C3

STATISTICS CANADA STATISTIQUE CANADA
JUL 7 1994
LIBRARY BIBLIOTHÈQUE

TECHNIQUES D' ENQUÊTE

Catalogue 12-001

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

Juin 1994

•

VOLUME 20

•

NUMÉRO 1

Statistique Canada Statistics Canada

Canada

TECHNIQUES D' ENQUÊTE

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

JUIN 1994 • VOLUME 20 • NUMÉRO 1

Publication autorisée par le ministre
responsable de Statistique Canada

© Ministre de l'Industrie, des Sciences
et de la Technologie, 1994

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistrement sur support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division de la commercialisation, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Juin 1994

Prix : Canada : 45 $
États-Unis : 50 $ US
Autres pays : 55 $ US

N° 12-001 au catalogue

ISSN 0714-0045

Ottawa

10101684 98
STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA

Statistique Canada Statistics Canada

Canada

SQQ 80

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

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

COMITÉ DE DIRECTION

Président G.J. Brackstone

Membres

B.N. Chinnappa
G.J.C. Hole
F. Mayda (Directeur de la Production)
R. Platek (Ancien président)
C. Patrick
D. Roy
M.P. Singh

COMITÉ DE RÉDACTION

Rédacteur en chef M.P. Singh, *Statistique Canada*

Rédacteurs associés

D.R. Bellhouse, *University of Western Ontario*
D. Binder, *Statistique Canada*
M.J. Colledge, *Statistique Canada*
J.-C. Deville, *INSEE*
J.D. Drew, *Statistique Canada*
W.A. Fuller, *Iowa State University*
M. Gonzalez, *U.S. Office of Management and Budget*
R.M. Groves, *U.S. Bureau of the Census*
D. Holt, *University of Southampton*
G. Kalton, *University of Michigan*
A. Mason, *East-West Center*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
I. Sande, *Bell Communications Research, U.S.A.*
C.-E. Särndal, *Université de Montréal*
W.L. Schaible, *U.S. Bureau of Labor Statistics*
F.J. Scheuren, *U.S. Internal Revenue Service*
J. Sedransk, *State University of New York*
J. Waksberg, *Westat Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

Rédacteurs adjoints N. Laniel, M. Latouche, L. Mach et H. Mantel, *Statistique Canada*

POLITIQUE DE RÉDACTION

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

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociales, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de Techniques d'enquête (n° 12-001 au catalogue) est de 45 $ par année au Canada, 50 $ (É.-U.) aux États-Unis, et de 55 $ (É.-U.) par année à l'étranger. Prière de faire parvenir votre demande d'abonnement à Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada
Volume 20, numéro 1, juin 1994

TABLE DES MATIÈRES

Dans ce numéro 1

Estimation pour petites régions

M.P. SINGH, J. GAMBINO et H.J. MANTEL
Les petites régions: problèmes et solutions 3
Commentaires: W.A. FULLER 16
G. KALTON 19
Réponse des auteurs 22

D. HOLT et D.J. HOLMES
Estimation pour petits domaines dans des plans de sondage avec probabilités inégales 25

A.C. SINGH, H.J. MANTEL et B.W. THOMAS
MPLSE à données chronologiques pour petites régions évalués à l'aide de données d'enquête 35

J.G. KOVAR et E.J. CHEN
Méthode du jackknife pour l'estimation de la variance en présence de données imputées 47

D.S. TRACY et S.S. OSAHAN
Estimation pour grappes chevauchantes lorsque la taille de la population est inconnue 57

N.G.N. PRASAD et J.E. GRAHAM
Échantillonnage avec PPT en deux occasions 63

R.R. SITTER et C.J. SKINNER
Stratification multidimensionnelle par programmation linéaire 69

W.A. FULLER, M.M. LOUGHIN et H.D. BAKER
Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988 79

E.A. STASNY, B.G. TOOMEY et R.J. FIRST
Estimation du taux de sans-abri dans les régions rurales: une étude des régions non urbaines de l'Ohio 91

Techniques d'enquête, juin 1994
Vol. 20, n° 1, p. 1
Statistique Canada

Dans ce numéro

Ce numéro de *Techniques d'enquête* commence par une section spéciale sur **l'estimation pour petites régions**. Les trois articles de cette section traitent le problème de l'estimation pour domaine dans des perspectives différentes. Je voudrais remercier tout particulièrement Jon Rao d'avoir coordonné la rédaction de cette section spéciale. Un ou deux autres articles sur le même sujet, qui n'étaient pas prêts au moment de la publication, pourront apparaître dans une prochaine édition.

Le premier article de la série, de Singh, Gambino et Mantel, examine la question des statistiques régionales du point de vue de la conception des plans de sondage. Les auteurs étudient comment certains aspects du plan de sondage, par exemple la stratification, la formation de grappes et la répartition de l'échantillon, influent sur la production de données régionales pour des domaines planifiés ou non planifiés. Ils font aussi un survol des méthodes d'estimation pour petites régions actuellement en usage. L'article est suivi des commentaires pénétrants de MM. Fuller et Kalton et d'une réponse des auteurs.

L'article de Holt et Holmes présente une méthode d'estimation pour petites régions fondée sur un modèle qui ne permet pas l'"emprunt d'information" à d'autres domaines et qui peut être utilisée quand on ne possède pas d'information supplémentaire comme des totaux ou des moyennes de population. Les estimations des paramètres de modèle sont combinées avec des estimations (fondées sur un plan) de moyennes ou de totaux de covariables. En se servant d'une étude de marché comme exemple, les auteurs montrent que la méthode proposée peut amener des gains d'efficacité appréciables dans les estimations pour petits domaines.

Le dernier article de la section spéciale, de Singh, Mantel et Thomas, présente une comparaison empirique de divers estimateurs pour petites régions faite à l'aide d'un échantillonnage simulé appliqué à une population de fermes. Les auteurs montrent que, pour les enquêtes à passages répétés, les estimateurs fondés sur un modèle chronologique peuvent – du point de vue tant du biais que de l'erreur quadratique moyenne – être plus efficaces que les estimateurs fondés sur un modèle s'appliquant à une seule période.

Kovar et Chen présentent les résultats d'une étude de simulation dans laquelle ils ont analysé les propriétés statistiques d'une méthode "jackknife" d'estimation de la variance pour des ensembles de données imputées. Selon cette méthode, la variance due à l'imputation est intégrée dans l'estimateur de la variance. Les auteurs ont recours à des ensembles de données réelles, à quatre méthodes d'imputation, à l'échantillonnage aléatoire simple et à un mécanisme de non-réponse uniforme. Ils étudient aussi l'efficacité de l'estimation suivant un plan à plusieurs degrés stratifié et un mécanisme de non-réponse non uniforme.

Tracy et Osahan proposent des estimateurs par quotient pour deux méthodes d'échantillonnage qui servent à estimer la moyenne d'une population en grappes chevauchantes lorsque la taille de la population est inconnue. De nombreux travaux de recherche ont été faits sur les grappes non chevauchantes, mais, dans la pratique, il y a beaucoup de cas où l'échantillonnage se fait avec des grappes chevauchantes. La première méthode proposée consiste en un échantillonnage avec remise avec probabilités égales, tandis que la seconde consiste en un échantillonnage avec probabilités inégales.

Prasad et Graham étendent la "méthode des groupes aléatoires" pour un échantillonnage avec probabilité proportionnelle à la taille (PPT) à un échantillonnage effectué en deux occasions. À cette fin, ils se servent de l'information relative à une variable observée à la première occasion pour prélever la portion appariée de l'échantillon à la deuxième occasion.

Sitter et Skinner montrent comment, dans les problèmes de stratification multidimensionnelle, on peut obtenir un plan de sondage optimal au moyen de la programmation linéaire. Ils comparent leur méthode avec les méthodes existantes, d'une part en examinant les plans de sondage produits pour des applications particulières et, d'autre part, en évaluant les erreurs quadratiques moyennes. L'estimation de la variance est également examinée.

Fuller, Loughin et Baker étudient la production de poids de régression en situation de non-réponse. Ils exposent les conditions dans lesquelles l'estimateur par régression demeure convergent en situation de non-réponse et examinent les facteurs qui peuvent déterminer le choix des variables explicatives. Pour illustrer leurs propos, les auteurs se servent de la Nationwide Food Consumption Survey de 1987-1988, réalisée par le Human Nutrition Information Service du Département de l'agriculture des É.-U.

Enfin, l'article de Stasny, Toomey et First décrit une enquête effectuée en 1990 dans le but d'estimer le taux de sans-abri dans les régions rurales de l'Ohio. Les auteurs évaluent la taille probable du biais de l'estimateur en simulant un échantillonnage dans diverses populations synthétiques. Ils constatent que le biais sera plutôt négligeable par rapport à l'écart type.

Le rédacteur en chef

Techniques d'enquête, juin 1994
Vol. 20, n° 1, pp. 3–23
Statistique Canada

Les petites régions: problèmes et solutions

M.P. SINGH, J. GAMBINO et H.J. MANTEL[1]

RÉSUMÉ

Dans cet article, nous discutons de questions techniques relatives à la fourniture de données régionales tirées de recensements, de fichiers administratifs et d'enquêtes. Bien qu'il s'agisse de questions d'ordre général, nous les analysons en relation avec des programmes de Statistique Canada. Nous faisons ressortir la nécessité d'élaborer une approche globale pour les estimations d'enquête et nous soulignons, dans la perspective du remaniement d'une enquête- ménages, les aspects de la conception de plans de sondage qui ont une incidence sur les données régionales. Enfin, nous faisons un survol des méthodes d'estimation en faisant ressortir leurs avantages et leurs inconvénients.

MOTS CLÉS: Stratégie de plan d'échantillonnage; estimateurs de plan; estimateurs de modèle.

1. INTRODUCTION

Pendant des décennies, les fichiers administratifs et les recensements étaient les principales sources de données destinées à l'élaboration des politiques et à la planification pour les grandes et les petites régions. Ils constituent encore aujourd'hui la source de données la plus précieuse pour les petites régions dans la plupart des pays. Dans les années quarante et cinquante toutefois, comme les enquêtes par sondage prenaient de l'importance, les estimations d'enquête sont venues compléter les sources habituelles car elles constituent des données plus à jour et d'un bon rapport coût-efficacité pour un grand nombre de domaines spécialisés. Bien qu'elles soient conçues tout d'abord pour produire des estimations fiables aux niveaux d'agrégation supérieurs, comme les niveaux national et provincial, les enquêtes par sondage visent de plus en plus à répondre à la demande croissante d'estimations plus actuelles pour des domaines de types et de tailles variés. Aucun problème technique ne se pose tant que les domaines sont suffisamment grands (ex.: groupes d'âge-sexe, grandes villes et régions infra-provinciales) pour produire des estimations raisonnablement fiables. Cependant, si on a besoin de données pour de petits domaines, particulièrement des domaines dont les éléments sont répartis dans plusieurs strates de plan, des problèmes d'estimation particuliers se posent et plusieurs méthodes, proposées récemment, existent pour les résoudre.

Cet article a principalement pour but de souligner la nécessité d'envisager le problème des données régionales dans sa globalité. On devrait prendre conscience de la question des petites régions dès le début de la conception des plans de sondage pour les grandes enquêtes. Les plans d'échantillonnage devraient être conçus de manière que l'on puisse obtenir des données régionales fiables à l'aide d'estimateurs de plan ou de modèle. La solution voulant que les organismes statistiques règlent cette difficulté grandissante à l'étape de l'estimation ne devrait être envisagée qu'en dernier recours.

Dans la section 2, nous examinons les besoins en information et les trois principales sources de données socio-économiques au Canada, c'est-à-dire le recensement, les fichiers administratifs et les enquêtes. Dans la section suivante, nous traitons quelques questions techniques concernant les trois sources de données et nous mettons en lumière le problème des mesures de qualité et de leur interprétation. Ensuite, nous faisons ressortir la nécessité d'élaborer une approche globale qui recouvre les trois grandes étapes de la conception d'une enquête: planification, élaboration du plan de sondage et estimation. Nous examinons particulièrement deux aspects du plan de sondage, à savoir la formation de grappes dans un plan à plusieurs degrés et la répartition de l'échantillon. Dans la section 5, nous présentons quelques caractéristiques de plan de sondage qui ont été introduites récemment dans l'Enquête sur la population active du Canada la plus grande enquête mensuelle auprès des ménages de Statistique Canada à l'occasion de son remaniement afin qu'elle produise des données régionales de meilleure qualité. Enfin, la section 6 a pour but d'examiner les nombreuses méthodes d'estimation qui s'appliquent aux petites régions. En même temps, nous proposons de nouveaux estimateurs et nous faisons des commentaires sur les avantages et les faiblesses de divers estimateurs pour domaine. Nous conseillons la prudence dans l'utilisation d'estimateurs de modèle.

2. BESOINS EN INFORMATION ET SOURCES DE DONNÉES

En tant qu'organisme statistique national, Statistique Canada joue un rôle intégral dans la marche de la société canadienne. Tandis que le caractère confidentiel des données fournies par les répondants est préservé, l'information produite par l'organisme décrit les conditions économiques et sociales du pays et de sa population. Les divers programmes de l'organisme (statistique économique, statistique démographique, statistique sociale et statistique

[1] M.P. Singh, J. Gambino et H.J. Mantel, Statistique Canada, 16ième étage, Immeuble R.H. Coats, Parc Tunney, Ottawa (Ontario), Canada K1A 0T6.

des établissements) produisent des données fiables de tous genres aux niveaux national, provincial et infra-provincial à l'intention des administrations publiques fédérale et provinciales, des institutions privées, des universités et des médias. Comme la planification, l'administration et la surveillance des programmes sociaux et financiers tendent de plus en plus à se faire au niveau régional, on cherche à obtenir un plus grand nombre de données et des données de meilleure qualité à ce niveau. Nous examinons brièvement ci-dessous trois grandes sources de données socio-économiques et démographiques en mettant l'accent sur les statistiques régionales.

Recensement de la population: Le recensement quinquennal de la population produit des données-repères et constitue, à tous les cinq ans, la source d'information la plus précieuse ayant trait aux petites régions et à divers caractères, domaines ou groupes cibles tels que les minorités ethniques, les personnes en état d'incapacité, les jeunes et les autochtones.

Fichiers administratifs: Les fichiers administratifs sont une source de données statistiques de plus en plus importante. Les organismes statistiques puisent largement dans ces fichiers en démographie pour produire des estimations régionales (Schmidt 1952, Verma et Basavarajappa 1987). Pour certains domaines, comme la statistique de l'état civil, les fichiers administratifs sont la seule source d'information qui permet de produire des statistiques à des niveaux d'agrégation variés. Pour d'autres domaines, l'usage des fichiers administratifs dépend de leur capacité d'offrir des données à jour et des données de qualité comparativement aux recensements ou aux enquêtes. Les fichiers administratifs ne servent pas uniquement à la production de totalisations directes; ils sont utilisés dans un certain nombre de programmes comme source d'information supplémentaire dans le but d'améliorer la qualité d'estimations d'enquête. Ils servent aussi à la construction de bases de sondage pour des enquêtes, par exemple le Registre des entreprises et le Registre des adresses de logements résidentiels à Statistique Canada.

Comme le recensement de la population, les fichiers administratifs sont caractérisés par des niveaux d'agrégation géographique très détaillés, ce qui en fait une source d'information précieuse pour l'établissement de statistiques régionales. Leur accessibilité est maintenant plus grande et grâce aux progrès technologiques récents, ils deviennent une source de données ayant un meilleur rapport coût-efficacité. Sur le plan du contenu toutefois, les fichiers administratifs ne sont pas aussi complets et les concepts sont définis pour des programmes plutôt qu'à des fins statistiques. Brackstone (1987a, 1987b) donne les détails d'un programme de Statistique Canada qui a pour objet l'élaboration et l'intégration d'un système de fichiers administratifs destiné à produire des données statistiques. Le recueil rédigé par Coombs et Singh (1987) décrit l'expérience d'autres pays en ce qui concerne l'utilisation de fichiers administratifs.

Programme des enquêtes-ménages: Les enquêtes-ménages sont depuis longtemps une source majeure de données économiques et sociales à Statistique Canada. Les enquêtes de ce programme peuvent être classées en trois catégories: i) l'Enquête sur la population active, ii) les enquêtes spéciales et les enquêtes supplémentaires et iii) les enquêtes longitudinales et cycliques. Nous décrivons brièvement ci-dessous ces enquêtes en indiquant leur rapport avec la question des données régionales.

Ayant été créée comme une enquête trimestrielle en 1945, l'Enquête sur la population active du Canada (EPA) est devenue une enquête mensuelle en 1952. L'information recueillie au moyen de cette enquête s'est diversifiée considérablement au fil des années et aujourd'hui, elle donne un portrait complet et détaillé du marché du travail au Canada. Outre les estimations nationales et provinciales, l'EPA produit régulièrement des estimations pour les régions infra-provinciales. Il existe aussi une forte demande pour des estimations d'indicateurs courants du marché du travail relatives à de petites régions comme les circonscriptions électorales fédérales, les divisions de recensement et les territoires des centres d'emploi du Canada. Ces estimations servent aux administrations fédérale et provinciales pour contrôler les programmes et répartir les ressources, monétaires ou autres, entre les divers champs de compétence politique et administrative.

Pour des considérations de coût, l'appareil de l'EPA est souvent utilisé pour effectuer des enquêtes ponctuelles et des enquêtes périodiques à l'échelle nationale et provinciale; ce peut être sous la forme d'enquêtes supplémentaires ou d'enquêtes spéciales. Dans le cas des enquêtes supplémentaires, on demande aux participants à l'EPA de répondre à des questions additionnelles tandis que dans le cas des enquêtes spéciales, on demande à un autre échantillon de ménages, tiré de la base de l'EPA, de répondre aux questions. En règle générale, les enquêtes spéciales et les enquêtes supplémentaires sont parrainées par d'autres ministères et réalisées selon le principe de recouvrement des coûts. Dans le cas de ces enquêtes, la demande de statistiques régionales varie beaucoup selon l'enquête et, en règle générale, elle semble moins pressante que celle provenant de l'EPA.

Statistique Canada effectue annuellement une Enquête sociale générale (ESG) pour répondre, modestement, aux besoins grandissants en données sur des sujets d'actualité sociale. Le programme de l'ESG (Norris et Paton 1991) comporte cinq cycles d'enquête chacun ayant une thématique principale qui reviennent tous les cinq ans. À cause de la taille limitée de l'échantillon (10,000 ménages à l'échelle nationale), l'ESG met l'accent sur les estimations nationales et les données analytiques.

Les enquêtes longitudinales (ou enquêtes par panel) sont nouvelles dans le contexte canadien. Statistique Canada a mis sur pied deux enquêtes longitudinales qui viendront enrichir largement le programme des enquêtes-ménages; ce sont l'Enquête sur la dynamique du travail et du revenu et l'Enquête nationale sur la santé de la population. Il s'agit de deux grandes enquêtes par panel dont on s'attend déjà qu'elles produisent des données au niveau infra-provincial et régional.

3. QUESTIONS RELATIVES À L'ESTIMATION POUR DOMAINE

La fourniture de données régionales soulève de nombreuses questions d'ordre stratégique ou technique. L'importance de ces questions peut varier d'un organisme à l'autre et d'une application à l'autre au sein du même organisme, selon la qualité des données et les politiques de diffusion. Ces questions intéressent les estimations nationales et provinciales, mais elles supposent une plus grande importance pour les données régionales. Selon les termes mêmes de Brackstone (1987a), "(...) en ce qui concerne l'évaluation de données régionales, il convient de noter que les utilisateurs peuvent observer plus facilement une erreur dans des estimations régionales que dans des agrégats nationaux (...) il se trouvera des détracteurs prompts à souligner les lacunes des données régionales (...) il est vrai qu'en ce qui concerne les petites régions, pour lesquelles l'estimation est plus difficile, les valeurs estimées font l'objet d'un examen plus intense." Deux ouvrages en particulier Platek et coll. (1987) et Platek et Singh (1986) présentent plusieurs études de recherche et d'élaboration sur l'estimation pour petites régions. Pour avoir un aperçu des méthodes d'estimation pour petites régions utilisées actuellement dans les programmes statistiques fédéraux aux É.-U., prière de se référer au document du U.S. Statistical Policy Office (1993).

Utilisation de fichiers administratifs: Dans la plupart des cas, l'offre et la demande de données régionales sont influencées en tout premier lieu par les politiques des gouvernements fédéral et provinciaux. Du côté de l'offre, les fichiers administratifs basés sur des programmes gouvernementaux renferment une masse de renseignements qui peuvent servir à produire des données régionales. Des exemples de ces fichiers au Canada: allocations familiales, assurance-chômage, impôt sur le revenu, santé, éducation, sécurité de la vieillesse. On produit déjà de façon régulière des données régionales sur le revenu. Tout **changement** dans la politique du gouvernement et les programmes qui s'y rattachent peut avoir un effet immédiat, positif ou négatif, sur le champ d'application, la disponibilité, l'actualité ou la qualité des données tirées des fichiers administratifs correspondants. En ce qui concerne la demande, les administrations publiques ont besoin de données régionales pour la planification, l'administration et la surveillance de leurs programmes.

Problèmes conceptuels: Dans des séries de données, on confond très souvent les problèmes conceptuels ou définitionnels et les problèmes d'échantillonnage et d'estimation. Prenons par exemple le Régime d'assurance-chômage (AC) du Canada. Le Règlement sur l'assurance-chômage prévoit différentes périodes d'admissibilité selon le taux de chômage qui sévit dans une région donnée, de telle sorte que les périodes ouvrant droit aux prestations sont plus courtes dans les régions qui affichent des taux de chômage relativement élevés. Les estimations du taux de chômage régional établies à l'aide des données de l'EPA servent à déterminer l'admissibilité d'une personne aux prestations. C'est pourquoi les médias et la population en général prêtent une attention soutenue à ces estimations régionales. Mais cet intérêt a trait plus souvent à des questions **conceptuelles** qu'à des questions d'estimation; on s'interroge, par exemple, sur la place des travailleurs découragés, des mises à pied ou des licenciements et des techniques de recherche d'emploi dans le questionnaire d'enquête.

Utilisation de modèles et des mesures de qualité correspondantes: Presque toutes les grandes enquêtes ont leurs estimations pour domaines et aucun problème ne se pose tant que les estimateurs de plan, c.-à-d. des estimateurs approximativement non biaisés selon le plan, sont de qualité acceptable. On considère deux catégories d'estimateurs de plan. D'après Schaible (1992), les estimateurs **directs** sont des estimateurs qui utilisent des valeurs de la variable étudiée qui se rapportent uniquement à des unités du domaine pour la période visée (ex.: l'estimateur par régression dont la pente est estimée au moyen de l'échantillon dans le domaine seulement). Les estimateurs de ce genre peuvent aussi utiliser - et cela est fréquent - de l'information relative à une ou à plusieurs variables auxiliaires pour d'autres domaines ou d'autres périodes, et ces estimateurs sont (approximativement) non biaisés selon le plan. Il existe une autre catégorie d'estimateurs, que nous appelons estimateurs **directs modifiés** et qui peuvent utiliser de l'information relative à la variable auxiliaire et à la variable étudiée pour d'autres domaines mais qui conservent leur propriété d'être (approximativement) sans biais selon le plan (ex.: l'estimateur par régression dont la pente est estimée au moyen de tout l'échantillon). Il existe de plus en plus d'ouvrages sur les **estimateurs indirects (ou estimateurs de modèle)**, c'est-à-dire des estimateurs qui utilisent de l'information relative à la variable étudiée et aux variables auxiliaires pour d'autres domaines ou d'autres périodes mais qui n'ont pas nécessairement la propriété d'être non biaisés selon le plan.

La plupart des producteurs et des utilisateurs de données d'enquête sont familiarisés avec les estimateurs de plan et les inférences correspondantes. Ils interprètent les données sur la base d'échantillons répétés formés à l'aide d'un plan d'échantillonnage probabiliste donné et se servent des **c.v.** de plan estimés (c.v. = coefficients de variation: racine carrée de la variance de plan divisée par la valeur estimée selon le plan) comme mesure de la qualité des données. Dans les cas où les domaines sont trop petits ou bien le plan de sondage ne prévoit pas la production d'estimations régionales, les estimations de plan peuvent donner des **c.v.** élevés et les estimations de modèle peuvent alors s'imposer comme la seule solution s'il faut produire des estimations d'enquête pour domaines. La tâche qui attend les statisticiens n'est pas simple: comment estimer et comparer la précision relative des estimations d'une enquête qui produit un grand nombre d'estimations nationales et infra-nationales et de nombreuses estimations pour de grands et de petits domaines en utilisant surtout des estimateurs de plan mais aussi, dans une moindre mesure, des estimateurs de modèle, et comment présenter cela aux utilisateurs. Les **c.v.** de modèle (racine carrée de la variance théorique de la valeur estimée selon le modèle, divisée par la valeur estimée selon le modèle) peuvent traduire une situation totalement

différente et être beaucoup moins élevés que les **c.v.** de plan correspondants pour la même petite région et, dans beaucoup de cas, moins élevés que les **c.v.** de plan pour des régions beaucoup plus grandes.

Il est généralement simple d'établir l'expression de l'erreur quadratique moyenne (c.-à-d., variance théorique + carré du biais théorique) pour les estimateurs de modèle. Il n'en va pas de même lorsqu'il s'agit d'estimer cette expression avec un degré de précision raisonnable. S'il est vrai que les données (ex.: taille d'échantillon) pour de petits domaines ne se prêtent pas à la production d'estimations de plan, il est peu probable qu'elles puissent servir à produire une estimation de la variance et du biais correspondants. Comme il est relativement plus difficile d'estimer le biais, certains auteurs recherchent chez les estimateurs de modèle la propriété de convergence selon le plan, ce qui laisse supposer peut-être qu'on peut faire abstraction du biais. Or, si la taille de l'échantillon à l'intérieur d'un domaine est suffisamment grande pour qu'un estimateur de modèle soit convergent, un estimateur de plan devrait produire des estimations fiables pour ce domaine. En ce qui concerne les estimateurs de modèle, on a proposé d'utiliser l'estimation de l'erreur quadratique moyenne calculé pour l'ensemble des domaines. Comme, en règle générale, on cherche à obtenir des estimations pour chacun des domaines pris individuellement parce qu'on croit que ces domaines sont différents les uns des autres, on doit s'efforcer d'expliquer pourquoi les estimations de ces divers domaines ont toutes le même degré de fiabilité. Une autre possibilité serait de calculer des estimations indirectes (basées sur un modèle) de la variance et du biais des estimateurs de modèle pour des domaines **pris individuellement**. De fait, l'élaboration de méthodes qui permettraient d'estimer l'erreur quadratique moyenne pour des domaines pris individuellement devrait figurer parmi les priorités de recherche. Une autre préoccupation majeure des expérimentateurs est de savoir comment parer aux défaillances de modèle. Peut-être faudrait-il effectuer des recherches sur la validation de modèle dans des enquêtes complexes. En outre, dans le cas des estimateurs de modèle qui utilisent des données relatives à la variable étudiée pour d'autres périodes que la période pertinente, les estimations de la **variation** d'une période à l'autre seraient de qualité douteuse; voir Schaible (1992). Par surcroît, les estimateurs de modèle qui empruntent de l'information à d'autres domaines de la région présenteraient la même lacune si on comparait des différences entre deux domaines de la région.

Protection de la vie privée: Afin de créer des bases de données complètes desquelles on peut tirer des statistiques régionales, on doit parfois combiner des données de recensement et d'enquête et des données de fichiers administratifs. Cette opération nécessite le couplage d'enregistrements de diverses sources. Or, compte tenu de l'intérêt qu'attache la population à la protection de la vie privée, on ne devrait effectuer des couplages d'enregistrements qu'après en avoir analysé minutieusement toutes les conséquences. La Loi sur la statistique permet à Statistique Canada d'avoir accès aux fichiers administratifs d'autres ministères ou organismes pour des raisons statistiques. Même dans ce cas, comme le souligne Fellegi (1987), "(. . .) nous devrions disposer de procédures de révision rigoureuses et vérifiables afin de nous assurer que nous ne combinons ces données que lorsque le bien public qui résulte de ces nouvelles informations statistiques l'emporte manifestement sur l'intrusion dans la vie privée que leur création entraîne."

4. NÉCESSITÉ D'UNE APPROCHE GLOBALE

Bien que les grandes enquêtes visent surtout à produire des estimations nationales et provinciales, il est rare que les estimations issues de ces enquêtes concernent uniquement les populations nationale ou provinciales dans leur entier. Autrement dit, les grandes enquêtes servent immanquablement à produire des estimations pour divers domaines **recoupés** et, parfois, pour des domaines **spatiaux** (ex.: régions infra-provinciales). Dans beaucoup de cas, on ne cherche pas particulièrement à obtenir un niveau de précision voulu pour le domaine, que ce soit à l'étape de la conception du plan de sondage ou à celle de l'estimation, pourvu que le degré de fiabilité soit (jugé) raisonnable. Les problèmes surgissent lorsqu'un domaine recoupé correspond à une sous-population rare ou qu'un domaine spatial correspond à une petite région, auxquels cas il n'existe aucune estimation ou, si elles existent, les estimations sont de qualité douteuse. Dans un certain nombre de cas, ces problèmes peuvent survenir pour la simple raison qu'on n'a pas été assez attentif à ces besoins au début du processus de planification. Si on doit répondre aux besoins en données régionales à l'aide de données d'enquête, il est nécessaire d'élaborer une approche globale qui visera particulièrement à répondre à ces besoins à chacune des étapes d'une enquête: planification, sondage et estimation. En ce qui regarde le plan de sondage et l'estimation, nous classons les domaines en deux catégories:

Domaines planifiés: Du point de vue de l'échantillonnage, les domaines planifiés sont des strates ou des groupes de strates qui ont été formés en vue de la création d'échantillons. Des exemples de ces domaines au Canada sont les régions infra-provinciales, par exemple les régions économiques, les régions de l'Assurance-chômage et les régions créées aux fins de la planification des services de santé. D'autres exemples de ces domaines seraient les grands comtés, les districts ou les régions infra-provinciales de ce type.

Domaines non planifiés: Ce sont des domaines qui n'ont pas été définis à l'étape de la conception du plan de sondage et qui, par conséquent, peuvent recouper des strates. Ils peuvent avoir n'importe quelle taille et créer des problèmes d'estimation particuliers.

Planification: Comme nous l'avons mentionné plus tôt, les données d'enquêtes périodiques comme l'EPA sont beaucoup plus en demande que les données d'enquêtes ponctuelles. En ce qui concerne les enquêtes périodiques qui sont remaniées à tous les cinq ou dix ans, on peut élaborer une stratégie adéquate au moment du remaniement puisque dans ces cas précis, les organismes statistiques sont généralement beaucoup plus en mesure de prévoir les

besoins en données régionales en se fondant sur la demande antérieure. Quant aux enquêtes ponctuelles, les concepteurs devraient inclure la définition des besoins dans le processus d'établissement des objectifs de l'enquête. Par conséquent, dans l'un et l'autre cas, les concepteurs devraient fixer un degré de précision non seulement pour les estimations nationales ou provinciales, mais aussi pour les domaines étudiés.

La première étape dans la fourniture de données régionales dépendra dans quelle mesure les domaines auront été définis à l'avance, de sorte qu'ils puissent être considérés comme des domaines planifiés au moment de la conception (ou du remaniement) de l'enquête. Si, pour des considérations budgétaires, on ne peut obtenir d'estimations fiables pour certains domaines particulièrement petits, les concepteurs d'une enquête devraient envisager sérieusement, de concert avec les promoteurs, la possibilité de regrouper des domaines, de réunir des estimations d'enquêtes différentes ou encore de ne pas produire d'estimations. Certains domaines ne peuvent être définis à l'avance. Ces domaines non planifiés appellent des méthodes d'estimation spéciales.

Plan de sondage: Il est rare dans la pratique de trouver un plan "optimal", que ce soit au niveau national ou provincial ou pour un domaine en particulier. Le plus souvent, on introduira des compromis à différentes étapes de l'échantillonnage et de la collecte des données pour répondre aux contraintes théoriques et opérationnelles. Selon les besoins en données, les estimations pour domaines devraient aussi faire partie intégrante de ces compromis. Nous allons étudier deux façons de tenir compte des besoins en données régionales à l'étape de l'élaboration du plan de sondage, notamment la répartition de l'échantillon et le degré de mise en grappes de l'échantillon.

Méthode de répartition: De façon générale, une méthode de répartition optimale pour des estimations nationales permet de répartir les échantillons dans les provinces au prorata de leur population. La fiabilité des estimations pour les petites provinces en souffre. Il est donc préférable d'effectuer une répartition intermédiaire. Il y a diverses façons d'effectuer cette répartition, selon l'importance que l'on accorde aux estimations provinciales ou aux estimations des États (dans le cas de É.-U.). Dans le cas des grandes provinces, une faible réduction de la taille de l'échantillon aura généralement peu d'effet sur la fiabilité des données pour ces provinces (ou des données nationales), mais une augmentation équivalente de la taille d'échantillon dans les petites provinces aura un effet positif appréciable sur la fiabilité des données relatives à ces provinces.

Le même raisonnement s'applique aux domaines planifiés à l'intérieur des provinces car dans la plupart des cas, les répartitions optimales autorisent un certain "jeu" et les concepteurs peuvent exploiter cette caractéristique en redistribuant l'échantillon par un transfert d'unités des grandes régions vers des domaines planifiés de taille moindre.

Mise en grappes: Les enquêtes-ménages de grande envergure comportent habituellement des plans à plusieurs degrés stratifiés avec des unités primaires d'échantillonnage relativement grandes afin d'assurer un bon rapport coût-efficacité pour les estimations nationales et provinciales. Ce genre de plans impliquent donc de nombreuses grappes et nuisent par conséquent à la production de statistiques pour des domaines spatiaux non planifiés en ce sens que, sous l'effet du hasard, certains domaines peuvent contenir de nombreuses unités d'échantillon alors que d'autres peuvent n'en contenir aucune. Étant donné l'importance des estimations pour domaines, on devrait s'efforcer de limiter le plus possible la formation de grappes dans l'échantillon. À cet égard, les facteurs suivants jouent un rôle important: choix de la base de sondage, choix des unités d'échantillonnage et de la taille de ces unités, nombre de strates et leur taille et degrés d'échantillonnage. L'objectif est de réduire au maximum les effets de plan étant donnée les contraintes opérationnelles.

Estimation: Quel que soit le degré d'attention que l'on prête aux estimations pour domaine dans les premières étapes de la planification et de la conception d'une enquête, il existera toujours de petits domaines pour lesquels il faudra appliquer des méthodes d'estimation spéciales afin d'obtenir des valeurs suffisamment justes. Depuis quelque temps, on s'intéresse d'une manière particulière aux estimateurs synthétiques; ceux-ci empruntent de l'information à des domaines qui ressemblent au domaine étudié. Or, comme les estimateurs synthétiques sont très sensibles à l'hypothèse de la similarité des domaines, il en faudrait peu pour que le biais théorique soit élevé et que l'utilisation de ces estimateurs soit remise en question. Préoccupés du biais théorique, les concepteurs de plan de sondage ont proposé de combiner des estimateurs directs et des estimateurs synthétiques dans le but de résoudre le problème du biais théorique tout en essayant de conserver les qualités de l'estimateur synthétique. On s'est servi de méthodes empiriques de Bayes et de méthodes similaires pour pondérer chaque composante des estimateurs combinés. Nous traiterons brièvement ce sujet dans la section 6, qui porte sur l'estimation.

5. CONSIDÉRATIONS RELATIVES AU PLAN D'ÉCHANTILLONNAGE

5.1 Introduction

En règle générale, on pense résoudre le problème des petits domaines (ou des petites régions) par l'estimation. Or, comme nous l'avons mentionné dans la section précédente, il est possible d'envisager la question dès l'étape de la conception du plan de sondage. Pour illustrer nos propos, nous allons nous servir de l'Enquête sur la population active (EPA) du Canada.

Voici les caractéristiques du plan d'échantillonnage actuel de l'EPA: l'Enquête sur la population active du Canada est une enquête mensuelle menée auprès de 59,000 ménages qui sont échantillonnés en plusieurs étapes par diverses méthodes. L'unité d'échantillonnage ultime, c.-à-d. le ménage, demeure dans l'échantillon pendant six mois, puis est remplacée. Les unités des degrés supérieurs (unités primaires d'échantillonnage (UPÉ), grappes) sont

aussi renouvelées périodiquement. Chacune des dix provinces du Canada est divisée en régions économiques (RÉ), lesquelles, en vertu de l'EPA, sont subdivisées en régions autoreprésentatives (grandes villes et villes de taille moyenne) et en régions non autoreprésentatives (le reste de la RÉ). La stratification et l'échantillonnage s'effectuent à l'intérieur de ces régions, et le nombre de degrés d'échantillonnage de même que les unités d'échantillonnage varient selon le type de région. Par exemple, le nombre de degrés d'échantillonnage est de trois dans les régions non autoreprésentatives mais seulement deux dans les régions autoreprésentatives. Pour une description détaillée du plan d'échantillonnage de l'EPA, prière de se référer à Singh et coll. (1990).

5.2 Degrés d'échantillonnage et unités d'échantillonnage

Les bases aréolaires ont rapport généralement à l'échantillonnage en grappes, c'est-à-dire que les unités d'échantillonnage du premier degré sont habituellement des superficies qui contiennent un certain nombre d'unités du deuxième degré. Si on dispose d'une liste des unités du deuxième degré, on peut alors effectuer un échantillonnage à même la liste, ce qui donne un échantillon comportant un moins grand nombre de grappes. Nous aurons ainsi non seulement des estimations de meilleure qualité (grâce à des effets de plan moindres), mais aussi de meilleures estimations régionales pour les domaines non planifiés. La seconde conséquence est vraie car une répartition plus uniforme de l'échantillon a de fortes chances de "placer" des unités échantillonnées dans un domaine spatial non planifié. Inversement, dans le cas d'un plan avec formation de grappes, il arrive souvent qu'un domaine renferme un nombre suffisant d'unités de l'échantillon du fait qu'il contient des grappes échantillonnées, tandis qu'un autre domaine semblable ne contiendra pas suffisamment de grappes pour produire de bonnes estimations.

Nous avons examiné deux solutions pour réduire la formation de grappes dans l'EPA: i) remplacer la base aréolaire (avec plan à deux degrés) pour les grandes villes par une nomenclature fondée sur le Registre des adresses ou ii) réduire le nombre de degrés d'échantillonnage pour les régions rurales et les petits centres urbains. Le Registre des adresses, qui a été créé dans le but d'améliorer le taux de couverture du recensement du Canada de 1991 (Swain, Drew, Lafrance et Lance 1992), contient des adresses et des numéros de téléphone ainsi que de l'information géographique se rapportant à des logements pour chaque secteur de dénombrement (SD). La première solution examinée consistait à tirer un échantillon aléatoire simple stratifié de logements dans le Registre des adresses. À cet échantillon pouvait s'ajouter un second tiré d'une base de croissance qui contient des logements qui ne figurent pas dans le registre des adresses post-censitaire. Seulement, la manière de traiter la croissance est devenue le principal obstacle dans l'application de la première solution car il n'a pas été possible d'élaborer et de tester une méthode d'un bon rapport coût-efficacité dans les délais requis par le projet de remaniement de l'EPA. Néanmoins, on travaille actuellement à l'élaboration d'une méthode de mise à jour du registre des adresses post-censitaire en vue des enquêtes et des recensements futurs.

En ce qui concerne la deuxième solution, on a envisagé la modification d'unités et la réduction du nombre de degrés d'échantillonnage pour les régions non urbaines en gardant en mémoire l'idée qu'une réduction de la formation en grappes se traduit par de meilleures estimations régionales. Étant donné l'évolution qu'ont connue les méthodes de collecte de données (de l'interview directe à l'interview téléphonique à l'interview assistée par ordinateur), les analyses qui visaient auparavant à trouver un équilibre entre le coût d'une enquête et la variance désirée sont révolues. Plus de 80 pour cent des interviews de l'EPA sont maintenant effectuées par téléphone. Avec l'accroissement du nombre des interviews téléphoniques et la diminution connexe du nombre des déplacements, il est presque toujours possible maintenant de supprimer le premier degré d'échantillonnage et de prélever directement des SD.

5.3 Stratification

Une méthode de stratification conforme à l'esprit de l'analyse qui s'est faite précédemment sur la taille de l'UPÉ consiste à remplacer les grandes strates par de nombreuses petites strates dans l'espoir qu'un domaine redéfini ou qu'un domaine non planifié renferment des strates presque complètes. La taille des domaines sera ainsi plus stable.

Il peut y avoir plusieurs régions en chevauchement pour lesquelles on a besoin d'estimations. Par exemple, chaque province canadienne est divisée en régions économiques (RÉ) et en régions de l'Assurance-chômage (RAC). Une solution consiste à considérer tous les secteurs créés par l'intersection des régions comme des strates. Ainsi, au Canada, le recoupement des 71 RÉ et des 61 RAC produit en tout 133 régions en chevauchement, ce qui est raisonnable. Dans certains cas toutefois, le nombre de régions chevauchantes pourrait être trop élevé pour que les opérations soient efficaces. En outre, certains des secteurs pourraient avoir un effectif minime, ce qui en ferait des strates inutilisables.

En combinant la réduction de la formation en grappes et la diminution de la taille des strates, nous espérons avoir un plan de sondage qui pourra mieux répondre aux exigences des petites régions. Par exemple, les plans de ce genre devraient offrir toute la souplesse voulue pour répondre efficacement aux exigences des RÉ et des RAC et traiter les changements de définition des régions.

5.4 Répartition

Si nous connaissons d'avance les définitions des petites régions, nous pourrons peut-être considérer ces régions comme des domaines planifiés et en tenir compte au moment de la conception de l'enquête. Le concepteur peut s'appliquer à répartir l'échantillon dans les petites régions de manière à rendre possible la production d'estimations fiables. Pour de grandes enquêtes comme l'Enquête sur la population active du Canada, cette approche peut, à

tout le moins en théorie, favoriser la production d'un grand nombre d'estimations régionales. Si nous avons un échantillon mensuel de 59,000 ménages et si nous supposons qu'il faut, par exemple, 100 ménages par mois pour produire des estimations trimestrielles fiables, on peut diviser le pays en quelque 600 régions non chevauchantes, chacune contenant avec certitude un nombre suffisant d'unités de l'échantillon. De la même manière, en groupant des régions, on obtiendra des territoires qui contiendront un nombre suffisant d'unités pour produire des estimations mensuelles fiables.

Il existe diverses méthodes de répartition de l'échantillon. Dans une perspective du haut vers le bas, on détermine tout d'abord la taille de l'échantillon provincial; on répartit ensuite l'échantillon entre les régions infra-provinciales. Il se peut toutefois que la taille de l'échantillon provincial empêche que le degré de fiabilité requis pour les estimations infra-provinciales soit respecté. Dans une perspective du bas vers le haut, on répartit tout d'abord l'échantillon entre les régions infra-provinciales de manière à répondre aux objectifs de fiabilité pour chaque région. En conséquence, ces régions devraient normalement avoir des tailles d'échantillon comparables. Avec cette méthode, il se peut que la taille de l'échantillon provincial soit plus grande que celle déterminée par la méthode précédente. Quelle que soit la méthode utilisée, il faudra normalement apporter des corrections aux chiffres initiaux. La répartition finale sera une sorte de compromis entre la répartition proportionnelle et la répartition uniforme. En pratique, le concepteur de plan de sondage doit se livrer à un exercice complexe où il doit tenir compte à la fois des conditions de fiabilité au niveau provincial, des exigences au niveau infra-provincial pour un ou plusieurs groupes de régions, du coût total de l'enquête et des conditions sur le terrain.

La méthode de répartition qui a été choisie pour la version remaniée de l'EPA peut être utile pour d'autres enquêtes. La répartition de l'échantillon s'est faite en deux étapes: on a tout d'abord réparti un échantillon de base de 42,000 ménages de manière à obtenir de bonnes estimations aux niveaux national et provincial; on a ensuite réparti le reste de l'échantillon de manière à obtenir des estimations infra-provinciales le plus justes possible. Cette formule de répartition permettra d'obtenir des estimations fiables pour presque tous les domaines planifiés. Cette solution de compromis s'est traduite par des pertes mineures au niveau provincial mais par des gains appréciables au niveau infra-provincial. Par exemple, le c.v. théorique pour les "personnes en chômage" en Ontario et au Québec est 3.2 et 3.0% respectivement, au lieu de 2.8 et 2.6%. Les pourcentages correspondants pour le Canada sont 1.51 et 1.36. Si on vise l'optimisation au niveau provincial, on obtient des c.v. pouvant atteindre 17.7% pour les régions de l'A.-C. Grâce à la formule de répartition décrite ci-dessus, ces mêmes c.v. ne dépassent pas 9.4%.

Redistribution de l'échantillon: Le concepteur jouit normalement d'une certaine marge de manoeuvre pour déplacer des unités de l'échantillon d'une région à une autre. Par exemple, si on réduit de 1,000 ménages la taille de l'échantillon d'une grande province et qu'on augmente d'autant la taille de l'échantillon d'une petite province, la qualité des estimations pour la grande province s'en trouvera peu affectée alors que celle des estimations pour la petite province en sera grandement améliorée. On peut effectuer des transferts semblables à l'intérieur d'une province.

5.5 Autres considérations

Modification de la définition des petites régions: Les concepteurs de plan de sondage doivent se faire à l'idée que les définitions de domaines planifiés peuvent changer au cours de l'existence d'une enquête et qu'ils peuvent devoir considérer de nouveaux domaines comme des domaines non planifiés. Par exemple, il se peut très bien que la définition des régions de l'Assurance-chômage soit modifiée dans les deux ou trois années qui vont suivre la mise en application du nouveau plan de sondage de l'EPA en 1995. La meilleure chose que le concepteur pourra faire pour parer à cette difficulté à l'étape de l'élaboration sera de choisir comme unités de base des régions dites normalisées (par ex., des régions propres au recensement dont la définition varie peu) et d'espérer que les régions redéfinies correspondent à un groupement de régions normalisées. C'est la solution qui est préconisée dans la version remaniée de l'EPA.

Une autre solution consisterait à adopter une méthode de mise à jour. Cela nécessiterait le tirage de nouvelles unités; ce tirage serait effectué de manière à maximiser le chevauchement des unités tirées initialement et des nouvelles unités. De cette manière, on réduirait au maximum le nombre de nouvelles unités à inscrire sur la liste. On réduirait aussi au maximum les autres contraintes opérationnelles, comme la nécessité d'embaucher d'autres intervieweurs.

6. ESTIMATION

Cette section a pour but d'exposer quelques-unes des méthodes d'estimation de totaux pour petites régions. Nous ne cherchons pas ici à faire une analyse exhaustive de ces méthodes; nous décrivons plutôt l'évolution des recherches sur l'estimation pour petites régions. Pour une analyse détaillée de la question, le lecteur est prié de consulter le tout récent article de Ghosh et Rao (1993). Pour faciliter notre exposé, nous allons classer les méthodes d'estimation pour petites régions en deux groupes. (Il existe beaucoup d'autres possibilités de classement). Le premier groupe comprend les estimateurs de plan, c.-à-d. des estimateurs (approximativement) non biaisés selon le plan; dans ce groupe on trouve les estimateurs directs et les estimateurs directs modifiés. Comme nous l'avons mentionné plus haut, les estimateurs de plan laissent souvent à désirer à cause de leur variance élevée, qui s'explique par le faible effectif des petites régions (ou même un effectif nul). Le second groupe est désigné comme le groupe des estimateurs indirects (ou estimateurs de modèle) et il comprend les estimateurs synthétiques et les estimateurs combinés. Certains de ces estimateurs font l'objet d'une comparaison empirique dans une version antérieure de cet article; voir Singh, Gambino et Mantel (1992).

6.1 Estimateurs de plan

Estimateurs directs: Les estimateurs directs pour petite région reposent sur des données d'enquête qui proviennent uniquement de la petite région; à ces données peuvent s'ajouter parfois des données auxiliaires tirées du recensement ou de fichiers administratifs. La forme la plus simple de l'estimateur direct d'un total est l'estimateur par facteur d'extension,

$$\hat{Y}_{e,a} = \sum_{i \in s_a} w_i y_i, \tag{6.1}$$

où s_a est la portion de l'échantillon contenue dans la petite région a et w_i est le poids de sondage de l'unité i. Cet estimateur est non biaisé; cependant, il peut être caractérisé par une forte variabilité du fait que la taille d'échantillon dans la région a est aléatoire.

Si la taille de la population, N_a, est connue, on peut se servir d'un estimateur de stratification a posteriori,

$$\hat{Y}_{pst,a} = N_a \sum_{i \in s_a} w_i y_i \Big/ \sum_{i \in s_a} w_i = N_a \hat{Y}_{e,a} / \hat{N}_{e,a} = N_a \bar{y}_{e,a}. \tag{6.2}$$

Cet estimateur est plus stable que l'estimateur par facteur d'extension; cependant, il peut être entaché d'un biais d'estimation par quotient dans les enquêtes complexes.

Si l'échantillonnage est stratifié et que $N_{h,a}$, l'effectif de la population contenu dans la strate h et la petite région a, est connu, on peut utiliser un autre estimateur de stratification a posteriori: $\hat{Y}_{st,pst,a} = \sum_h (N_{h,a} \sum_{i \in s_{h,a}} w_i y_i / \sum_{i \in s_{h,a}} w_i) = \sum_h N_{h,a} \hat{Y}_{h,e,a} / \hat{N}_{h,e,a} = \sum_h N_{h,a} \bar{y}_{h,a}$. Il peut aussi s'agir de strates formées a posteriori au lieu de strates initiales.

L'estimation par quotient ressemble à l'estimation de stratification a posteriori avec cette différence qu'on utilise une variable auxiliaire au lieu des effectifs N_a et $N_{h,a}$. Par exemple, si x est une covariable pour laquelle on connaît le total de petite région, X_a, ou le total de strate de petite région, $X_{h,a}$, nous pouvons définir les estimateurs par quotient

$$\hat{Y}_{r,a} = X_a \hat{R}_a \quad \text{et} \quad \hat{Y}_{st,r,a} = \sum_h X_{h,a} \hat{R}_{h,a}, \tag{6.3}$$

où $\hat{R}_a = \hat{Y}_{e,a} / \hat{X}_{e,a}$ est une estimation du rapport Y_a / X_a et $\hat{R}_{h,a} = \hat{Y}_{h,e,a} / \hat{X}_{h,e,a}$.

Par ailleurs, l'estimateur par régression tente d'expliquer la différence entre la valeur de covariables pour l'effectif d'une petite région et celle pour un sous-échantillon au moyen d'une équation de régression estimée qui met en relation la variable étudiée y et les covariables x. Un des avantages de l'estimation par régression est qu'elle s'étend facilement à des covariables vectorielles. L'estimateur est défini

$$\hat{Y}_{reg,a} = \hat{Y}_a + \hat{\beta}_a (X_a - \hat{X}_a), \tag{6.4}$$

où $\hat{Y}_a$ peut être un estimateur par facteur d'extension ou un estimateur de stratification a posteriori, $\hat{X}_a$ doit être calculé de la même manière que $\hat{Y}_a$, et $\hat{\beta}_a = \sum_{i \in s_a} v_i^{-1} w_i y_i x_i' \{ \sum_{i \in s_a} v_i^{-1} w_i x_i x_i' \}^{-1}$, où v_i désigne les poids donnés pour la régression. Notons que $\hat{\beta}_a = \hat{R}_a$ lorsque x est une grandeur scalaire et $v_i = x_i$. Lorsque $\hat{Y}_a$ et $\hat{X}_a$ sont des estimateurs par facteur d'extension, l'estimateur (6.4) est aussi appelé estimateur par régression généralisé. Il est approximativement non biaisé selon le plan à la condition qu'il en soit de même pour $\hat{Y}_a$ et $\hat{X}_a$.

Comme l'estimateur par quotient, l'estimateur par régression peut être appliqué à des strates initiales ou à des strates formées a posteriori.

Estimateurs directs modifiés: Les estimateurs directs modifiés peuvent utiliser des données d'enquête qui ne proviennent pas du domaine étudié; cependant, ils demeurent des estimateurs approximativement non biaisés selon le plan. Selon notre définition, l'estimateur direct modifié est un estimateur direct qui comporte un facteur de correction synthétique pour biais de modèle; comme l'espérance du facteur de correction serait approximativement nulle par rapport au plan, l'estimateur modifié est approximativement non biaisé selon le plan si l'estimateur direct est aussi approximativement non biaisé selon le plan. On a un exemple d'estimateur direct modifié si on remplace $\hat{\beta}_a$ dans (6.4) par un estimateur synthétique, $\hat{\beta} = \sum_{i \in s} v_i^{-1} w_i y_i x_i' \{ \sum_{i \in s} v_i^{-1} w_i x_i x_i' \}^{-1}$; nous désignerons cet estimateur par $\hat{Y}_{sreg,a}$. $\hat{\beta}$ sera généralement plus stable que $\hat{\beta}_a$; le choix entre l'un et l'autre dépendra du rapport entre la grandeur de la variance de $\hat{\beta}_a$ et la variation des valeurs $\beta_a s$ pour les régions a. Une solution intermédiaire serait d'opter pour une moyenne pondérée, $\lambda_a \hat{\beta}_a + (1 - \lambda_a)\hat{\beta}$, où la valeur de λ_a est choisie judicieusement; le choix d'une valeur de λ_a est examiné dans la partie de la section 6.2 qui porte sur les estimateurs combinés. On a un autre exemple d'estimateur direct modifié si on remplace $\hat{\beta}_a$ dans (6.4) par $\hat{R} = \hat{Y}_e / \hat{X}_e$; notons que $\hat{R}$ est un cas particulier de $\hat{\beta}$ lorsque x est une grandeur scalaire et $v_i = x_i$.

6.2 Estimateurs indirects

Estimateurs synthétiques: Les méthodes d'estimation synthétique reposent sur l'hypothèse selon laquelle la petite région ressemble d'une certaine manière à une autre région, souvent plus grande, dans laquelle elle est contenue. Les estimations relatives à cette autre région seraient généralement plus précises que les estimations relatives à la petite région. L'estimateur synthétique serait alors un estimateur à faible variance, qui pourrait toutefois être fortement biaisé si l'hypothèse mentionnée plus haut ne se vérifiait pas.

L'une des formes les plus simples d'estimateur synthétique repose sur l'hypothèse que la moyenne pour petite région est égale à la moyenne globale. On a ainsi l'estimateur synthétique de moyenne

$$\hat{Y}_{syn,m,a} = N_a \sum_{i \in s} w_i y_i \Big/ \sum_{i \in s} w_i = N_a \bar{y}. \tag{6.5}$$

Une forme plus courante d'estimateur synthétique est celle qui repose sur la stratification ou la stratification a posteriori, soit

$$\hat{Y}_{syn,st,m,a} = \sum_h N_{h,a} \sum_{i \in s_h} w_i y_i \Big/ \sum_{i \in s_h} w_i = \sum_h N_{h,a} \bar{y}_h.$$

Comme dans le cas des estimateurs directs, les estimateurs synthétiques par quotient peuvent utiliser des données auxiliaires en plus des effectifs N_a et $N_{h,a}$. Par exemple, les estimateurs synthétiques par quotient basés sur une covariable x sont définis le plus souvent sous la forme

$$\hat{Y}_{syn,r,a} = X_a \hat{Y}_e / \hat{X}_e \quad \text{et} \quad \hat{Y}_{syn,st,r,a} = \sum_h X_{h,a} \hat{Y}_{h,e} / \hat{X}_{h,e}, \tag{6.6}$$

où $\hat{Y}_e = \sum_{i \in s} w_i y_i$ est l'estimateur par facteur d'extension du total de population pour y et $\hat{Y}_{h,e} = \sum_{i \in s_h} w_i y_i$. $\hat{X}_e$ et $\hat{X}_{h,e}$ sont définis de la même manière. Ces estimateurs ont été étudiés par Gonzalez (1973), Gonzalez et Waksburg (1973) et Ghangurde et Singh (1977,1978), pour n'en nommer que quelques-uns.

Singh et Tessier (1976) ont proposé une autre forme d'estimateur synthétique par quotient, où X remplace $\hat{X}_e$:

$$\tilde{Y}_{syn,r,a} = X_a \hat{Y}_e / X. \tag{6.7}$$

Les deux estimateurs, $\hat{Y}_{syn,r,a}$ et $\tilde{Y}_{syn,r,a}$, ont le même biais synthétique et le biais de $\hat{Y}_{syn,r,a}$ est négligeable pour de grands échantillons. Le choix entre l'un et l'autre de ces estimateurs dépend de ρ, le coefficient de corrélation de $\hat{Y}_e$ et $\hat{X}_e$. On peut montrer que pour de grands échantillons, $V(\hat{Y}_{syn,r,a}) \leq V(\tilde{Y}_{syn,r,a})$ si $\rho \geq 0.5 c_x / c_y$, où c_x et c_y sont les coefficients de variation de $\hat{X}_e$ et de $\hat{Y}_e$ respectivement. Dans la plupart des cas, lorsque ρ est élevé ou que la population est asymétrique, on préférera $\hat{Y}_{syn,r,a}$; cependant, si c_x est élevé et que la corrélation n'est pas trop forte, $\tilde{Y}_{syn,r,a}$ pourra être un meilleur choix.

Dans certains cas, il peut exister de l'information sur une autre variable auxiliaire (z) en plus de x. On peut alors construire un estimateur synthétique par quotient à deux variables:

$$\hat{Y}^{(2)}_{syn,r,a} = \gamma_a X_a \hat{Y}_e / \hat{X}_e + (1 - \gamma_a) Z_a \hat{Y}_e / \hat{Z}_e, \tag{6.8}$$

où la valeur de γ_a est choisie judicieusement. On peut aussi envisager un estimateur synthétique par quotient à plusieurs variables si l'on se fonde sur l'article d'Olkin (1958).

L'estimation synthétique par régression ressemble à l'estimation synthétique par quotient,

$$\hat{Y}_{syn,reg,a} = \hat{\beta} X_a,$$

$$\hat{\beta} = \sum_{i \in s} v_i^{-1} w_i y_i x_i' \left\{ \sum_{i \in s} v_i^{-1} w_i x_i x_i' \right\}^{-1}. \tag{6.9}$$

Comme dans le cas des estimateurs directs, l'estimateur synthétique par régression peut être appliqué à des strates initiales ou à des strates formées a posteriori. Royall (1979) a proposé un variante de cet estimateur, soit $\hat{Y}_{syn,Roy,a} = \sum_{i \in s_a} y_i + \hat{\beta}(X_a - \sum_{i \in s_a} x_i)$, où la somme des valeurs y qui se rapportent uniquement aux unités non incluses dans l'échantillon est estimée de façon synthétique.

Remarque: Les estimateurs directs modifiés présentés dans la section 6.1 peuvent aussi être exprimés comme des estimateurs synthétiques par quotient ou des estimateurs synthétiques par régression comportant un facteur de compensation du biais basé sur le plan. Nous pouvons écrire par exemple l'expression $\hat{Y}_{sreg,a} = \hat{Y}_{syn,reg,a} + (\hat{Y}_a - \hat{\beta}\hat{X}_a)$, où $\hat{Y}_a - \hat{\beta}\hat{X}_a$ est la valeur estimée du biais de $\hat{Y}_{syn,reg,a}$. De même, nous pouvons exprimer $\hat{Y}_{sreg,a}$ comme l'estimateur de Royall, $\hat{Y}_{syn,Roy,a}$, accompagné d'un facteur de compensation du biais basé sur le plan.

Purcell et Kish (1980) étudient une autre forme d'estimation synthétique qu'ils appellent SPREE (pour "structure preserving estimation") et qui s'applique à l'estimation d'effectifs pour petites régions. Selon cette méthode, on combine des données antérieures détaillées, provenant peut-être d'un recensement, avec des estimations d'enquête courantes moins détaillées dans le but d'obtenir des estimations détaillés de valeurs courantes. Cette méthode suppose que certains rapports parmi les données détaillées ne varient pas dans le temps.

Estimateurs combinés: L'estimateur combiné est défini comme la moyenne pondérée d'un estimateur de plan et d'un estimateur synthétique, c'est-à-dire

$$\hat{Y}_{com,a} = \lambda_a \hat{Y}_{des,a} + (1 - \lambda_a) \hat{Y}_{syn,a}, \tag{6.10}$$

où la valeur de λ_a est choisie judicieusement. L'estimateur combiné vise à faire l'équilibre entre la possibilité pour l'estimateur synthétique d'être biaisé et l'instabilité de l'estimateur de plan. Il existe trois grandes méthodes pour déterminer les poids λ_a en (6.10): ces poids peuvent être établis à l'avance, ils peuvent dépendre de la taille de l'échantillon ou encore, ils peuvent dépendre des données.

La première méthode de pondération, et la plus simple, consiste à fixer des poids à l'avance (par exemple, utiliser une moyenne simple). Or, cette méthode ne tient aucunement compte de la précision réelle de l'estimateur de plan. L'estimateur de plan pour la petite région a est plus précis pour certains échantillons réels que pour d'autres. Le poids attribué à l'estimateur de plan devrait refléter cette différence.

La deuxième grande méthode de pondération des composantes de l'estimateur combiné a un rapport de dépendance avec la taille de l'échantillon; en effet, les poids sont fonction du rapport $\hat{N}_{e,a}/N_a$. Une variante de cette méthode, que nous n'examinons pas ici, a plutôt un rapport de dépendance avec les valeurs empiriques d'une covariable x; par exemple, les poids peuvent être fonction de $\hat{X}_{des,a}/X_a$ ou de $S^2_{x,a}/\sigma^2_{x,a}$, où $S^2_{x,a}$ est la variance empirique de $\hat{X}_{des,a}$, étant donné $\hat{N}_{e,a}$ ou une autre caractéristique pertinente de l'échantillon réel, et $\sigma^2_{x,a}$ est la variance inconditionnelle de $\hat{X}_{des,a}$.

Des estimateurs de ce genre ont été proposés antérieurement. Drew, Singh et Choudhry (1982) ont proposé, par exemple, l'estimateur suivant (tributaire de la taille de l'échantillon):

$$\hat{Y}_{ssd,r,a} = \lambda_a \hat{Y}_{r,a} + (1 - \lambda_a)\hat{Y}_{syn,r,a}, \qquad (6.11a)$$

où

$$\lambda_a = \begin{cases} 1 & \text{si} \quad \hat{N}_{e,a} \geq \delta N_a \\ \hat{N}_{e,a}/\delta N_a & \text{dans le cas contraire} \end{cases} \qquad (6.11b)$$

et la valeur δ est choisie subjectivement pour régler la contribution de la composante synthétique. Särndal (1984) a proposé l'estimateur suivant

$$\hat{Y}_{ssd,reg,a} = \lambda_a \hat{Y}_{sreg,a} + (1 - \lambda_a)\hat{Y}_{syn,reg,a}, \qquad (6.12)$$

où $\lambda_a = \hat{N}_{e,a}/N_a$. Rao (1986) a proposé de modifier cet estimateur dans le sens suivant: λ_a aurait la valeur 1 si $\hat{N}_{e,a} \geq N_a$. Särndal et Hidiroglou (1989) ont amélioré la proposition de Rao en ajoutant la relation suivante: $\lambda_a = (\hat{N}_{e,a}/N_a)^{h-1}$ si $\hat{N}_{e,a} < N_a$, la valeur h étant choisie de façon rationnelle pour régler la contribution de la composante synthétique.

Dans la pratique, ce qui compte le plus lorsqu'on utilise des estimateurs de ce genre est le biais de la composante synthétique. Cette composante devrait être affectée d'un poids tel que le biais se situe dans des limites raisonnables. Par exemple, on se sert actuellement de l'estimateur tributaire de la taille d'échantillon de Drew, Singh et Choudhry (1982) dans l'Enquête sur la population active du Canada pour produire des estimations pour domaine, compte tenu de ce que l'estimation par régression généralisée remplace l'estimation par quotient et que $\delta = 2/3$. Pour une majorité de domaines, le poids attribué à la composante synthétique est nul puisque l'estimateur direct satisfait à lui seul les exigences de fiabilité. Pour les autres domaines, le poids de la composante synthétique est d'environ 10% en moyenne et ne dépasse jamais 20%. Selon le biais qu'on est prêt à accepter, la valeur de δ peut se situer dans l'intervalle [2/3,3/2] pour la plupart des cas.

La troisième méthode de pondération met les poids dans un rapport de dépendance avec les données. Les poids optimaux nécessaires à la combinaison de deux estimateurs dépendent généralement de l'erreur quadratique moyenne des estimateurs et de leur covariance. Ces valeurs sont le plus souvent inconnues mais elles peuvent être estimées à partir des données. Dans le cas des estimateurs combinés, il faudrait normalement modéliser le biais de la composante synthétique. On trouve dans Fay et Herriot (1979) un exemple très connu d'application de cette méthode, qui est un des premiers du genre dans la littérature statistique. Fay et Herriot modélisent les biais des estimateurs synthétiques pour les petites régions comme des effets aléatoires indépendants ayant une variance inconnue mais fixe. En termes plus précis, si $\hat{Y}_{des,a}$ est l'estimateur de plan, ils considèrent alors le modèle $Y_a = X_a\beta + \alpha_a$ et l'estimateur $\hat{Y}_{des,a} = Y_a + \epsilon_a$, où $\alpha_a \sim (0,\sigma^2)$, $\epsilon_a \sim (0,\nu_a^2)$, et α_a et ϵ_a sont indépendants et non corrélés pour a, σ^2 est inconnu et les ν_a^2 sont supposés connus (en pratique, ils devraient être estimés). Pour une valeur donnée de σ^2, on peut calculer les poids optimaux nécessaires à la combinaison de $\hat{Y}_{des,a}$ et de $X_a\hat{\beta}$. On calcule une estimation de σ^2 à l'aide de la méthode des constantes d'ajustement, puis on introduit cette estimation dans la formule des poids optimaux. Pour parer à la défaillance du modèle, on prévoit une troncation de l'estimation calculée si l'écart entre celle-ci et l'estimation directe est supérieur à un multiple déterminé de ν_a. Schaible (1979) et Battese et Fuller (1981) considèrent aussi des poids optimaux λ_a (en 6.12) estimés empiriquement selon des modèles à effets aléatoires semblables pour les totaux de petites régions.

Prasad et Rao (1990) définissent un estimateur de l'erreur quadratique moyenne de l'estimateur de Fay-Herriot qui tient compte de l'estimation des composantes de la variance. Kott (1989) propose un estimateur convergent selon le plan de l'erreur quadratique moyenne mais constate que cet estimateur est très instable.

Une autre méthode consiste à calculer les poids au moyen de données historiques; l'avantage de cette solution est que les poids calculés de cette manière peuvent être plus stables que des poids estimés à l'aide des données d'enquête courantes; néanmoins, cela suppose l'hypothèse que les poids optimaux sont stables.

Remarque: Dans l'estimation tributaire de la taille de l'échantillon, les poids peuvent dépendre de la taille observée du sous-échantillon s_a mais non des valeurs de la variable étudiée. Cette absence de relation de dépendance entre les poids et la variable étudiée présente des avantages et des inconvénients. Un des avantages est que les mêmes poids servent à l'estimation de totaux pour toutes les variables étudiées. On n'a à les calculer qu'une seule fois. Un avantage plus important encore est que l'estimation de la somme de deux variables est égale à la somme des estimations de chaque variable. Du côté des inconvénients, les poids ne tiennent pas compte directement de la précision de l'estimateur de plan pour la variable étudiée ou de la grandeur probable du biais de l'estimateur synthétique.

Combinaison de données de diverses périodes: Dans les enquêtes à passages répétés, il est courant de grouper les données de plusieurs passages dans le but d'accroître la précision des estimations. Selon le plan de renouvellement utilisé dans ces enquêtes, on peut réussir à accroître sensiblement la fiabilité des estimations. Cette technique de groupement ou de mise en moyenne sur plusieurs périodes suscite donc un intérêt particulier en ce qui concerne l'estimation pour domaine, caractérisée par un degré de précision modeste. En ce qui regarde l'estimation pour domaine dans l'Enquête sur la population active du Canada, il est pratique courante d'utiliser un estimateur tributaire de la taille de l'échantillon basé sur une moyenne trimestrielle des estimations du nombre de personnes occupées et du nombre de personnes en chômage. Grâce au plan de renouvellement de six mois utilisé dans cette enquête, la mise en moyenne sur trois mois permet d'accroître du tiers

la taille de l'échantillon. Si les échantillons des différentes périodes se chevauchent parfaitement, la mise en moyenne ne se traduit par aucun gain d'efficacité. Cependant, pour d'autres plans de renouvellement, ce procédé pourrait plus que doubler la taille de l'échantillon pour des estimations pour domaine. Mais les estimations groupées posent un problème théorique en ceci qu'elles représentent une valeur moyenne du paramètre étudié (par ex., niveau de chômage) pour une période de trois mois par exemple.

Dans l'estimation composite, on combine l'estimateur de plan de la période courante avec l'estimateur composite de la période précédente, ce dernier étant redressé au moyen d'une estimation de la variation basée sur l'échantillon commun. Ce procédé a été appliqué pour autre chose que l'estimation pour petites régions par Jessen (1942) et Patterson (1950), pour ne nommer que ceux-là. Binder et Hidiroglou (1988) font un examen de la question. Les poids nécessaires à la combinaison sont ordinairement des estimations des poids optimaux dans l'hypothèse que ceux-ci ne varient pas dans le temps. Ces poids tributaires des données ont pour inconvénient de produire des estimations incohérentes pour diverses caractéristiques et les sommes de ces caractéristiques.

Dans l'estimation pour petites régions, on a recours depuis peu à des méthodes basées sur des séries temporelles pour les enquêtes périodiques. On modélise la relation entre les paramètres étudiés pour différentes périodes, puis on se sert du modèle afin d'accroître l'efficience des estimations pour le passage courant. Dans la plupart des cas, il faut aussi tenir compte, par la modélisation ou autrement, de la non-indépendance des échantillons des divers passages, condition attribuable au plan de renouvellement de l'échantillon. Cette approche fondée sur les séries chronologiques est traitée dans les ouvrages suivants: Choudhry et Rao (1989); Pfeffermann et Burck (1990); Singh, Mantel et Thomas (1994) et Singh et Mantel (1991). Tous ces ouvrages présentent des versions généralisées du modèle de Fay-Herriot selon lesquelles les paramètres de régression, les effets de petite région et les erreurs d'enquête varient selon divers modèles chronologiques. Le vecteur d'estimations régionales calculé par cette méthode peut être exprimé comme la moyenne pondérée du vecteur des estimations de plan et du vecteur des estimations synthétiques, celles-ci étant fondées sur des données de périodes antérieures et les valeurs courantes des covariables; cependant, la matrice des poids sera très rarement une matrice diagonale, de sorte que l'estimateur pour une petite région quelconque dépendra généralement aussi bien des estimations de plan et des estimations synthétiques d'autres petites régions.

7. CONCLUSION

Les concepteurs de plan de sondage doivent surmonter beaucoup d'obstacles pour produire des estimations pour domaine qui soient précises et actuelles. Ils doivent tout d'abord convaincre les promoteurs d'enquêtes et les chargés de programme qu'un système conçu spécialement pour produire des estimations nationales et infra-nationales ne peut répondre à tous les besoins en données régionales. On peut réaliser des gains d'efficience appréciables (selon les enquêtes) au niveau du domaine avec très peu de perte aux niveaux supérieurs. Nous avons souligné la nécessité d'élaborer une approche globale, par laquelle on pourrait définir un degré de fiabilité pour les domaines planifiés comme pour les régions de niveau supérieur par des répartitions intermédiaires et par laquelle on pourrait réduire la formation de grappes dans le but d'améliorer les estimations relatives aux domaines non planifiés. Il convient de noter que bon nombre de domaines dits planifiés à l'étape de la conception du plan de sondage peuvent se transformer en domaines non planifiés (révisés) au fil du temps lorsqu'il s'agit d'enquêtes permanentes.

L'approche globale devrait aussi permettre de considérer les estimateurs de plan pour de grands domaines aussi bien que les estimateurs de modèle pour de petits domaines. L'estimateur de modèle devrait être préféré à l'estimateur de plan seulement si l'erreur quadratique moyenne du premier (variance de plan + biais^2) peut être estimée et qu'elle est inférieure à la variance correspondante de l'estimateur de plan **par une marge suffisante**. Il devrait y avoir des estimations de l'erreur quadratique moyenne pour chacun des domaines **pris individuellement**. Les organismes statistiques peuvent choisir de grouper des domaines semblables ou de grouper des estimations de périodes différentes qui se rapportent au même domaine. Ils peuvent même supprimer les estimations touchant certains domaines à cause de leur faible degré de fiabilité ou pour des raisons de protection du secret statistique.

La deuxième tâche qui attend les statisticiens est d'expliquer aux utilisateurs les divers types de mesures de fiabilité qui peuvent s'appliquer à différentes séries d'estimations provenant de la même enquête. Espérons qu'avec la poursuite des recherches sur la validation de modèle et de meilleurs estimations de l'erreur quadratique moyenne, les concepteurs hésiteront moins à utiliser des estimateurs de modèle pour les petits domaines. D'ici là, on devra se servir des estimateurs de modèle avec circonspection même si leurs coefficients de variation sont beaucoup plus petits.

Les recensements, avec les fichiers administratifs en appui, demeureront vraisemblablement la principale source de données socio-économiques pour les petites régions, surtout dans les pays où il se fait un recensement de la population et du logement à tous les cinq ans. En outre, on continuera de se préoccuper des problèmes conceptuels que posent les données des fichiers administratifs tant que les organismes statistiques n'auront pas la chance de participer à l'élaboration des formulaires qui servent à recueillir ces données. En attendant, on ne peut exploiter pleinement cette source de données extrêmement riche que constituent les fichiers administratifs; ce qui est dommage pour l'analyse statistique et surtout pour l'estimation pour domaine.

REMERCIEMENTS

Nous sommes reconnaissants à Jon Rao d'avoir agi comme rédacteur associé pour cet article et nous remercions aussi les rapporteurs pour leurs nombreux conseils utiles.

BIBLIOGRAPHIE

BATTESE, G.E., et FULLER, W.A. (1981). Prediction of county crop areas using survey and satellite data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 500-505.

BINDER, D., et HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, (Éds. P.R. Krishnaiah et C.R. Rao). New York: Elsevier Science, 187-211.

BRACKSTONE, G.J. (1987a). Small area data: Policy issues and technical challenges. *Small Area Statistics*, (Éds. R. Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh), New York: Wiley and Sons, 3-20.

BRACKSTONE, G.J. (1987b). Utilisations statistiques des données administratives: questions et défis. *Recueil: Symposium sur les utilisations statistiques des données administratives*, (Éds. J.W. Coombs et M.P. Singh), Statistique Canada, 5-17.

CHOUDHRY, G.H., et RAO, J.N.K. (1989). Estimation de données régionales à l'aide de modèles qui combinent des séries chronologiques et des données transversales. *Recueil: Symposium sur l'analyse des données dans le temps*, (Éds. A.C. Singh et P. Whitridge), Statistique Canada, 71-80.

COOMBS, J.W., et SINGH, M.P. (Éds.) (1987). *Recueil: Symposium sur les utilisations statistiques des données administratives*, Statistique Canada.

DREW, J.D., SINGH, M.P., et CHOUDHRY, G.H. (1982). Evaluation of small area estimation techniques for the Canadian Labor Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 545-550.

FAY, R.E., et HERRIOT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

FELLEGI, I.P. (1987). Allocution d'ouverture. *Recueil: Symposium sur les utilisations statistiques des données administratives*, (Éds. J.W. Coombs et M.P. Singh), Statistique Canada, 1-2.

GHOSH, M., et RAO, J.N.K. (1993). Small area estimation: An appraisal. À paraître dans *Statistical Science*.

GHANGURDE, P.D., et SINGH, M.P. (1977). Synthetic estimates in periodic household surveys. *Techniques d'enquête*, 3, 152-181.

GHANGURDE, P.D., et SINGH, M.P. (1978). Evaluation of efficiency of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, 53-61.

GONZALEZ, M.E. (1973). Use and evaluation of synthetic estimators. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.

GONZALEZ, M.E., et WAKSBERG, J. (1973). Estimation of the error of synthetic estimates. Présenté à la première réunion de l'Association Internationale des Statisticiens d'Enquête, Vienna, Austria.

JESSEN, R.J. (1942). Statistical investigation of a farm survey for obtaining farm facts. *Iowa Agricultural Station Research Bulletin*, 304, 54-59.

NORRIS, D., et PATON, D. (1991). L'enquête sociale générale canadienne: bilan des cinq premières années. *Techniques d'enquête*, 17, 245-260.

OLKIN, I. (1958). Multivariate ratio estimation for finite populations. *Biometrika*, 45, 154-165.

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

PFEFFERMANN, D., et BURCK, L. (1990). Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales. *Techniques d'enquête*, 16, 229-249.

PLATEK, R., RAO, J.N.K., SÄRNDAL, C.-E., et SINGH, M.P. (1987). *Small Area Statistics*. Présentations invitées. New York: Wiley.

PLATEK, R., et SINGH, M.P. (1986). Small Area Statistics, An International Symposium' 85 (articles contribués), Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University, University of Ottawa, Canada.

PRASAD, N.G.N., et RAO, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

PURCELL, N.J., et KISH, L. (1980). Postcensal estimates for local areas (or Domains). *Revue Internationale de Statistique*, 48, 3-18.

RAO, J.N.K. (1986). Synthetic estimates, SPREE and best model based predictors. *Proceedings of the Conference on Survey Research Methodology in Agriculture*, American Statistical Association and National Agricultural Statistics Service, USDA, 1-6.

ROYALL, R.M. (1979). Prediction models in small area estimation. Dans *Synthetic Estimates for Small Area*, NIDA Research Monograph Series 24, U.S. Department of Health, Education and Welfare.

SÄRNDAL, C.-E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association*, 79, 624-631.

SÄRNDAL, C.-E., et HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

SCHAIBLE, W.L. (1979). A composite estimator for small area statistics. *Synthetic Estimates for Small Area*, NIDA Research Monograph Series 24, U.S. Department of Health, Education and Welfare, Library of Congress catalogue number 79-600067, 36-53.

SCHAIBLE, W.L. (1992). Use of small area estimators in U.S. federal programs. *Small Area Statistics and Survey Designs*, (Vol. 1), (Éds. G. Kalton, J. Kordos et R. Platek). Warsaw: Central Statistical Office, 95-114.

SCHMIDT, R.C. (1952). Short-cut methods for estimating county populations. *Journal of the American Statistical Association*, 47, 232-238.

SINGH, A.C., et MANTEL, H.J. (1991). Estimation composite par espace d'états pour les petites régions. *Recueil: Symposium 91, Questions spatiales liées aux statistiques*, Statistique Canada, 21-30.

SINGH, A.C., MANTEL, H.J., et THOMAS, B.W. (1994). MPLSE à données chronologiques pour petites régions évalués d'après des données d'enquête. *Techniques d'enquête*, 20, 35-46.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., et MAYDA, F. (1990). *Méthodologie de l'enquête sur la population active du Canada*. N° 71-526 au catalogue, Statistique Canada.

SINGH, M.P., GAMBINO, J.G., et MANTEL, H. (1992). Issues and options in the provision of small area statistics. *Small Area Statistics and Survey Designs*, (Vol. 1), (Éds. G. Kalton, J. Kordos et R. Platek), Warsaw: Central Statistical Office, 37-75.

SINGH, M.P., et TESSIER, R. (1976). Some estimators for domain totals. *Journal of the American Statistical Association*, 71, 322-325.

SWAIN, L., DREW, J.D., LAFRANCE, B., et LANCE, K. (1992). La création d'un registre des adresses résidentielles pour améliorer la couverture du recensement du Canada de 1991. *Techniques d'enquête*, 18, 139-155.

U.S. STATISTICAL POLICY OFFICE (1993). Indirect Estimators in Federal Programs. Statistical Policy Working Paper 21. Préparé par le sous-comité sur Small Area Estimation, Federal Committee on Statistical Methodology.

VERMA, R.B.P., et BASAVARAJAPPA, J.G. (1987). Recent developments in the regression method for estimation of population for small areas in Canada. *Small Area Statistics*, (Éds. R. Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh), New York: Wiley and Sons, 46-61.

COMMENTAIRES

W.A. FULLER[1]

Les auteurs méritent des félicitations pour avoir exposé avec brio les problèmes de conception de plan et d'estimation qui ont rapport aux domaines. Les auteurs examinent l'estimation pour domaines planifiés en particulier les cas où les éléments d'un domaine peuvent être identifiés dans la base de sondage de même que l'estimation pour domaines non planifiés, notamment les domaines dont les éléments ne peuvent être identifiés à l'aide de la base de sondage. Cet article est un élément positif supplémentaire dans la littérature de plus en plus abondante qui existe sur la question de l'estimation pour domaine.

Les auteurs font une description particulièrement bonne des opérations de planification, de collecte des données et de traitement qui sont exécutées dans le cadre des enquêtes de Statistique Canada. Ils mentionnent les problèmes qui sont liés depuis toujours à la conception des plans de sondage: arbitrage entre la nécessité de recourir à l'estimation pour domaine et le désir d'obtenir une certaine efficience aux niveaux (d'agrégation) supérieurs, importance de la confidentialité dans l'utilisation de dossiers administratifs en vue d'établir des estimations pour domaine, et importance de la compatibilité des définitions lorsqu'on tente de combiner des données de diverses sources.

Les auteurs font ressortir très bien l'importance de prendre en considération l'estimation pour domaine à l'étape de la conception du plan de sondage, ce que font rarement d'autres auteurs qui s'intéressent particulièrement à l'estimation pour petite région. Comme le soulignent Singh, Gambino et Mantel, le statisticien pourra souvent construire des estimateurs pour domaine qui sont directs et convergents selon le plan s'il exécute avec soin l'étape de la conception. Je suis sûr que les concepteurs de plan de sondage tiennent compte de l'importance de la formation de grappes lorsqu'ils élaborent des enquêtes qui doivent servir à l'estimation pour domaine, mais il est réjouissant de voir qu'on en fait mention explicitement.

Les auteurs décrivent plusieurs types d'estimateur pour domaine. La classification de ces estimateurs fait ressortir le nombre de possibilités qui s'offrent au praticien. Les erreurs quadratiques moyennes théoriques peuvent nous renseigner sur les avantages relatifs des estimateurs. Afin d'illustrer ce genre de comparaison, supposons un échantillon aléatoire simple de taille n prélevé dans une population qui est divisée en k domaines. Supposons qu'on connaît la taille des domaines et la valeur moyenne d'une variable auxiliaire, X, pour le domaine. Considérons les trois estimateurs par régression de la moyenne de domaine

$$\hat{\mu}_{(1)yi} = \bar{y}_{i.} + (\mu_{xi} - \bar{x}_{i.})b_i,$$

$$\hat{\mu}_{(2)yi} = \bar{y}_{i.} + (\mu_{xi} - \bar{x}_{i.})b_.$$

et

$$\hat{\mu}_{(3)yi} = \bar{y}_{..} + (\mu_{xi} - \bar{x}_{..})b_.,$$

où

$$(\bar{x}_{..}, \bar{y}_{..}) = \sum_{i=1}^{k} N^{-1}N_i(\bar{x}_{i.}, \bar{y}_{i.}),$$

$$(\bar{x}_{i.}, \bar{y}_{i.}) = n_i^{-1} \sum_{j=1}^{n_i} (X_{ij}, Y_{ij}),$$

$$b_i = \left[\sum_{j=1}^{n_i} (X_{ij} - \bar{x}_{i.})^2 \right]^{-1} \times \sum_{j=1}^{n_i} (X_{ij} - \bar{x}_{i.})(Y_{ij} - \bar{y}_{i.}),$$

$$b_. = \left[\sum_{i=1}^{k} N^{-1}N_i\, n_i^{-1} \sum_{j=1}^{n_i} (X_{ij} - \bar{x}_{i.})^2 \right]^{-1} \times \sum_{i=1}^{k} N^{-1}N_i\, n_i^{-1} \sum_{j=1}^{n_i} (X_{ij} - \bar{x}_{i.})(Y_{ij} - \bar{y}_{i.}),$$

n_i est le nombre d'observations dans le domaine i, N_i est le nombre d'éléments de la population contenus dans le domaine i, μ_{xi} est la moyenne de population de X pour le domaine i et $\mu_{x.}$ est la moyenne de X pour l'ensemble de la population. Selon la terminologie utilisée par les auteurs, le premier estimateur est un estimateur par régression direct, le deuxième, un estimateur direct modifié et le troisième, un estimateur synthétique. Nous avons

$$\mathrm{EQM}\{\hat{\mu}_{(1)yi} \mid n_i\} = n_i^{-1}(1 + n_i^{-1})V\{Y_{\ell j} - \beta_\ell X_{\ell j} \mid \ell = i\} + O(n_i^{-2}),$$

$$\mathrm{EQM}\{\hat{\mu}_{(2)yi} \mid n_i\} = n_i^{-1}(1 + n^{-1})V\{Y_{\ell j} - \beta X_{\ell j} \mid \ell = i\} + O(n^{-2}),$$

[1] W.A. Fuller, Distinguished Professor, Statistical Laboratory and Department of Statistics, Iowa State University, Snedecor Hall, Ames, Iowa.

$$\text{EQM}\{\hat{\mu}_{(3)yi} | n_i\} = (1 + n^{-1})$$

$$\times \sum_{i=1}^{k} N^{-2} N_i^2 n_i^{-1} V\{Y_{\ell j} - \beta X_{\ell j} | \ell = i\}$$

$$+ (\mu_{xi} - \mu_{x.})^2 V\{b_.\}$$

$$+ [\mu_{yi} - \mu_{y.} - \beta(\mu_{xi} - \mu_{x.})]^2 + O(n^{-2}),$$

où $V\{b_.\} = E\{(b_. - \beta)^2\}$, $V\{a_\ell | \ell = i\}$ est la variance de la variable a pour le domaine i,

$$\beta_i = [V\{X_{\ell j} | \ell = i\}]^{-1} C\{Y_{\ell j}, X_{\ell j} | \ell = i\}$$

et

$$\beta = \left[\sum_{i=1}^{k} N^{-1} N_i V\{X_{\ell j} | \ell = i\}\right]^{-1}$$

$$\times \sum_{i=1}^{k} N^{-1} N_i C\{Y_{\ell j}, X_{\ell j} | \ell = i\}.$$

L'estimateur $\hat{\mu}_{(1)yi}$ utilise uniquement l'information contenue dans l'échantillon de n_i observations. Par conséquent, toutes les propriétés de l'estimateur sont fonction de n_i et des paramètres du domaine. Le biais de régression est d'ordre n_i^{-1}, comme la variance. L'estimateur $\hat{\mu}_{(2)yi}$ utilise les moyennes de domaine, mais utilise aussi l'échantillon complet pour l'estimation du coefficient de régression. En conséquence, la variance fondamentale est aussi d'ordre n_i^{-1} et sera plus élevée que la variance fondamentale de $\hat{\mu}_{(1)yi}$ dans les cas où $\beta_i \neq \beta$. Or, la contribution de deuxième ordre à la variance est d'ordre $n_i^{-1}\, n^{-1}$ pour $\hat{\mu}_{(2)yi}$ et d'ordre n_i^{-2} pour $\hat{\mu}_{(1)yi}$. De plus, le biais de régression pour $\hat{\mu}_{(2)yi}$ est d'ordre n^{-1}. Si les domaines étaient des strates, on pourrait désigner $\hat{\mu}_{(1)yi}$ comme l'estimateur par régression simple et $\hat{\mu}_{(2)yi}$ comme l'estimateur par régression combiné.

L'estimateur $\hat{\mu}_{(3)yi}$ est un estimateur synthétique qui a une variance d'ordre n^{-1}, contrairement aux deux premiers estimateurs, qui ont une variance d'ordre n_i^{-1}. Cependant, cette réduction de la variance a un inconvénient: le biais est d'ordre un. On aura un biais nul seulement si la droite de régression est la même pour le domaine et pour la population entière.

On peut estimer la moyenne des erreurs quadratiques moyennes des trois estimateurs pour n'importe quel sous-groupe de petites régions. Si les n_i sont faibles, les variances estimées ne donneront qu'une information restreinte pour faire la différence entre les estimateurs. En outre, il n'existe qu'un degré de liberté pour le carré du biais pour un domaine particulier. Cependant, si l'écart pour un domaine devait être élevé par rapport à l'erreur-type, il faudrait reconsidérer l'estimateur synthétique.

Dans leur analyse de modèles, les auteurs soulignent l'importance de produire des estimateurs de la fiabilité des estimateurs pour petites régions. Ils laissent entendre que les principaux estimateurs de l'erreur quadratique moyenne pour des méthodes fondées sur un modèle sont des estimateurs de la valeur moyenne de l'erreur quadratique moyenne. Bien que ce soit vrai, il convient de mentionner que la méthode des "composantes de la variance" ne suppose pas que l'erreur quadratique moyenne est la même dans chaque domaine. En outre, dans une enquête classique, les estimateurs de l'erreur quadratique moyenne ne sont pas nécessairement les mêmes dans tous les domaines. Par exemple, un des termes de l'estimateur de l'erreur quadratique moyenne utilisé dans la méthode des "composantes de la variance" est l'estimateur de la variance de l'estimateur direct. La variance estimée de l'estimateur direct sera une fonction de l'effectif de l'échantillon contenu dans le domaine et pourra aussi être une fonction de la variance estimée (directement) de l'estimateur direct pour ce domaine. Voir Battese, Harter et Fuller (1988), Harville (1976), Prasad et Rao (1990) et Ghosh et Rao (1993).

Dans leur analyse de plans de sondage, les auteurs expliquent que la courbe de la fonction de variance est souvent "aplatie" dans le voisinage de la répartition optimale des unités entre les strates. Une légère redistribution de l'échantillon peut avoir pour effet d'accroître sensiblement l'efficience des estimations pour domaine sans réduire de façon notable l'efficience des estimations globales. Il en va de même de la combinaison d'un estimateur direct et d'un estimateur synthétique. Par conséquent, si quelqu'un a une idée assez juste de la composante de variance liée aux petites régions, soit grâce à une étude antérieure qui portait sur la même population ou grâce à une étude qui portait sur une population similaire, et s'il doit produire des estimations dans un délai très limité, il est raisonnable, dans les circonstances, d'attribuer des poids fixes pour former la combinaison linéaire. Il y aura vraisemblablement peu de perte d'efficience et les exigences de programmation pour la construction d'estimateurs seront beaucoup moins grandes. L'estimateur synthétique est un estimateur indiqué dans les circonstances, et beaucoup de praticiens l'utilisent.

Les auteurs évoquent la question de la propriété de fermeture liée à la construction d'estimateurs pour petites régions. Comme ils l'affirment, si on utilise une méthode dépendante des données comme la méthode des composantes de la variance, on obtient, pour chaque variable dépendante, des estimations qui n'ont pas la propriété de fermeture. Dans les circonstances, on peut toujours recourir à des méthodes multidimensionnelles. Voir, par exemple, Fuller et Harter (1987) et Fay (1987). Une autre méthode proposée par Fuller (1990) consiste à construire des estimateurs de composantes de la variance pour un sous-ensemble limité de variables, puis de se servir des estimations correspondantes comme variables de contrôle dans une méthode de régression. Celle-ci produit des poids pour les observations individuelles. Une fois que les poids sont déterminés, on peut construire un nombre indéfini de tableaux de données et toutes les estimations ont la propriété de fermeture.

J'estime que dans la plupart des cas d'estimation pour domaine, les gains d'efficience découlent surtout de l'utilisation judicieuse de l'information supplémentaire. Par conséquent, les efforts consacrés à l'obtention d'information supplémentaire de qualité sont justifiés. Si nous pouvons trouver une variable x qui soit étroitement corrélée avec la variable y, il n'y aura plus autant de variabilité à répartir entre la variance d'une région à l'autre et la variance d'échantillonnage.

REMERCIEMENTS

Je remercie Jay Breidt pour ses commentaires.

BIBLIOGRAPHIE

BATTESE, G.E., HARTER, R.M., et FULLER, W.A. (1988). An error components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

FAY, R.E. (1987). Application of multivariate regression to small domain estimation. Dans *Small Area Statistics*. (Éds. R. Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh). New York: Wiley.

FULLER, W.A. (1990). Analyse d'enquêtes à passages répétés. *Techniques d'enquête*, 16, 177-190.

FULLER, W.A., et HARTER, R.M. (1987). The multivariate components of variance model for small area estimation. Dans *Small Area Statistics*. (Éds. R. Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh). New York: Wiley.

HARVILLE, D.A. (1976). Extension of the Gauss-Markov Theorem to include the estimation of random effects. *The Annals of Statistics*, 4, 384-395.

GHOSH, M., et RAO, J.N.K. (1993). Small area estimation: An appraisal. Manuscrit non-publié. Carleton University, Ottawa, Ontario, Canada.

PRASAD, N.G.N., et RAO, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

COMMENTAIRES

GRAHAM KALTON[1]

Comme le mentionnent Singh, Gambino et Mantel (SGM), on cherche de plus en plus à réaliser des enquêtes qui vont produire des estimations pour des domaines de tous genres et de toutes tailles. Cette tendance est observée dans de nombreux pays dans le monde. D'une part, elle peut être simplement le résultat d'une évolution naturelle des exigences des analystes d'enquête, qui auparavant se contentaient d'estimations nationales et d'estimations touchant un petit nombre de domaines majeurs mais qui, aujourd'hui, veulent comparer des estimations relatives à de nombreux types de domaines différents. D'autre part, cette tendance vient des besoins des responsables de l'action gouvernementale, qui recherchent de l'information pour domaine afin d'analyser les effets des programmes courants sur différents domaines, de prévoir les conséquences d'une révision des politiques et de suivre l'exécution des programmes. Les données relatives à des divisions administratives (par ex., province ou État, comté, arrondissement scolaire) présentent un intérêt particulier pour l'élaboration de programmes (par ex., lorsqu'il s'agit de définir les régions à faible revenu aux fins de l'élaboration des programmes de soutien public).

Dans certains cas, on peut réussir à obtenir des estimations pour domaine suffisamment précises à l'aide des méthodes d'inférence fondée sur un plan qui sont utilisées habituellement dans l'analyse de données d'enquête. C'est le cas notamment pour les grands domaines, où la taille des échantillons est suffisamment grande pour qu'on ait le degré de précision voulu. Cela est possible aussi pour de petits domaines, pourvu qu'ils soient définis à l'avance et que le plan de sondage soit élaboré de manière à produire des échantillons de taille acceptable. Ainsi, aux États-Unis par exemple, la National Health and Nutrition Examination Survey et la Continuing Survey of Food Intakes by Individuals utilisent des fractions de sondage qui varient selon l'âge, le sexe ou l'origine ethnique dans le premier cas et selon l'âge et le sexe ou le niveau de faible revenu dans le second cas, afin de produire des échantillons acceptables pour les domaines créés par le recoupement de ces variables. La Current Population Survey des É.-U. utilise des fractions de sondage différentes selon les États afin de produire des estimations de l'emploi pour l'État. La faiblesse de cette méthode ressort clairement lorsqu'il y a un grand nombre de petits domaines, auquel cas la somme des effectifs de chaque domaine dans l'échantillon donne une taille d'échantillon global très élevée. On observe souvent cela pour les petites divisions administratives comme les comtés, les arrondissements scolaires et les territoires de bureaux de placement. Dans de telles circonstances, on peut devoir remplacer les méthodes classiques d'inférence fondée sur un plan par des méthodes à modèles, qui utilisent pour l'estimation un modèle statistique par lequel on renforce les estimations d'une petite région par des données recueillies dans l'enquête. On peut devoir recourir aussi aux méthodes fondées sur un modèle pour de petits domaines non planifiés, pour lesquels on n'avait pas prévu la nécessité d'un suréchantillonnage à l'étape de la conception.

Compte tenu de l'intérêt pour les estimations régionales, un nombre appréciable d'ouvrages sont parus sur les méthodes d'estimation pour petit domaine fondées sur un modèle. Cependant, peu d'ouvrages ont été écrits jusqu'à maintenant sur les questions que traite l'article de SGM, lesquelles méritent une plus grande attention. Comme les auteurs, je crois qu'il faut envisager la question de l'utilisation d'estimateurs pour petite région fondés sur un modèle avec circonspection. C'est pourquoi je suis heureux de voir que SGM examinent des méthodes qui permettraient d'établir des estimations régionales au moyen de l'inférence fondée sur un plan.

À mon point de vue, la première étape dans le calcul d'estimations régionales est de vérifier si on peut obtenir des estimations ayant un degré de précision acceptable avec des méthodes fondées sur un plan. Si les domaines sont définis à l'avance, on devrait envisager de concevoir le plan d'échantillonnage de manière à faciliter la production d'estimations régionales. Cela peut signifier de faire en sorte que les petites régions ne chevauchent pas de strates et qu'il y ait un échantillon de taille suffisamment grande pour chaque petite région. Une autre solution proposée par SGM consiste à réduire au maximum la formation de grappes. Plus la formation de grappes est limitée, moins la taille de l'échantillon dans chaque petite région est laissée au hasard. À cet égard, le principal avantage de la limitation de la formation de grappes est de permettre la production d'estimations pour de petites régions qui n'étaient pas définies à l'étape de la conception du plan de sondage. Lorsque de petites régions pour lesquelles des estimations sont prévues ne chevauchent pas des strates, la taille de l'échantillon dans chaque petite région devrait être contrôlée raisonnablement même lorsqu'il s'agit d'un échantillon en grappes (pourvu que les mesures de taille utilisées dans l'échantillonnage avec PPT soient raisonnables). Or, même dans le cas d'estimations prévues, il arrivera souvent qu'on se demande comment calculer les estimations de la variance pour une

[1] Graham Kalton, Westat, 1650 Research Blvd., Rockville, Maryland, U.S.A. 20850.

petite région étant donné un plan avec grappes, puisque les UPÉ tirées dans chaque petite région seront vraisemblablement peu nombreuses. Une estimation de la variance fondée sur les UPÉ d'une petite région sera donc imprécise, étant donné quelques degrés de liberté, et on préférera probablement une fonction de variance généralisée (en supposant, par exemple, que l'effet de plan au niveau national vaut pour chaque petite région). Autrement dit, même si l'estimation proprement dite est fondée sur un plan, l'estimation de la variance correspondante peut être une estimation indirecte, renforcée par les données d'autres régions. C'est pourquoi il est préférable d'avoir un plan de sondage qui prévoit le moins de grappes possible, même lorsqu'il s'agit de petits domaines planifiés. Toutefois, il est plus important de penser à modéliser les estimations proprement dites que de se concentrer sur la modélisation des variances.

Un des principes fondamentaux de l'approche fondée sur un plan est que l'information supplémentaire qui existe sur la population peut servir à l'étape de la conception du plan de sondage, à celle de l'analyse ou aux deux à la fois. Lorsqu'il existe de l'information sur des variables auxiliaires qui ont un rapport étroit avec la variable étudiée, on peut réaliser des gains appréciables en précision. L'utilisation d'information supplémentaire à l'étape de l'analyse, par l'intermédiaire de techniques comme la stratification a posteriori, l'estimation par quotient ou par régression et l'estimation de différences, a un attrait particulier en ce qui concerne l'estimation pour petite région. Il convient de souligner que les estimateurs par quotient et par régression peuvent trouver leur justification dans des hypothèses portant sur le modèle qui met en relation la variable étudiée (Y) et les variables auxiliaires (X), mais les estimateurs seront convergents selon le plan peu importe si le modèle convient ou non. L'utilisation d'un modèle approprié permet de réaliser les meilleurs gains en précision mais quel que soit le modèle choisi, les estimations sont approximativement non biaisées. On peut illustrer cela par un exemple simple, où la valeur des variables $X_1, X_2, \ldots, X_p$ est connue pour chaque élément de la population et la combinaison linéaire $\tilde{Y}_i = B_0 + B_1X_{1i} + \ldots + B_pX_{pi}$ est utilisée pour estimer Y_i, la valeur de la variable Y pour l'élément i de la population. Supposons pour des raisons de simplicité que les valeurs de B sont déterminées à l'aide de données externes, indépendantes de l'échantillon. Étant donné l'équation $Y_i = \tilde{Y}_i + e_i$, le total de domaine est $Y_a = \sum_{i\epsilon a}\tilde{Y}_i + \sum_{i\epsilon a} e_i = \tilde{Y}_a + E_a$. Comme $\tilde{Y}_a$ est connu, le problème revient à estimer E_a. À partir d'un échantillon d'éléments du domaine a, on peