d'estimation que l'imputation multiple, fondées sur un modèle ainsi que fondées sur un plan de sondage, peuvent être mises au point et appliquées aux données provenant de plans d'échantillonnage matriciel.

L'approche de l'échantillonnage matriciel a été appliquée ou explorée dans diverses circonstances, comme l'évaluation des acquis scolaires (Sirotnik et Wellington 1977; Beaton et Zwick 1992; Zeger et Thomas 1997), la recherche dans le domaine de la santé (Wacholder, Carroll, Pee et Gail 1994; Raghunathan et Grizzle 1995; Houseman et Milton 2006), le Recensement des États-Unis (Navarro et Griffin 1993), et l'étude des entreprises (Shoemaker 1973). En outre, un genre d'échantillonnage matriciel a été utilisé avant 1997 dans la National Health Interview Survey réalisée par le National Center for Health Statistics (NCHS) des Centers for Disease Control and Prevention. Dans cette enquête, les problèmes de santé chroniques ont été répartis en six listes et, pour chaque liste, l'information a été obtenue auprès d'environ un sixième des répondants (Schenker, Gentleman, Rose, Hing et Shimizu 2002). Par contre, dans le contexte de la collecte de données pour une enquête nationale générale sur la santé, aucune approche de création de plans d'échantillonnage matriciel exploitant les associations inhérentes entre les questions n'a été étudiée.

Le présent article décrit la mise au point et l'évaluation d'une méthode de conception de questionnaires d'échantillonnage matriciel, chacun contenant un sous-ensemble des questions devant être posées à un échantillon aléatoire de répondants. La méthode peut être appliquée dans des conditions complexes, y compris les situations où il existe des enchaînements de questions. Les questionnaires sont conçus de telle façon que chacun comprenne des questions qui sont prédictives des questions exclues, de sorte que, lors des analyses subséquentes fondées sur l'imputation multiple, il soit possible de recouvrer l'information sur les questions exclues qui aurait été recueillie si l'on n'avait pas recouru à l'échantillonnage matriciel. La méthode suppose que l'on dispose d'un échantillon d'apprentissage. Ce dernier peut provenir de l'administration antérieure d'un questionnaire complet ou d'un échantillon pilote utilisé pour appuyer la conception du questionnaire. La méthode d'échantillonnage matriciel est évaluée dans le cadre d'une étude portant sur des données provenant de la National Health and Nutrition Examination Survey (NHANES), l'une des nombreuses enquêtes représentatives de la population nationale réalisée par le NCHS (http://www.cdc.gov/nchs/nhanes. htm). La NHANES, une enquête transversale qui a été répétée plusieurs fois au cours de diverses périodes, permet de recueillir une grande quantité de données auprès des répondants au moyen d'un questionnaire sur les membres des ménages, un examen médical dans un centre d'examen mobile et l'analyse en laboratoire de prélèvements biologiques. Il est intéressant d'étudier la faisabilité de plans d'échantillonnage matriciel pour des enquêtes telles que la NHANES, qui est caractérisée par des dépendances structurelles complexes entre les questions, ce que reflètent les nombreux enchaînements, ainsi que de multiples composantes. Par souci de vraisemblance, la méthode de conception du questionnaire est appliquée à des données pilotes

provenant de la deuxième NHANES (NHANES II), puis le plan résultant et les méthodes d'imputation multiple sont évalués grâce à une étude par simulation portant sur les données de la NHANES III. La section 2 décrit la méthode de conception de questionnaires d'échantillonnage matriciel. La section 3 décrit le plan d'échantillonnage et les résultats de l'étude fondés sur la NHANES. Enfin, l'article se conclut par une discussion à la section 4.

2. Conception de questionnaires d'échantillonnage matriciel

La présente section décrit l'élaboration d'une méthode de conception de questionnaires d'échantillonnage matriciel, chacun contenant un sous-ensemble de questions destinées à être posées à un échantillon de répondants.

Lors de la conception d'un échantillon matriciel, il faut décider quelles questions seront considérées comme des questions communes qui seront incluses dans tous les questionnaires et quelles questions seront traitées comme des questions échantillonnées qui ne seront incluses que dans certains questionnaires. Habituellement, les questions communes sont sélectionnées en se basant sur un jugement de fonds et sur d'autres considérations quant à l'importance relative des questions. Les questions clés, pour lesquelles la précision de certains estimateurs doit être maximisée, devraient être traitées comme des questions communes, tandis que celles de moins grande importance peuvent être choisies comme questions échantillonnées. En outre, il est utile de sélectionner des questions communes qui sont prédictives d'un grand nombre de questions échantillonnées, afin que l'information sur les questions échantillonnées qui sont exclues d'un questionnaire puisse être récupérée d'après les questions communes combinées aux questions échantillonnées incluses dans le questionnaire. Enfin, le coût et le fardeau de réponse associés à une question sont aussi des éléments à prendre en considération, car il peut être avantageux de désigner des questions dont l'administration est coûteuse et (ou) représente un fardeau comme étant des questions échantillonnées. L'accent étant mis ici sur la façon de répartir les questions échantillonnées entre les questionnaires après que l'on ait choisi les questions communes, nous supposons que ces dernières ont déjà été sélectionnées. Cependant, nous verrons que la méthode de répartition des questions échantillonnées repose sur une mesure qui tient également compte de l'utilité des questions communes pour la prédiction des questions échantillonnées. Le pouvoir de prédire les valeurs des questions échantillonnées est estimé à l'aide d'un échantillon d'apprentissage.

Il faut aussi choisir un format pour l'organisation des questions échantillonnées. Afin de s'assurer que chaque paire possible de questions échantillonnées figure dans un questionnaire, pour que l'estimation directe de toutes les associations bidirectionnelles entre variables soit possible, les questions échantillonnées sont réparties en blocs, et des questionnaires d'échantillonnage matriciel sont créés en assemblant deux ou plusieurs blocs de questions échantillonnées (Raghunathan et Grizzle 1995). La taille des blocs et leur nombre déterminent la longueur des questionnaires et leur nombre. Par exemple, dans l'étude portant sur la NHANES dont nous discuterons à la section 3, les questions échantillonnées sont réparties en quatre blocs, et chaque questionnaire contient deux blocs (ainsi que les questions communes), de sorte qu'il en existe six en tout (combinaisons de 2 parmi 4). Dans la méthode élaborée ici, les blocs sont approximativement de taille égale, et chaque question échantillonnée est attribuée à un seul bloc. L'utilisation de blocs de même longueur donne une réduction identique du fardeau de réponse pour tous les participants à l'étude. Elle produit aussi la même précision pour les questions de même type. Notons cependant que ces caractéristiques ne sont pas obligatoires pour tous les plans d'échantillonnage matriciel. Si une estimation de plus grande précision est souhaitée pour une question, celle-ci pourrait être incluse dans plus d'un questionnaire, ou être désignée comme une question commune devant figurer dans tous les questionnaires.

Un bon plan d'échantillonnage matriciel répartit les questions échantillonnées entre les blocs de façon que, pour toute question échantillonnée exclue d'un bloc, il existe des questions échantillonnées incluses dans le bloc qui, regroupées avec les questions communes, sont prédictives de la question exclue; cela facilite le recouvrement de l'information au sujet de la question exclue à l'étape de l'analyse des données. La discussion qui suit porte sur l'élaboration d'une méthode en vue d'atteindre cet objectif. Cette élaboration comporte deux volets. Premièrement, à la section 2.1, nous formulons un indice permettant de déterminer la mesure dans laquelle chaque question échantillonnée est prédite correctement par chaque autre question échantillonnée, l'utilité prédictive étant mesurée en tant que gain relatif de précision conditionnellement à l'inclusion des questions communes. Nous présentons aussi des méthodes d'estimation des valeurs de l'indice à partir d'un échantillon d'apprentissage. Deuxièmement, à la section 2.2, nous décrivons un algorithme pour l'affectation des questions échantillonnées aux blocs d'après l'indice de valeur prédictive.

2.1 Indice de valeur prédictive

2.1.1 Notation préliminaire pour les plans d'échantillonnage matriciel

Soit Y, une question échantillonnée pouvant être prédite, $\boldsymbol{X} = (X_1, ..., X_c)$, les questions communes et Z, une question échantillonnée utilisée pour prédire Y.

Comme nous l'avons mentionné plus haut, un plan d'échantillonnage matriciel crée ce que l'on peut considérer comme des données manquantes. Donc, dans un plan d'échantillonnage matriciel possible, les sujets doivent être classés de telle sorte que les n_{obs} sujets pour lesquels il existe des valeurs observées de Y soient énumérés pour commencer, les valeurs de Y étant dénotées par Y_1, ..., $Y_{n_{\text{obs}}}$, et que les n_{mis} sujets pour lesquels les valeurs de Y manquent suivent, les valeurs de Y étant dénotées par $Y_{n_{\text{obs}}+1}$, ..., $Y_{n_{\text{tot}}}$, où $n_{\text{tot}} = n_{\text{obs}} + n_{\text{mis}}$ est le nombre total d'observations.

L'espérance et la variance de Y dans la population visée par l'échantillonnage matriciel sont dénotées par $\text{E}(Y)$ et $\text{V}(Y)$.

2.1.2 Hypothèses simplificatrices

Plusieurs hypothèses sont faites pour simplifier le calcul de l'indice. Elles sont utilisées durant ce calcul, mais non dans les analyses subséquentes des données. Toute hypothèse peut être affaiblie ou éliminée si des études supplémentaires indiquent qu'elle altère sensiblement l'évaluation des plans d'échantillonnage matriciel possibles.

1. Chaque prédicteur échantillonné Z est considéré individuellement lors de l'ajout aux questions communes $\boldsymbol{X}$. S'il existe plusieurs questions présentant une forte corrélation mutuelle, l'algorithme d'affection s'efforce d'attribuer ces questions à des blocs différents, comme le requiert un plan matriciel efficace. Une approche multivariée basée sur les corrélations partielles tenant compte d'autre questions échantillonnées produirait, en principe, des propriétés semblables, mais nécessiterait un beaucoup plus grand nombre de calculs.
2. Nous supposons que chaque prédicteur échantillonné Z est entièrement observé, alors qu'en pratique, il ne sera pas toujours possible de prédire Y, parce que Z est aussi une question échantillonnée. En outre, l'occurrence de données manquantes non prévues (c'est-à-dire, des données manquantes qui ne sont pas créées par l'échantillonnage matriciel) n'est pas pris en compte. Bien que ces hypothèses puissent donner lieu à une surestimation de l'utilité de Z pour l'amélioration des estimations de $\text{E}(Y)$, ce genre de surestimation peut être corrigé à l'aide de méthodes multivariées utilisant plusieurs variables Z. De surcroît, chaque question Z sera posée le même nombre de fois, de sorte que tout biais systématique dans la valeur prédictive devrait être approximativement le même pour chaque question échantillonnée; l'indice est utilisé principalement pour établir le classement des questions, lequel ne sera pas modifié par un biais commun.
3. Pour le calcul des valeurs de l'indice, nous supposons que les répondants ont été sélectionnés par échantillonnage aléatoire simple dans l'échantillon matriciel ainsi que dans l'échantillon d'apprentissage. De nouveau, nous ne pensons pas qu'une surestimation cohérente de la précision réduira considérablement la performance de l'indice.
4. Nous supposons que l'échantillonnage matriciel produit des données manquantes selon le mécanisme de données manquant entièrement au hasard. Cette hypothèse est satisfaite pour tous les plans d'échantillonnage matriciel pris en considération.
5. En vue de dériver les approximations qui suivent, nous supposons que n_{tot} est grand et que le ratio $n_{\text{obs}} / n_{\text{tot}}$ est constant quand n_{tot} augmente. Cette approximation devrait être adéquate pour la plupart des paramètres à estimer dans les enquêtes nationales.

2.1.3 Estimation basée sur l'imputation multiple

L'indice de valeur prédictive est établi en vue d'estimer $\text{E}(Y)$ par imputation multiple appliquée à l'échantillon matriciel. Un estimateur fondé sur l'imputation multiple, $\bar{\bar{y}}$, de $\text{E}(Y)$ est approximé, sous l'hypothèse d'un nombre infini d'imputations, par

$$\bar{\bar{y}} = \lim_{M \to \infty} M^{-1} \sum_{j=1}^{M} \bar{y}_j .$$

Dans cette expression, M est le nombre d'imputations, $\bar{y}_j$ est la moyenne calculée d'après le j^{e} ensemble de données complet avec les valeurs imputées $Y_{i,j}$, $i = n_{\text{obs}} + 1$, ..., n_{tot}, et les valeurs observées $Y_{i,j} = Y_i$, $i = 1$, ..., n_{obs} (qui ne varient pas d'un ensemble de données complet à l'autre), c'est-à-dire

$$\bar{y}_j = n_{\text{tot}}^{-1} \sum_{i=1}^{n_{\text{tot}}} Y_{i,j} = n_{\text{tot}}^{-1} \left(\sum_{i=1}^{n_{\text{obs}}} Y_i + \sum_{i=n_{\text{obs}}+1}^{n_{\text{tot}}} Y_{i,j} \right).$$

Un estimateur de la variance de $\bar{\bar{y}}$ lorsque les imputations sont créées en utilisant $\boldsymbol{X}$ et Z, en se servant de la formule courante de la variance de Rubin (1987, section 3.1), est

$$\text{V}_{\text{IM}} = \text{V}_{\text{comp}} + \text{V}_{\text{imp}}, \qquad (1)$$

où le premier terme est une estimation de la variance que l'on obtiendrait avec les données complètes et le deuxième, une estimation de la variance entre les ensembles de données imputés. Dans le cas de grands échantillons, pour lesquels la variance de Y peut être traitée comme étant connue, $\text{V}_{\text{comp}} = \text{V}(Y) / n_{\text{tot}}$, et

$$V_{imp} = \lim_{M\to\infty} M^{-1} \sum_{j=1}^{M} (\bar{y}_j - \bar{\bar{y}})^2. \qquad (2)$$

Nous supposons tout au long de l'exposé que le modèle d'imputation est compatible avec le modèle à données complètes, de sorte que l'estimateur de la variance (2) est convergent (Rubin 1987, section 3.6; Meng 1994).

2.1.4 Définition de l'indice

Lorsque des données sont recueillies auprès d'échantillons matriciels, il est possible d'obtenir des estimateurs simples, mais éventuellement inefficaces, de valeurs sommaires univariées d'une question échantillonnée, Y, à partir des données observées sans aucune imputation (autrement dit, en utilisant uniquement les valeurs observées de Y), parce que les valeurs manquantes de Y manquent entièrement au hasard; la variance de l'estimateur sans imputation de $E(Y)$ est dénotée $V_{SI} = V(Y)/n_{obs}$.

L'indice proposé est la proportion de l'écart entre V_{SI} et V_{comp} qui est recouvrée par l'estimateur sous imputation multiple, dans lequel est intégrée l'information contenue dans $\boldsymbol{X}$ et Z :

$$I(Y \mid \boldsymbol{X}, Z) = \frac{V_{SI} - V_{IM}}{V_{SI} - V_{comp}}. \qquad (3)$$

L'indice $I(Y \mid \boldsymbol{X}, Z)$ prend la valeur 1 quand $\boldsymbol{X}$ et Z prédisent parfaitement les valeurs omises de Y (de sorte que $V_{IM} = V_{comp}$), et la valeur 0 quand $\boldsymbol{X}$ et Z ne prédisent pas les valeurs omises de Y du tout, de sorte que l'estimateur en présence d'imputation multiple n'est pas une amélioration par rapport à l'estimateur sans imputation (c'est-à-dire $V_{IM} = V_{SI}$).

L'indice peut être utilisé pour évaluer la contribution éventuelle de chaque question échantillonnée Z à l'estimation de la moyenne de chaque autre question échantillonnée Y. Un plan d'échantillonnage matriciel désirable assure que, pour chaque question échantillonnée Y qui est exclue d'un bloc, il existe dans ce dernier d'autres questions échantillonnées Z dont l'indice de valeur prédictive de Y est élevé, de sorte que l'information sur Y peut être recouvrée durant les analyses des données provenant de l'échantillon matriciel.

Nota :

1. Les variances V_{SI}, V_{comp} et V_{imp} sont proportionnelles à n_{tot}^{-1}, , donc $I(Y \mid \boldsymbol{X}, Z)$ est indépendant de n_{tot}.
2. Si les questions communes $\boldsymbol{X}$ sont fortement prédictives de Y, l'indice ne fera pas de grande distinction entre les questions échantillonnées restantes Z; mais, dans cette situation, la sélection de la question échantillonnée Z appropriée pour prédire Y est moins importante, puisque Y est déjà bien prédite par $\boldsymbol{X}$.

2.1.5 Approximation de V_{imp}

Pour faciliter le calcul de l'indice $I(Y \mid \boldsymbol{X}, Z)$, il est utile d'obtenir une approximation de la variance V_{imp}. Celle qui est développée ici a trait à un plan d'échantillonnage matriciel particulier, en supposant qu'un plan ait été choisi.

Supposons que la loi de Y sachant $(\boldsymbol{X}, Z)$ suit un modèle linéaire généralisé avec une fonction de lien μ qui dépend des paramètres inconnus $\boldsymbol{\beta}$,

$$E(Y \mid \boldsymbol{X}, Z) = \mu((\boldsymbol{X}^T, Z)\boldsymbol{\beta}),$$

où la fonction de lien est égale à l'identité pour la variable continue Y, $\mu(Y) = Y$, et à la fonction logistique pour la variable binaire Y, $\mu(Y) = \text{logit}^{-1}(Y)$. Si la variable Y est continue, nous supposons aussi que la variance résiduelle, σ^2, est constante. Bien qu'elles ne le soient pas ici, des extensions de ces modèles et méthodes peuvent être élaborées pour des variables catégoriques ou catégoriques ordonnées. Les catégories individuelles peuvent être représentées par des variables binaires, ou peuvent être regroupées en catégories sommaires lorsqu'elles sont nombreuses.

Schafer et Schenker (2000) ont dérivé une approximation de la variance entre les ensembles de données imputés, c'est-à-dire V_{imp}, quand l'estimation calculée pour chaque ensemble de données complété est une fonction lisse g des moyennes des variables concernées. (Dans le cas qui nous occupe, g est l'identité.) Cette approximation, qui est basée sur des développements en série de Taylor de premier ordre de g et μ, ainsi que sur des résultats sur grand échantillon tirés de la théorie des sondages (par exemple, Wolter 1985, chapitre 6), est celle que nous utiliserons ici.

L'estimation du maximum de vraisemblance (ou de quasi-vraisemblance) (McCullagh et Nelder 1989) de $\boldsymbol{\beta}$ fondée sur les n_{obs} sujets pour lesquels des valeurs de Y sont observées donne un estimateur, $\hat{\boldsymbol{\beta}}$, dont la matrice de variance-covariance est $V_{obs}(\hat{\boldsymbol{\beta}})$ (rappelons l'hypothèse simplificatrice 4 de la section 2.1.2). Soit $\bar{\mu}_{mis}(\hat{\boldsymbol{\beta}}) \equiv n_{mis}^{-1} \sum_{i=n_{obs}+1}^{n_{tot}} \mu((\boldsymbol{X}_i^T, Z_i)\hat{\boldsymbol{\beta}})$, et dénotons sa dérivée par rapport à la j^e composante de $\boldsymbol{\beta}$ évaluée à $\hat{\boldsymbol{\beta}}$ par $\bar{\mu}'_{mis,\,j}(\hat{\boldsymbol{\beta}})$, $j = 1, \ldots, (c+1)$. La dérivée est de la forme

$$\bar{\mu}'_{mis,\,j}(\hat{\boldsymbol{\beta}}) = n_{mis}^{-1} \sum_{i=n_{obs}+1}^{n_{tot}} X_{ij}\, f(\hat{\boldsymbol{\beta}}, \boldsymbol{X}_i, Z_i) \qquad j = 1, \ldots, c$$

et

$$\bar{\mu}'_{mis,\,c+1}(\hat{\boldsymbol{\beta}}) = n_{mis}^{-1} \sum_{i=n_{obs}+1}^{n_{tot}} Z_i\, f(\hat{\boldsymbol{\beta}}, \boldsymbol{X}_i, Z_i),$$

où $f(\hat{\boldsymbol{\beta}}, \boldsymbol{X}_i, Z_i) = \mu((\boldsymbol{X}_i^T, Z_i)\hat{\boldsymbol{\beta}})[1 - \mu((\boldsymbol{X}_i^T, Z_i)\hat{\boldsymbol{\beta}})]$ quand Y est binaire. Si Y est continue, $f(\hat{\boldsymbol{\beta}}, \boldsymbol{X}_i, Z_i) = 1$,

ce qui implique que les dérivées $\bar{\mu}'_{\text{mis},\,j}(\hat{\boldsymbol{\beta}})$ sont égales aux moyennes des questions communes $\boldsymbol{X}$ et de la question échantillonnée Z.

Maintenant, représentons par $\bar{\mu}'_{\text{mis}}(\hat{\boldsymbol{\beta}})$ le vecteur des dérivées et par P_{mis} la proportion de sujets pour lesquels des valeurs de Y manquent. En appliquant l'équation (10) de Schafer et Schenker (2000), avec leur fonction g égale à l'identité, et leur paramètre général θ égal à β, nous obtenons

$$\mathrm{V}_{\text{imp}} \approx P_{\text{mis}}^2 \left[\begin{array}{l} n_{\text{mis}}^{-1} \left\{ n_{\text{mis}}^{-1} \displaystyle\sum_{i=n_{\text{obs}}+1}^{n_{\text{tot}}} \mu((\boldsymbol{X}_i^T, Z_i)\,\hat{\boldsymbol{\beta}}) \right. \\ \qquad\qquad \left. [1 - \mu((\boldsymbol{X}_i^T, Z_i)\,\hat{\boldsymbol{\beta}})] \right\} \\ + (\bar{\mu}'_{\text{mis}}(\hat{\boldsymbol{\beta}}))^T\, \mathrm{V}_{\text{obs}}(\hat{\boldsymbol{\beta}})\, \bar{\mu}'_{\text{mis}}(\hat{\boldsymbol{\beta}}) \end{array} \right] \quad (4)$$

si Y est binaire, et

$$\mathrm{V}_{\text{imp}} \approx P_{\text{mis}}^2 [\sigma^2 / n_{\text{mis}} + (\bar{\mu}'_{\text{mis}}(\hat{\boldsymbol{\beta}}))^T\, \mathrm{V}_{\text{obs}}(\hat{\boldsymbol{\beta}})\, \bar{\mu}'_{\text{mis}}(\hat{\boldsymbol{\beta}})] \quad (5)$$

si Y est continue.

2.1.6 Estimation de l'indice de valeur prédictive d'après un échantillon d'apprentissage

Puisque nous supposons que les données manquantes prévues dans nos plans d'échantillonnage matriciel manquent entièrement au hasard, les moments d'échantillon et les estimations d'autres paramètres provenant d'un échantillon d'apprentissage peuvent être utilisés pour estimer les moments et paramètres correspondants dans les sous-échantillons contenant des valeurs observées de Y, ainsi que dans ceux contenant des valeurs manquantes de Y, sous l'hypothèse que l'échantillon d'apprentissage est tiré à partir de la même population cible. Les moments et paramètres incluent : $\mathrm{V}(Y)$; la variance résiduelle σ^2, qui peut être estimée par $\hat{\sigma}^2_{\text{tr}}$, la variance résiduelle estimée d'après la régression ajustée à l'échantillon d'apprentissage; les coefficients de régression, avec les estimations $\hat{\boldsymbol{\beta}}_{\text{tr}}$ provenant de l'échantillon d'apprentissage; et la matrice de variance-covariance des coefficients de régression, qui peuvent être approximés par rééchelonnement de l'estimation $\mathrm{V}_{\text{tr}}(\hat{\boldsymbol{\beta}}_{\text{tr}})$ provenant de l'échantillon d'apprentissage pour obtenir $\mathrm{V}_{\text{obs}}(\hat{\boldsymbol{\beta}}) \approx (n_{\text{tr}} / n_{\text{obs}})\mathrm{V}_{\text{tr}}(\hat{\boldsymbol{\beta}}_{\text{tr}})$, où n_{tr} est la taille de l'échantillon d'apprentissage. Les dérivées $\bar{\mu}'_{\text{mis}}(\hat{\boldsymbol{\beta}})$ et la fonction faisant intervenir μ dans (4) sont également sous la forme de moyennes de sous-échantillon et peuvent donc être estimées par les moyennes correspondantes dans l'échantillon d'apprentissage. En dénotant les dérivées dans l'échantillon d'apprentissage par $\bar{\mu}'_{\text{tr}}(\hat{\boldsymbol{\beta}}_{\text{tr}})$, et en introduisant les estimateurs provenant de l'échantillon d'apprentissage par substitution dans (4) et (5), nous obtenons

$$\mathrm{V}_{\text{imp}} \approx P_{\text{mis}}^2 \left[\begin{array}{l} n_{\text{mis}}^{-1} \left\{ n_{\text{tr}}^{-1} \displaystyle\sum_{i=1}^{n_{\text{tr}}} \mu((\boldsymbol{X}_i^T, Z_i)\,\hat{\boldsymbol{\beta}}_{\text{tr}}) \right. \\ \qquad\qquad \left. [1 - \mu((\boldsymbol{X}_i^T, Z_i)\,\hat{\boldsymbol{\beta}}_{\text{tr}})] \right\} \\ + \dfrac{n_{\text{tr}}}{n_{\text{obs}}} (\bar{\mu}'_{\text{tr}}(\hat{\boldsymbol{\beta}}_{\text{tr}}))^T\, \mathrm{V}_{\text{tr}}(\hat{\boldsymbol{\beta}}_{\text{tr}})\, \bar{\mu}'_{\text{tr}}(\hat{\boldsymbol{\beta}}_{\text{tr}}) \end{array} \right] \quad (6)$$

pour les variables binaires, et

$$\mathrm{V}_{\text{imp}} = P_{\text{mis}}^2 (\hat{\sigma}^2_{\text{tr}} / n_{\text{mis}} + \hat{\sigma}^2_{\text{tr}} / n_{\text{obs}}), \quad (7)$$

pour les variables continues, cette dernière expression découlant du fait que $(\bar{\mu}'_{\text{tr}}(\hat{\boldsymbol{\beta}}_{\text{tr}}))^T \mathrm{V}_{\text{tr}}(\hat{\boldsymbol{\beta}}_{\text{tr}})\, \bar{\mu}'_{\text{tr}}(\hat{\boldsymbol{\beta}}_{\text{tr}})$ se réduit à la forme simple $\hat{\sigma}^2_{\text{tr}} / n_{\text{tr}}$.

2.2 Affectation des questions échantillonnées aux blocs

2.2.1 Critères de conception des questionnaires

Les questionnaires d'échantillonnage matriciel sont créés en affectant les questions échantillonnées à divers blocs, comme il est décrit au début de la section 2. Quatre objectifs de conception orientent l'affectation des questions : 1) affecter chaque question échantillonnée à un seul bloc; 2) affecter un nombre approximativement égal de questions à chaque bloc; 3) affecter à un même bloc les questions reliées logiquement et 4) affecter à chaque bloc une ou plusieurs questions qui prédisent les questions omises dans le bloc. Dénotons le nombre de blocs par n_{block} ($n_{\text{block}} = 4$ dans l'étude par simulation de la NHANES).

Un critère quantitatif pour le quatrième objectif est spécifié séparément pour chaque question échantillonnée Y en trouvant les $(n_{\text{block}} - 1)$ autres questions échantillonnées Z possédant les valeurs d'indice prédictif $I(Y \mid \boldsymbol{X}, Z)$ les plus élevées, pour l'affectation éventuelle de l'une d'entre elles aux $(n_{\text{block}} - 1)$ blocs ne contenant pas Y. Les questions Z ne comprennent pas celles liées à la question Y, qui doivent figurer avec cette dernière dans un bloc. Les $(n_{\text{block}} - 1)$ valeurs de $I(Y \mid \boldsymbol{X}, Z)$ pour les questions Z fournissent une borne supérieure sur les indices prédictifs qui pourraient être réalisés pour Y. Comme ces valeurs d'indice optimales sont déterminées séparément pour chaque question échantillonnée Y, elles pourraient ne pas être réalisables simultanément pour toutes les questions Y.

Pour évaluer un plan d'échantillonnage matriciel particulier, nous déterminons la valeur de l'indice $I(Y \mid \boldsymbol{X}, Z)$ la plus élevée effectivement obtenue pour chacun des $(n_{\text{block}} - 1)$ blocs ne contenant pas de question échantillonnée Y. Puis, nous calculons la moyenne des $(n_{\text{block}} - 1)$ écarts entre ces indices et les indices prédictifs optimaux correspondants pour Y. Enfin, nous calculons la moyenne de ces écarts moyens sur l'ensemble des questions échantillonnées Y pour obtenir une mesure globale pour le plan.

2.2.2 Un algorithme d'affectation

Les critères de la section 2.2.1 doivent être maximisés sur un ensemble d'entrées entières (affectations aux blocs) d'une fonction, sous un ensemble de contraintes linéaires imposées par la nécessité de créer des blocs de longueur approximativement égale avec des questions éventuellement reliées. Bien que les méthodes de programmation entières puissent être appliquées à cette maximisation, l'algorithme qui suit est nettement plus simple et donne des résultats presque optimaux pour l'application de la NHANES, comme nous le démontrons à la section 3.1.

Étape 1. Ordonner aléatoirement les questions échantillonnées. L'affectation des questions aux blocs se fait séquentiellement, par répétition des étapes 2 et 3 qui suivent, jusqu'à ce que toutes les questions aient été affectées.

Étape 2. Affecter la prochaine (ou la première) question non affectée, disons $Y^{(0)}$, au bloc contenant le plus petit nombre de questions. Si plusieurs blocs sont à égalité, affecter $Y^{(0)}$ à celui pour lequel l'indice prédictif maximum $I(Y^{(0)} \mid \boldsymbol{X}, Z)$ est le plus faible pour $Y^{(0)}$. S'il persiste un ex æquo, affecter $Y^{(0)}$ à n'importe lequel des blocs admissibles. S'il existe des questions liées à $Y^{(0)}$, les affecter également au bloc sélectionné.

Étape 3. Pour chaque question affectée à l'étape 2 ($Y^{(0)}$ ou les questions qui y sont liées), trouver les questions non affectées restantes, disons $Y^{(1)}$, qui en sont les plus prédictives. Affecter $Y^{(1)}$ (et toute question liée à $Y^{(1)}$) à un autre bloc que celui sélectionné à l'étape 2, en suivant la même procédure que celle utilisée pour $Y^{(0)}$ à l'étape 2.

L'expérience avec les données de la NHANES donne à penser que l'algorithme est moyennement sensible au classement initial des questions (à l'étape 1). Pour réduire la dépendance, nous avons généré 1 000 plans d'expérience avec classements sélectionnés aléatoirement, puis nous avons choisi celui donnant la meilleure mesure globale de la valeur prédictive (telle qu'elle est définie à la section 2.2.1).

3. Étude au moyen de données provenant de la NHANES

Pour évaluer la faisabilité d'un plan d'échantillonnage matriciel pour une enquête telle que la NHANES, nous avons réalisé une étude d'évaluation. Pour commencer, nous avons utilisé la NHANES II (c'est-à-dire, la deuxième NHANES) pour créer un plan d'échantillonnage matriciel par la méthode décrite à la section 2. Ce plan simule la situation réelle dans laquelle les données provenant d'une enquête antérieure sont utilisées pour concevoir le questionnaire d'une nouvelle enquête. Puis, nous avons appliqué le plan ainsi établi à plusieurs échantillons simulés créés d'après les données de la NHANES III. Les participants à la NHANES III pour lesquels existaient des données complètes pour un ensemble sélectionné de variables ont été traités comme une grande population finie. Nous avons tiré 100 échantillons à partir de la population finie de la NHANES III selon un plan d'échantillonnage stratifié à deux degrés avec probabilités de sélection inégales. Les données complètes étant disponibles pour chacun des échantillons simulés, ceux-ci constituent notre « étalon ». Nous avons ensuite imposé le plan d'échantillonnage matriciel à chaque échantillon et procédé à l'imputation multiple des valeurs manquantes dues à l'échantillonnage matriciel. Enfin, nous avons réalisé plusieurs analyses en utilisant les échantillons matriciels sans imputation, les échantillons matriciels en présence d'une imputation multiple et les échantillons avec données complètes (c'est-à-dire l'étalon). Les résultats résumés sur l'ensemble des échantillons simulés produisent des estimations des propriétés des diverses méthodes sous échantillonnage répété.

Le plan d'échantillonnage matriciel créé en utilisant les données de la NHANES II est résumé à la section 3.1. Le plan de l'étude par simulation au moyen des données de la NHANES III est décrit à la section 3.2. Les résultats de l'étude sont présentés à la section 3.3. Certaines limites de l'étude qui ne sont pas abordées aux sections 3.1 à 3.3 sont décrites à la section 3.4.

3.1 Plan d'échantillonnage matriciel fondé sur les données d'apprentissage provenant de la NHANES II

Étant donné le temps qu'aurait pris l'extraction et l'analyse de toutes les variables de la NHANES III, nous n'avons inclus qu'un sous-ensemble dans l'étude afin que celle-ci demeure gérable, quoique le logiciel utilisé permettrait de traiter un beaucoup plus grand nombre de variables. Les variables retenues pour l'étude comprennent les questions représentant nombre des sujets couverts dans l'enquête et ont été sélectionnées en consultation avec les spécialistes du domaine. Les types de données comprennent des variables binaires et continues représentant des questions d'enquête et des mesures de laboratoire. Une paire de questions formant un enchaînement a été incluse : « Avez-vous fumé 100 cigarettes ou plus au cours de votre vie? » suivie de « Fumez-vous à l'heure actuelle? » L'algorithme utilisé pour affecter les questions échantillonnées aux blocs, décrit à la section 2.2, a forcé ces deux questions à être dans le même bloc.

Le tableau 1 donne une brève description des variables incluses. Les variables qui figuraient dans la NHANES III mais non dans la NHANES II (de nouveau, une situation réaliste) sont indiquées par un astérisque à côté de leur nom.

Comme nous l'avons mentionné plus haut, le plan d'échantillonnage matriciel a été construit avec quatre blocs.

Chacun contenait toutes les questions communes. En outre, les questions échantillonnées qui figuraient dans la NHANES II ont été affectées aux blocs en appliquant les méthodes élaborées à la section 2 aux données provenant de la NHANES II. (Lors de l'estimation des indices nécessaires, nous avons traité la question des valeurs manquantes dans les données de la NHANES II en analysant uniquement les cas complets.) Les questions échantillonnées qui ne figuraient pas dans la NHANES II ont été réparties et affectées aléatoirement aux blocs de façon à ce que la longueur de ceux-ci demeure approximativement égale. La colonne « Type » du tableau 1 indique quelles sont les variables communes et les variables échantillonnées, ainsi que le bloc affecté pour les variables échantillonnées.

Pour chaque question échantillonnée qui figurait dans la NHANES II, le tableau 2 donne les renseignements suivants : le bloc auquel la question a été affectée (« Bloc », les trois indices prédictifs les plus élevés pour les autres questions échantillonnées en tant que prédicteurs de la question concernée (« Optimaux ») et les valeurs les plus élevées de l'indice effectivement réalisées au moyen du plan sélectionné dans les trois blocs ne contenant pas la question en question (« Réalisés »). Les valeurs de l'indice sont triées de la plus faible à la plus élevée pour chaque question considérée, de sorte que les colonnes du tableau contenant les valeurs d'indice ne correspondent pas à des questions ni à des blocs particuliers. Le tableau 2 montre que le plan choisi est presque optimal pour le critères de la section 2.2.1. Par exemple, l'écart moyen entre les indices prédictifs optimaux et les indices correspondants effectivement réalisés n'est que de 0,002.

La colonne du tableau 2 étiquetée « Inférieur » sous « Réalisés » donne les bornes inférieures sur l'amélioration attendue des estimateurs des moyennes univariées pour les questions échantillonnées. Dix-neuf des 21 indices prédictifs figurant dans cette colonne sont inférieurs à 0,20, ce qui donne à penser que l'efficacité des estimateurs multi-imputés est relativement faible dans ce plan d'échantillonnage matriciel. Pour une discussion plus approfondie de cette question, voir les sections 3.3 et 4.

Tableau 1
Variables de la NHANES III qui ont été incluses dans l'évaluation.
Les questions marquées d'un astérisque ne figuraient pas dans la NHANES II

Nom de la variable	Description de la variable	Type
BMPBMI	Indice de masse corporelle	Commune
CHP*	Cholestérol sérique (mg/dL)	Commune
DMARETHN	Race-ethnicité	Commune
DMPCREGN	Région de recensement, pondération (Texas dans le Sud)	Commune
DMPMETRO	Code de région rurale/urbaine basé sur le code Usda	Commune
GHP*	Hémoglobine glycatée (%)	Commune
HAB1	Votre santé est-elle, en général, excellente, ..., mauvaise	Commune
HAB2*	Se rend dans un endroit particulier pour obtenir des soins de santé	Commune
HAB5*	Au cours des 12 derniers mois, nombre de visites chez un docteur	Commune
HAC1C	Un docteur a déclaré : insuffisance cardiaque congestive	Commune
HAC1L*	Un docteur vous a-t-il déjà dit que vous aviez : lupus	Commune
HAC1M	Un docteur vous a-t-il déjà dit que vous aviez : goutte	Commune
HAD1	Vous a-t-on déjà dit que vous aviez du sucre/diabète	Commune
HAD10	Pour le moment, prenez-vous des pilules contre le diabète	Commune
HAE3	On vous a dit au moins deux fois que vous faisiez de l'hypertension	Commune
HAF10	Un docteur vous a-t-il déjà dit que vous aviez fait une crise cardiaque	Commune
HAF26	Étourdissement grave pendant plus de 5 minutes	Commune
HAL1	Toux la plupart des jours, 3 mois consécutifs ou plus dans l'année	Commune
HAL6	A eu des sifflements dans la poitrine au cours des 12 derniers mois	Commune
HAL14E	Symptômes déclenchés par le pollen	Commune
HAZMNK1R	Pa K1 moyenne d'après questionnaire-ménage et CEM	Commune
HAZMNK5R	Pa K5 moyenne d'après questionnaire-ménage et CEM	Commune
HFA12	État matrimonial	Commune
HFA8R	Grade le plus élevé ou nombre d'années d'école achevées	Commune
HSAGEIR	Âge au moment de l'entrevue (questionnaire de présélection)	Commune
HSSEX	Sexe	Commune
I1P	Insuline sérique (uU/mL)	Commune
G1P	Glucose sérique (mg/dL)	Échantillonnée – 1
HAC1J	Un docteur vous a-t-il déjà dit que vous aviez : goitre	Échantillonnée – 1
HAC1N*	Un docteur vous a-t-il déjà dit que vous aviez : cancer de la peau	Échantillonnée – 1
HAC1O	Un docteur vous a-t-il déjà dit que vous aviez : autre cancer	Échantillonnée – 1
HAF14*	Ressent de la douleur dans les deux jambes en marchant	Échantillonnée – 1
HAL11A	Lourdeur, chatouillement ou écoulement nasal au cours des 12 derniers mois	Échantillonnée – 1

Tableau 1 (suite)
Variables de la NHANES III qui ont été incluses dans l'évaluation.
Les questions marquées d'un astérisque ne figuraient pas dans la NHANES II

Nom de la variable	Description de la variable	Type
BMPWHR*	Ratio tour de taille-tour de hanche	Échantillonnée – 2
HAC1E	Un docteur vous a-t-il déjà dit que vous aviez : asthme	Échantillonnée – 2
HAC1K	Un docteur vous a-t-il déjà dit que vous aviez : maladie thyroïdienne	Échantillonnée – 2
HAF24	Engourdissement, etc., d'un côté du visage/corps pendant plus de 5 minutes	Échantillonnée – 2
HAL11B	Yeux larmoyants, qui chatouillent au cours des 12 derniers mois	Échantillonnée – 2
HAL19A*	Au cours des 12 derniers mois, a eu : rhume ou grippe	Échantillonnée – 2
HAL19C*	Au cours des 12 derniers mois, a eu : pneumonie	Échantillonnée – 2
HAT28	Actif(ve) comparativement aux hommes/femmes de votre âge	Échantillonnée – 2
PBP	Plomb (ug/dL)	Échantillonnée – 2
SPPFVC*	CVF, valeur la plus grande (mL)	Échantillonnée – 2
FEP	Fer sérique (ug/dL)	Échantillonnée – 3
HAF1	A déjà eu une douleur ou une gène dans la poitrine	Échantillonnée – 3
HAF23	Faiblesse/paralysie dans le visage, le bras, la jambe pendant plus de 5 minutes	Échantillonnée – 3
HAL19B	Au cours des 12 derniers mois, a eu : sinusite/problème de sinus	Échantillonnée – 3
HAR1	Avez-vous fumé 100 cigarettes ou plus au cours de votre vie	Échantillonnée – 3
HAR3	Fumez-vous des cigarettes à l'heure actuelle	Échantillonnée – 3
SPPPEAK*	Débit maximal expiratoire	Échantillonnée – 3
BDPTOBMD*	Densité minérale osseuse, région totale (g/cm^2)	Échantillonnée – 4
HAB4	Au cours des 12 derniers mois, nombre d'hospitalisations	Échantillonnée – 4
HAC1D	Un docteur vous a-t-il déjà dit que vous aviez : accident vasculaire cérébral	Échantillonnée – 4
HAC1F	Un docteur vous a-t-il déjà dit que vous aviez : bronchite chronique	Échantillonnée – 4
HAC1H	Un docteur vous a-t-il déjà dit que vous aviez : rhume des foins	Échantillonnée – 4
HAC1I	Un docteur vous a-t-il déjà dit que vous aviez : cataracte	Échantillonnée – 4
HAE6*	A-t-on déjà vérifié votre cholestérol sanguin	Échantillonnée – 4
HAM11*	Considère que son poids est excessif/insuffisant/correct	Échantillonnée – 4
HAE7*	Un docteur a dit que le taux de cholestérol sanguin était élevé	Échantillonnée – 4

Tableau 2
Indices de valeur prédictive basés sur les données de la NHANES II pour les questions échantillonnées dans le plan d'échantillonnage matriciel

		Optimaux			Réalisés		
Question	Bloc	Inférieur	Moyen	Supérieur	Inférieur	Moyen	Supérieur
HAC1J(GOITRE)	1	0,04	0,04	0,15	0,04	0,04	0,15
HAC1O(AUTRE CANCER)	1	0,05	0,06	0,13	0,05	0,06	0,13
HAL11A(SYMPTÔMES NASAUX)	1	0,17	0,27	0,29	0,17	0,27	0,29
G1P(GLUCOSE SÉRIQUE)	1	0,26	0,30	0,43	0,26	0,30	0,43
HAC1E(ASTHME)	2	0,09	0,10	0,13	0,08	0,09	0,13
HAC1K(MALADIE THYROÏDIENNE)	2	0,07	0,07	0,15	0,07	0,07	0,15
HAF24(ENGOURDISSEMENT)	2	0,12	0,12	0,12	0,11	0,12	0,12
HAL11B(YEUX LARMOYANTS)	2	0,14	0,15	0,25	0,14	0,15	0,25
HAT28(ACTIF(VE) POUR SON ÂGE)	2	0,12	0,13	0,16	0,11	0,13	0,16
PBP(PLOMB (ug/dL))	2	0,19	0,20	0,21	0,18	0,20	0,21
HAF1(DOULEUR DANS LA POITRINE)	3	0,25	0,29	0,29	0,23	0,25	0,29
HAF23(FAIBLESSE/PARALYSIE)	3	0,08	0,12	0,12	0,08	0,12	0,12
HAL19B(SINUSITE/SINUS)	3	0,07	0,12	0,21	0,07	0,12	0,21
HAR1(100 CIGARETTES OU PLUS)	3	0,13	0,14	0,14	0,13	0,14	0,14
HAR3(FUME À L'HEURE ACTUELLE)	3	0,10	0,11	0,12	0,10	0,11	0,12
FEP(FER SÉRIQUE)	3	0,05	0,05	0,08	0,05	0,05	0,08
HAB4(NOMBRE D'HOSPITALISATIONS)	4	0,07	0,11	0,19	0,07	0,11	0,19
HAC1D(ACCIDENT VASCULAIRE CÉRÉBRAL)	4	0,19	0,20	0,24	0,18	0,20	0,24
HAC1F(BRONCHITE)	4	0,10	0,12	0,12	0,10	0,10	0,12
HAC1H(RHUME DES FOINS)	4	0,07	0,07	0,09	0,04	0,07	0,09
HAC1I(CATARACTE)	4	0,08	0,09	0,12	0,08	0,09	0,12

Nota : Les indices prédictifs optimaux sont déterminés pour chaque question séparément et ne sont pas nécessairement réalisables pour toutes les questions simultanément.

3.2 Conception de l'étude par simulation fondée sur les données de la NHANES III

3.2.1 Population et plan d'échantillonnage

Le plan d'échantillonnage matriciel et l'analyse en présence d'imputation multiple pourraient être appliqués à l'échantillon complet de la NHANES III. Cela serait certes informatif, mais une étude basée sur un seul ensemble de données ne permettrait pas d'évaluer les propriétés statistiques des méthodes étudiées sous échantillonnage répété. Par conséquent, les 11 759 sujets de la NHANES III qui avaient fourni des données complètes sur les variables énumérées au tableau 1 ont été traités comme une population finie et des échantillons répétés ont été tirés à partir de cette population. Nous avons utilisé pour sélectionner les échantillons un plan d'échantillonnage complexe plutôt qu'un échantillonnage aléatoire simple afin de créer une étude par simulation plus réaliste. Pour réaliser cet objectif, nous avons ajouté trois variables de plan d'échantillonnage à la population finie, à savoir 1) strate de simulation, 2) grappe de simulation et 3) poids d'échantillonnage de simulation (ici, le qualificatif « de simulation » est utilisé pour faire la distinction entre ces quantités et les variables du plan d'échantillonnage original de la NHANES III).

1. **Strates de simulation :** L'échantillon du fichier de données à grande diffusion de la NHANES III comporte 49 strates contenant chacune 2 grappes. La stratégie, pour l'étude en simulation, consistait à créer un plus petit nombre de strates contenant chacune un plus grand nombre de grappes, pour s'assurer que la variation d'un échantillon à l'autre entre les échantillons simulés soit suffisante. Nous avons regroupé les 49 strates originales en 20 strates de simulation de la façon suivante. Chacune des 49 strates originales a été classée dans l'une de huit catégories formées par recoupement de la région de recensement (quatre niveaux) et de la situation de région rurale/urbaine basée sur le code du United States Department of Agriculture (deux niveaux). Dans chacune de ces huit catégories, nous avons procédé à une analyse typologique en utilisant les proportions de non-Blancs au niveau de la strate pour sélectionner les strates originales qu'il convenait de combiner. La combinaison des strates originales a créé deux ou trois strates de simulation dans chacune des huit catégories, ce qui a donné en tout 20 strates de simulation. Cette méthode de création de strates plus grandes a également augmenté l'hétérogénéité raciale entre les strates de simulation résultantes, ce qui a accru l'importance de la pondération dans les analyses.

2. **Grappes de simulation :** L'échantillon du fichier de données à grande diffusion de la NHANES III comporte 98 grappes, c'est-à-dire 2 grappes dans chacune des 49 strates originales. Après avoir regroupé les 49 strates originales en 20 strates de simulation, nous avons réparti les grappes originales en nous fondant sur une autre analyse typologique axée sur les lectures de la pression artérielle systolique et diastolique et sur l'indice de masse corporelle (IMC). Les sujets pour lesquels les valeurs étaient similaires ont été regroupés afin de créer des conditions de corrélation intraclasse pour ces trois variables dans chaque grappe de simulation. Le nombre de grappes de simulation par strate de simulation variait de 3 à 25, et le nombre de sujets par grappe de simulation variait de 30 à 98.

3. **Poids d'échantillonnage de simulation :** Nous avons déterminé les poids d'échantillonnage de simulation au moyen du plan d'échantillonnage à deux degrés suivant. Premièrement, pour chaque strate de simulation, nous avons tiré deux grappes de simulation par échantillonnage aléatoire simple sans remise. Comme les nombres de grappes de simulation dans les 20 strates de simulation étaient inégaux, le poids d'échantillonnage de simulation correspondant à cette étape était $w_{1h} = A_h / 2, h = 1, 2, ..., 20,$ où A_h est le nombre de grappes de simulation dans la strate de simulation h.
   Deuxièmement, à partir de chaque grappe de simulation sélectionnée, nous avons tiré 30 sujets au hasard sans remise avec des probabilités de sélection variables. Si la taille de la grappe était égale à 30, alors tous les sujets ont été inclus dans l'échantillon. Pour les grappes contenant plus de 30 sujets, nous avons calculé les probabilités de sélection de premier tirage en normalisant les inverses des poids originaux provenant de l'échantillon du fichier de données à grande diffusion de la NHANES III de sorte que leur somme soit égale à 1 dans chaque grappe de simulation, l'inverse normalisé étant utilisé pour chaque sujet comme probabilité de sélection de ce sujet. Les probabilités de sélection de premier tirage dans les grappes de simulation variaient de 0,0003 à 0,2756.
   Soit i l'indice représentant les sujets échantillonnés dans une grappe de simulation, c les grappes échantillonnées dans une strate de simulation et h les strates de simulation telles qu'elles sont décrites plus haut, $i = 1, 2, ..., 30,$ $c = 1, 2,$ $h = 1, 2, ..., 20.$ Si la taille de la grappe c dans la strate h était égale à 30, alors le poids d'échantillonnage de simulation de deuxième degré pour le sujet i dans la grappe c

était $w_{2\text{ich}} = 1$. Si la taille c dans la strate h était supérieure à 30, alors le poids d'échantillonnage de simulation de deuxième degré pour le sujet i dans la grappe c était $w_{2\text{ich}} \propto \pi_{\text{ich}}^{-1}$, où π_{ich} représente la probabilité de sélection de premier tirage du sujet i. Pour chaque sujet échantillonné, le poids d'échantillonnage de simulation final était $w_{\text{ich}} = w_{1h} \times w_{2\text{ich}}$, $i = 1, 2, ..., 30, c = 1, 2, h = 1, 2, ..., 20$.

Les effets de plan pour l'estimation des moyennes de population étaient de l'ordre de 2,1, en moyenne, dans cette étude par simulation. Les caractéristiques du plan d'échantillonnage complexe utilisées dans l'étude sont informatives en ce sens qu'ignorer les caractéristiques du plan dans les analyses des données pourrait produire des estimations biaisées et une sous-estimation des variances d'échantillonnage. Cela est dû en particulier à l'utilisation de données sur la race, la pression artérielle et l'IMC dans le plan d'échantillonnage de simulation et au lien bien établi entre la race ou l'ethnicité et la pression artérielle ou l'IMC.

3.2.2 Simulation d'échantillons matriciels

Nous avons tiré 100 échantillons probabilistes indépendants à partir de la population finie. Chaque échantillon simulé comptait 1 200 sujets (20 strates de simulation, 2 grappes de simulation par strate de simulation, 30 sujets par grappe de simulation).

L'échantillonnage matriciel a été superposé à chaque échantillon simulé en affectant chacun des 1 200 sujets aléatoirement à l'un des six questionnaires contenant les questions communes et l'une des paires de blocs (1, 2), (1, 3), (1, 4), (2, 3), (2, 4) ou (3, 4). L'affectation aléatoire a été exécutée de façon que 200 sujets soient affectés à chaque questionnaire. Donc, pour chaque échantillon matriciel, les questions communes étaient disponibles pour l'ensemble des 1 200 sujets échantillonnés, tandis que chaque question échantillonnée était disponible pour 600 sujets échantillonnés.

3.2.3 Comparaison des méthodes d'estimation

Nous avons obtenu des estimations ponctuelles d'après chaque échantillon de l'étude en simulation selon trois méthodes, à savoir l'analyse des données complètes à titre d'étalon, l'analyse des données échantillonnées matriciellement sans imputation et l'application de l'imputation multiple pour remplacer les valeurs manquantes causées par l'échantillonnage matriciel, suivies de l'analyse des données multi-imputées. Pour l'analyse des données complètes et des données échantillonnées sans imputation, les estimations ponctuelles ont été pondérées. Pour les analyses en présence d'imputation multiple, nous avons utilisé les mêmes poids pour calculer l'estimation ponctuelle à partir de chacun des ensembles de données complétés par imputation multiple, puis nous avons calculé la moyenne habituelle des estimations ponctuelles en présence d'imputation multiple (Rubin et Schenker 1986; Rubin 1987, section 3.1).

L'imputation multiple des valeurs des questions échantillonnées manquantes a été exécutée par la méthode de régression séquentielle (Kennickell 1991; Oudshoorn, Van Buuren et Van Rijckevorsel 1999; Raghunathan, Lepkowski, Van Hoewyk et Solenberger 2001), telle qu'elle est implémentée dans le progiciel IVEware (http://www.isr.umich.edu/src/smp/ive). Nous avons créé cinq ensembles d'imputations en appliquant indépendamment la méthode de régression séquentielle cinq fois, avec dix itérations de l'algorithme de régression séquentielle pour chaque ensemble d'imputations. Le nombre d'imputations est basé sur la théorie et l'expérience indiquant que cinq imputations sont habituellement suffisantes, surtout si la fraction d'information manquante n'est pas importante (Rubin 1996). Pour des taux de données manquantes de 50 % pour les questions échantillonnées, la fraction d'information manquante, qui est approximativement $1 - V_{\text{comp}} / V_{\text{imp}}$, devrait, en principe, être au plus de 50 %, ce que confirment les résultats des simulations. Selon Rubin (1987, tableau 4.1), l'efficacité relative de cinq imputations sur grand échantillon comparativement à un nombre infini d'imputations est de 90 % lorsque 50 % d'information manque. Un grand nombre d'imputations augmenterait la précision de l'estimation de la variance entre imputations (V_{imp}) et la fraction d'information manquante.

Pour tenir compte du plan d'échantillonnage simulé complexe, les effets principaux ont été inclus dans le modèle d'imputation pour la strate de simulation et la grappe de simulation emboîtée dans la strate de simulation. Le logarithme du poids d'échantillonnage de simulation a également été inclus comme prédicteur dans le modèle d'imputation, ainsi que les questions communes et les questions échantillonnées.

3.3 Résultats de l'étude par simulation

Pour évaluer les estimations basée sur le plan d'échantillonnage matriciel, nous avons considéré deux types de problèmes analytiques, à savoir l'estimation des moyennes de population des questions échantillonnées et les analyses par la régression faisant intervenir les questions échantillonnées et les questions communes. Nous avons comparé les propriétés des estimateurs sans imputation, en présence d'imputation multiple et pour les données complètes sur l'ensemble des 100 ensembles de données simulées, afin d'évaluer le biais et la perte d'efficacité dus à l'échantillonnage matriciel conjugué à l'imputation multiple.

3.3.1 Estimation des moyennes de population pour les questions échantillonnées

Pour la moyenne de population d'une question échantillonnée, nous avons défini le biais standardisé simulé de

l'estimateur sans imputation comme étant $(\text{Moy.}_{\text{SI}} - \text{Moy.}_{\text{comp}}) / \text{É.-T.}_{\text{SI}}$, où Moy._{SI}, $\text{Moy.}_{\text{comp}}$ et É.-T._{SI} dénotent, respectivement, les moyennes des estimations sans imputation et sur les données complètes, ainsi que l'écart-type des estimations sans imputation sur l'ensemble des 100 ensembles de données simulées. Nous avons défini un biais standardisé simulé analogue pour l'estimateur en présence d'imputation multiple (IM). Le tableau 3 résume les biais standardisés simulés pour les 32 questions échantillonnées.

Tableau 3
Biais standardisés simulés des estimateurs sans imputation et en présence d'imputation multiple des moyennes de population pour les 32 questions échantillonnées

Biais standardisé	Fréquence	
	Sans imputation	Imputation multiple
– 1,4		1
(– 1, – 0,6]		4
(– 0,6, – 0,4]		5
(– 0,4, – 0,2]		4
(– 0,2, 0)	15	10
(0, 0,2)	17	4
[0,2, 0,4)		
[0,4, 0,6)		2
[0,6, 1)		
1,4		1
4,6		1
Total	32	32

Puisque notre mécanisme d'échantillonnage matriciel produit des données manquantes qui manquent entièrement au hasard, les estimateurs sans imputation sont presque sans biais, ce qui est reflété dans les résultats des simulations par le fait qu'aucun biais standardisé absolu n'est supérieur à 0,2. Les estimateurs en présence d'imputation multiple sont généralement affectés d'un biais standardisé simulé un peu plus élevé que les estimateurs sans imputation, quoique le biais standardisé absolu soit inférieur à 1 pour toutes les questions échantillonnées sauf trois et inférieur à 0,6 pour toutes, sauf sept. En guise de ligne directrice pour évaluer les biais standardisés, Cochran (1977, page 14) montre qu'un biais standardisé de 0,6 produit des intervalles de confiance à 95 % nominaux dont la couverture réelle est d'environ 91 %. Tout biais important observé dans le cadre de la présente étude lorsqu'on utilise l'échantillonnage matriciel conjugué à l'imputation multiple est vraisemblablement dû à des défauts dans les modèles d'imputation et non à l'échantillonnage matriciel proprement dit, puisque nous avons constaté que les analyses sans imputation étaient approximativement sans biais. En utilisant de plus grandes tailles d'échantillon dans une application à une enquête réelle, les biais standardisés correspondants auraient tendance à subir un mouvement à la hausse, à cause des dénominateurs plus faibles; mais ils pourraient aussi subir un mouvement à la baisse, à cause de l'amélioration des approximations sur grand échantillon.

La perte d'efficacité due à l'échantillonnage matriciel plutôt qu'à l'utilisation d'un questionnaire complet peut être évaluée en comparant l'erreur d'échantillonnage des estimateurs en l'absence d'imputation, en présence d'imputation multiple et avec données complètes (calculée comme étant l'écart-type sur les 100 ensembles de données simulées). Le tableau 4 résume les ratios des écarts-types simulés des estimateurs en présence d'imputation multiple à ceux des estimateurs sans imputation, et les ratios des écarts-types simulés des estimateurs sur données complètes à ceux des estimateurs en présence d'imputation multiple (nous utilisons l'expression « écart-type simulé » d'un estimateur plutôt que « erreur-type simulée » pour éviter la confusion avec l'erreur-type estimée que l'on pourrait obtenir d'après l'analyse de chaque ensemble de données simulé).

Tableau 4
Ratios des écarts-types simulés des estimateurs sans imputation (SI), en présence d'imputation multiple (IM) et sur les données complètes (comp) des moyennes de population pour les 32 questions échantillonnées

Ratios	Fréquence	
	$\text{É.-T.}_{\text{IM}}/\text{É.-T.}_{\text{SI}}$	$\text{É.-T.}_{\text{comp}}/\text{É.-T.}_{\text{IM}}$
(0,5, 0,6]		2
(0,6, 0,7]		9
(0,7, 0,8]		14
(0,8, 0,9]		6
(0,9, 0,95]	7	
(0,95, 1]	18	1
(1, 1,03]	7	
Total	32	32

Habituellement, les estimateurs en présence d'imputation multiple sont plus efficaces que les estimateurs sans imputation, mais le gain d'efficacité n'est que moyen, comme l'indique le fait que la plupart des ratios $\text{É.-T.}_{\text{IM}}/\text{É.-T.}_{\text{SI}}$ du tableau 4 sont compris entre 0,9 et 1. Des gains aussi modestes d'efficacité peuvent être prédits grossièrement d'après les indices de valeur prédictive basés sur les données provenant de la NHANES II (présentées au tableau 2), comme il suit. Puisque chaque question échantillonnée n'est incluse que dans la moitié des questionnaires d'échantillonnage matriciel, il s'ensuit que la variance d'un estimateur basé sur les données complètes de la moyenne d'une question échantillonnée devrait être égale à environ la moitié de la variance de l'estimateur sans imputation correspondant. Diviser le numérateur et le dénominateur de l'expression (3) par V_{SI}, et fixer $V_{comp} / V_{SI} = 0,5$, donne $2(1 - V_{IM} / V_{SI})$ comme expression approximative de l'indice de valeur prédictive dans cette étude en simulation. Pour un indice de 0,12, qui est la médiane des indices « réalisés moyens » du tableau 2, il s'ensuit que V_{IM} / V_{SI} devrait être d'environ 0,94. Ce ratio des variances est équivalent à un ratio des écarts-types d'environ $\sqrt{0,94} = 0,97$, qui est proche du milieu de la fourchette des ratios résumés au tableau 4. Dans la présente étude, parce que les estimateurs en présence

d'imputation multiple ne sont que modérément plus efficaces que les estimateurs sans imputation, et que les estimateurs en présence d'imputation multiple sont affectés d'un certain biais, leurs erreurs quadratiques moyennes sont plus grandes que celles des estimateurs sans imputation dans 22 des 32 cas.

Les résultats des simulations en ce qui a trait à l'efficacité relative des estimateurs en présence d'imputation multiple comparativement aux estimateurs sur données complètes concordent aussi avec la théorie. Puisque V_{comp} / V_{SI} devrait être égal à environ 0,5, et que V_{IM} devrait être un peu plus faible que V_{SI}, il s'ensuit que V_{comp} / V_{IM} devrait être légèrement supérieur à 0,5, ou de façon équivalente, que le ratio typique des écarts-types standardisés É.-T.$_{comp}$/É.-T.$_{IM}$ devrait être un peu plus grand que $\sqrt{0,5} = 0,71$. En effet, la médiane des ratios résumés au tableau 4 est 0,75. Une alternative à l'estimation en présence d'imputation multiple est la pondération en deux phases basée sur les estimateurs des questions communes et leurs différences entre les blocs. Tout gain d'efficacité comparativement à l'estimation en présence d'imputation multiple serait dû à l'information supplémentaire provenant des questions échantillonnées.

3.3.2 Estimation des coefficients de régression

Nous avons également évalué les méthodes d'échantillonnage matriciel et d'imputation multiple dans le cas de l'estimation des coefficients de huit modèles de régression, que nous avons spécifiés de sorte qu'ils soient similaires aux modèles décrits dans la littérature. Les modèles de régression, qui sont énumérés au tableau 5, comptaient, en tout, 115 coefficients. Les estimateurs sans imputation pour les coefficients de régression n'ont pas été inclus dans l'étude par simulation, mais nous discutons de certains résultats théoriques relatifs à leur efficacité à la présente section.

Pour chaque coefficient de régression, nous adoptons une définition du biais standardisé simulé analogue à celle utilisée pour chaque moyenne à la section 3.3.1. Le tableau 6 résume les biais standardisés pour les 115 coefficients de régression. La plupart de ces biais sont faibles, les valeurs absolues n'étant supérieures à 1 que pour cinq d'entre eux, et égales ou supérieures à 0,6 pour sept seulement.

Le tableau 7 résume les ratios des écarts-types des estimations avec les données complètes sur les 100 ensembles de données simulés à ceux des estimations en présence d'imputation multiple, pour les 115 coefficients de régression. Des résumés distincts sont présentés selon que les modèles de régression contiennent des variables échantillonnées provenant d'un seul bloc (modèles 1, 2, 6 et 7) ou de deux blocs (modèles 3, 4, 5 et 8). Une plus forte proportion de ratios s'approchent de 1 que dans le cas de l'estimation des moyennes (tableau 4). En outre, pour plusieurs coefficients de régression (particulièrement provenant des modèles 3, 6, 7 et 8), les écarts-types simulés des estimateurs sur les données complètes sont modérément plus grands que ceux des estimateurs en présence d'imputation multiple, et pour un coefficient, le ratio est d'environ 2. Enfin, il existe quatre coefficients de régression pour lesquels il semble y avoir une perte importante d'efficacité due à l'échantillonnage matriciel, les ratios étant inférieurs à 0,3 (provenant, respectivement, des modèles 1, 2, 5 et 8). Les ratios proches de l'unité ou supérieurs à celle-ci pourraient être dus en partie à un mauvais ajustement de certains modèles de régression sur les données complètes et à un meilleur ajustement des modèles sur les données complétées par imputation, ce dernier résultant d'un processus d'imputation fondé sur des modèles de régression. En outre, les deux ratios les plus faibles sont observés pour des modèles de régression contenant des variables échantillonnées provenant de deux blocs, pour lesquelles la fraction de sujets dans l'échantillon matriciel sans données manquantes est seulement un sixième, comme nous en discutons plus loin.

Pour les modèles de régression comprenant des variables échantillonnées ne provenant que d'un seul bloc, l'efficacité théorique de l'estimateur sur données complètes relativement à l'estimateur sans imputation, c'est-à-dire le ratio de la variance du second à celle du premier, est approximativement égale à 2, parce qu'il n'existera des données complètes sur ces variables que pour approximativement la moitié des sujets compris dans l'échantillon matriciel; pour les modèles de régression contenant des variables échantillonnées provenant de deux blocs, l'efficacité théorique relative est d'environ 6. Par contre, les efficacités relatives simulées respectives de l'estimateur sur données complètes comparativement à l'estimateur en présence d'imputation multiple, c'est-à-dire les inverses des carrés des ratios résumés au tableau 7, sont inférieurs à 2 pour 64 des 69 coefficients lorsqu'un seul bloc de questions entre en jeu et ils sont inférieurs à 6 pour 44 des 46 coefficients lorsque deux blocs sont utilisés. Donc, les estimateurs en présence d'imputation multiple sont généralement plus efficaces que les estimateurs sans imputation pour les problèmes de régression. Néanmoins, les importantes pertes d'efficacité des estimateurs en présence d'imputation multiple relativement aux estimateurs sur données complètes observées pour certains coefficients, ainsi que les gains apparents d'efficacité pour d'autres coefficients justifieraient une étude plus poussée.

Tableau 5
Modèles de régression utilisés dans l'évaluation

Type de modèle de régression	Variable dépendante	Variables recodées pour créer des prédicteurs incluant des termes d'interaction. Chaque modèle comprend aussi un terme d'ordonnée à l'origine. Pour chaque variable, le chiffre entre parenthèses indique le nombre de coefficients de régression associés à la variable	Variables échantillonnées dans les modèles de régression. Pour chaque variable, le chiffre entre parenthèses indique le bloc contenant la variable
1. Linéaire	G1P	HSSEX(1) ,HSAGEIR(1), DMARETHN(3), et GHP(1)	G1P(1)
2. Logistique	HAF10	HSSEX(1), HSAGEIR(1), DMARETHN(3), FEP(1), et BMPBMI(1)	FEP(3)
3. Logistique	HAF10	HSSEX(1), HSAGEIR(1), DMARETHN(3), HAD1(1), HAE3(1), PBP(1), FEP(1), CHP(1), et G1P(1)	FEP(3) et G1P(1)
4 et 5. Linéaire	SPPFVC	HSAGEIR(1), DMARETHN (3), HFA8R(2), et BMPBMI(1) [Selon le sexe (HSSEX), et limité aux personnes n'ayant jamais fumé (HAR1, HAR3)]	SPPFVC(1), HAR1(3), et HAR3(3)
6 et 7. Logistique	HCHP (1 SI CHP>=240 ET 0 SINON)	HSAGEIR(2), DMARETHN(3), HFA8R(1), BMPBMI(3), (HAR3,HAR1)(2), BMPBMI*HSAGEIR(6), et DMARETHN*BMPBMI(9) [Selon le sexe (HSSEX)]	(HAR3, HAR1)(3)
8. Logistique	HAC1E	HSAGEIR(5), HSSEX(1), DMARETHN(3), BMPBMI(4), (HAR3, HAR1)(2), SPPPEAK(1), et SPPFVC(1)	HACIE(2), (HAR1,HAR3)(3), SPPPEAK(3), et SPPFVC(2)

Tableau 6
Biais standardisés simulés des estimateurs en présence d'imputation multiple pour les 115 coefficients de régression

Fourchette des biais standardisés	Fréquence
– 5,2	1
– 1,5	1
– 1,3	1
– 1,1	1
(– 1, -0,6]	2
(– 0,6, – 0,4]	2
(– 0,4, – 0,2]	3
(– 0,2, 0)	52
(0, 0,2)	44
[0,2, 0,4)	6
[0,4, 0,6)	1
[0,6, 1)	
3,7	1
Total	115

Tableau 7
Ratios des écarts-types simulés des estimateurs sur les données complètes à ceux des estimateurs en présence d'imputation correspondants, pour les 115 coefficients de régression, selon que les modèles de régression comportent des variables échantillonnées provenant d'un seul bloc ou de deux blocs

	Fréquence	
Fourchette des ratios	Un bloc	Deux blocs
(0, 0,1]		1
(0,1, 0,2]		1
(0,2, 0,3]	2	
(0,3, 0,4]		
(0,4, 0,5]	1	
(0,5, 0,6]		3
(0,6, 0,7]	2	3
(0,7, 0,8]	2	7
(0,8, 0,9]	4	4
(0,9, 0,95]	2	3
(0,95, 1]	29	8
(1, 1,05]	20	5
(1,05, 1,1]	4	2
(1,1, 1,2]	3	2
(1,2, 1,4]		4
(1,4, 1,6]		2
2,0		1
Total	69	46

3.4 Limites supplémentaires de l'étude par simulation

À la présente section, nous discutons brièvement de certaines limites supplémentaires de l'étude par simulation et des ajustements qui ont été requis durant sa mise en œuvre.

Au départ, deux questions au sujet de deux problèmes de santé, la goutte et le lupus (HAC1M et HAC1L) ont été désignées comme questions échantillonnées. À cause de la faible prévalence de ces deux problèmes de santé dans la population finie construite, un grand nombre des échantillons simulés ne contenaient pas de sujets qui en étaient atteints. Après quelques passages préliminaires, nous avons modifié la désignation de ces deux questions et en avons fait des questions communes. En général, dans les situations où les tailles d'échantillon sont limitées, il pourrait être nécessaire de considérer les questions sur les problèmes de santé dont la prévalence est très faible comme des questions communes. En outre, à cause de problèmes tels que l'existence de questions échantillonnées dans la NHANES III, mais pas dans la NHANES II, ainsi que l'existence de liens logiques entre certaines questions échantillonnées, le nombre de ces dernières par bloc dans l'étude de simulation variait un peu plus que prévu (de 6 à 10).

Dans les modèles de régression énumérés au tableau 3, le nombre de prédicteurs variaient de 8 à 27, parce que certains modèles comprenaient des termes d'interaction en tant que prédicteur. Même avec une taille de 1 200 pour les échantillons simulés, certains estimateurs sur données complètes étaient instables. Cela était dû en partie à la petite taille d'échantillon pour certaines combinaisons de variables qui affectaient l'estimation des interactions. Il convient de souligner que, dans de nombreuses applications de l'échantillonnage matriciel à de grandes enquêtes, les tailles d'échantillon avec données complètes seraient sensiblement plus grandes que celle de 1 200 utilisée dans notre étude par simulation.

Les erreurs-types de Monte Carlo des moyennes simulées dans la présente étude sont égales au dixième environ des écarts-types des quantités individuelles sur l'ensemble des 100 échantillons. Cependant, les écarts-types sur l'ensemble des échantillons variaient fortement d'un paramètre estimé à l'autre, à cause de différences d'échelonnage. Par exemple, les écarts-types simulés des estimateurs sur données complètes des 115 coefficients de régression variaient de $9{,}8 \times 10^{-5}$ à 1 169,6. Des estimations plus précises du biais et de l'efficacité pourraient être calculées en se fondant sur un plus grand nombre d'échantillons simulés que celui utilisé dans l'étude.

4. Discussion

Le présent article décrit l'élaboration d'une méthode en vue de créer des plans d'échantillonnage matriciel ayant la

propriété de produire des questionnaires contenant des questions qui sont prédictives de celles qui ont été exclues. La faisabilité de l'application de plans de ce genre à une enquête sur la santé complexe, à grande échelle, est démontrée au moyen d'un exemple portant sur la National Health and Nutrition Examination Survey. Il est possible d'utiliser des plans d'échantillonnage matriciel, conjugués à l'imputation multiple, pour étendre la portée d'une enquête, sans accroître le fardeau de réponse, ni sans augmenter excessivement le fardeau subséquent des analystes des données.

Dans l'étude portant sur les données de la NHANES, les analyses en présence d'imputation multiple des données provenant des échantillons matriciels étaient modérément efficaces, les preuves de biais étant mineures et l'efficacité, plus grande que celle de l'analyse des données provenant des échantillons matriciels uniquement, sans imputation. Le gain d'efficacité est particulièrement évident dans le contexte des analyses de régression.

Cependant, dans l'exemple fondé sur la NHANES, l'échantillonnage matriciel entraînait généralement une grande perte de précision comparativement aux résultats que l'on aurait obtenus au moyen d'un questionnaire complet, plus long (c'est à dire sans échantillonnage matriciel). Cette constatation, qui est en contradiction avec les résultats plus prometteurs obtenus dans les autres applications de l'échantillonnage matriciel, fait ressortir combien il est important d'intégrer de bons prédicteurs des questions échantillonnées dans un questionnaire. Par exemple, une application de l'échantillonnage matriciel à une enquête sur les acquis scolaires (par exemple, Beaton et Zwick 1992) a donné de nettement meilleurs résultats, parce que les questions échantillonnées étaient fortement corrélées à des questions conçues pour mesurer le même trait. Raghunathan et Grizzle (1995) ont également donné la preuve d'un recouvrement beaucoup plus important de l'information sur les questions omises dans le contexte d'une enquête sur la santé.

Dans l'exemple de la NHANES, les questions ont été choisies principalement en vue de représenter une gamme de caractéristiques importantes de la santé, sans trop tenir compte de leur capacité à prédire ou à être prédites par d'autres variables. Un grand nombre de questions échantillonnées correspondaient à des maladies rares qui ne sont pas bien prédites par les problèmes de santé courants et les mesures standard de laboratoire; rétrospectivement, ces variables n'étaient pas de bonnes candidates pour les questions échantillonnées. Les variables représentant des événements rares peuvent aussi poser des difficultés dans le cas de nombreuses méthodes statistiques courantes qui s'appuient sur des approximation en grand échantillon, de sorte qu'elles se prêtent moins bien à l'imputation fondée sur un modèle.

Les « panels » de questions interdépendantes sont de meilleurs candidats à l'échantillonnage matriciel. Par exemple, les techniques d'échantillonnage matriciel peuvent être utilisées dans des situations où sont faites des mesures multiples des mêmes quantités (ou de quantités étroitement liées), et où il est souhaitable de recueillir certaines mesures auprès de sous-ensembles des répondants à l'enquête à cause de contraintes de coûts et de temps. Certaines formes rudimentaires d'échantillonnage matriciel sont déjà appliquées dans de telles circonstances et des améliorations sensibles pourraient être réalisées en appliquant des méthodes, telles que celle élaborée dans le présent article, qui visent à exploiter les associations entre les variables.

Remerciements

Les travaux de Neal Thomas et de Trivellore E. Raghunathan ont été financés en partie par un contrat de services professionnels conclu par le NCHS et Datametrics Research, Inc. Les auteurs remercient Randy Curtin du NCHS pour ses conseils utiles. Les résultats et les conclusions présentés dans le présent article sont ceux des auteurs et ne représentent pas forcément les opinions du National Center for Health Statistics des Centers for Disease Control and Prevention.

Bibliographie

Beaton, A., et Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics*, 17, 95-109.

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

Houseman, E., et Milton, D. (2006). Partial questionnaire designs, questionnaire nonresponse, and attributable fraction: Applications to adult onset asthma. *Statistics in Medicine*, 25, 1499-1519.

Kennickell, A.B. (1991). Imputation of the 1989 Survey of Consumer Finances: Stochastic relaxation and multiple imputation. *Proceedings the Survey Research Methods Section*, American Statistical Association, 112-121.

McCullagh, P., et Nelder, J. (1989). *Generalized Linear Models*. Deuxième édition, London : Chapman Hall.

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

Navarro, A., et Griffin, R. (1993). Matrix sampling designs for the year 2000 Census. *Proceedings the Survey Research Methods Section*, American Statistical Association, 480-485.

Oudshoorn, K., Van Buuren, S. et Van Rijckevorsel, J. (1999). Flexible multiple imputation by chained equations of the AVO-95 Survey. Leiden: TNO Prevention and Health, Rapport PG/VGZ/99.045.

Raghunathan, T.E., et Grizzle, J.E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90, 54-63.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J. et Solenberger, P. (2001). Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression. *Techniques d'enquête*, 27, 91-103.

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

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

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

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

Schafer, J.L., et Schenker, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association*, 95, 144-154.

Schenker, N., Gentleman, J.F., Rose, D., Hing, E. et Shimizu, I.M. (2002). Combining estimates from complementary surveys: A case study using prevalence estimates from national health surveys of households and nursing homes. *Public Health Reports*, 117, 393-407.

Shoemaker, D.M. (1973). *Principles and Procedures of Matrix Sampling*. Cambridge, MA : Ballinger.

Sirotnik, K., et Wellington, R. (1977). Incidence sampling: An integrated theory for matrix sampling. *Journal of Educational Measurement*, 14, 343-399.

Wacholder, S., Carroll, R.J., Pee, D. et Gail, M.H. (1994). The partial questionnaire design for case-control studies (avec discussion). *Statistics in Medicine*, 13, 623-649.

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

Zeger, L.M., et Thomas, N. (1997). Efficient matrix sampling for correlated latent traits: Examples from the National Assessment of Educational Progress. *Journal of the American Statistical Association*, 92, 416-425.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont fourni de l'aide ou ont fait la critique d'un article ou plus durant l'année 2006.

J.-F. Beaumont, *Statistique Canada*
Y. Berger, *The University of Reading*, *UK*
C. Boudreau, *Medical College of Wisconsin*
L. Burck, *Central Bureau of Statistics*, *Israël*
F. Butar, *Sam Houston University*
J. Chipperfield, *Australian Bureau of Statistics*
S.R. Chowdhury, *Westat Inc.*
G. Datta, *University of Georgia*
P. Duchesne, *Université de Montréal*
K. Duncan, *Dominican University*, *Chicago*
F. Dupont, *INSEE*
G.B. Durrant, *Southampton Statistical Sciences Research Institute*, *University of Southampton*, *UK*
M. Elliott, *University of Michigan*
J. Eltinge, *United States Bureau of Labor Statistics*
M. Feder, *Research Triangle Institute*
R. Folsom, *Research Triangle Institute*
O. Frank, *Stockholm University*
J. Gambino, *Statistique Canada*
C. Girard, *Statistique Canada*
M. Gosh, *University of Florida*
B. Graubard, *National Cancer Institute*
G. Griffiths, *Australian Bureau of Statistics*
D. Haziza, *Statistique Canada*
J. Horgan, *Dublin City University*
V.G. Iannacchione, *RTI International*
J. Jiang, *University of California at Davis*
J.-K. Kim, *Department of Applied Statistics*, *Korea*
P. Kokic, *Australian Bureau of Agriculture and Resource Economics*
P. Kott, *United States Department of Agriculture*
M. Kovačevič, *Statistique Canada*
F. Kreuter, *Joint Program in Survey Methodology*
M.D. Larsen, *Iowa State University*
P. Lavallée, *Statistique Canada*
H. Lee, *Westat*, *Inc*.
R. Lehtonen, *University of Helsinki*
C. Leon, *Statistique Canada*
W.W. Lu, *Department of Mathematics and Statistics*
A. Matei, *Université de Neuchâtel*, *Suisse*
D. Melec, *United States Bureau of the Census*
J.M. Montaquila, *Westat*, *Inc*.
R. Munnich, *University of Tubingen*
J. Opsomer, *Iowa State University*
Z. Patak, *Statistique Canada*
D. Pfeffermann, *Israël and University of Southampton*
N. Prasad, *University of Alberta*
L. Qualité, *Université de Neuchâtel*, *Suisse*
M.G. Ranalli, *Universita' degli Studi di Perugia*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
O. Sautory, *Insee-Cepe*
J. Schafer, *Pennsylvania State University*
A. Scott, *University of Auckland*
R. Singh, *U.S. Census Bureau*
C. Skinner, *University of Southampton*
E. Stuart, *Mathematica Policy Research Inc*
C.J. Swartz, *Simon Fraser University*
R. Valliant, *University of Michigan*
Z. Wang, *Wilfrid Laurier University, Waterloo*
M. Winglee, *Westat*
C. Wu, *University of Waterloo*
W. Yung, *Statistique Canada*
E. Zanutto, *Department of Statistics, The Wharton School*, *University of Pennsylvania*

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2006: Cécile Bourque, Louise Demers, Anne-Marie Fleury, Roberto Guido, Liliane Lanoie, Micheal Pelchat et Isabelle Poliquin (Division de la diffusion), Nadine Lacroix (Division des services à la clientèle), Sheri Buck (Division du développement de systèmes), François Beaudin (Division des langues officielles et traduction) et Sophie Chartier (Division des méthodes d'enquêtes auprès des entreprises). Finalement nous désirons exprimer notre reconnaissance à Christine Cousineau, Céline Ethier et Denis Lemire de la Division des méthodes d'enquêtes auprès des ménages, pour leur apport à la coordination, la dactylographie et la rédaction.


PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 22, No. 1, 2006

Volume 22, No. 2, 2006

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 34, No. 2, June/juin 2006

CONTENTS TABLE DES MATIÈRES

Volume 34, No. 3, September/septembre 2006

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, N° 1) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word. Une version papier pourrait être requise pour les formules et graphiques.

1. Présentation

1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.

1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.

1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.

1.4 Les remerciements doivent paraître à la fin du texte.

1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.

3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) *etc*.

3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.

3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.

3.5 Distinguer clairement les caractères ambigus (comme w, ω ; o, O, 0; l, 1).

3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).

4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple: Cochran (1977, page 164).

5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, *etc.* à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

6. Communications brèves

6.1 Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots.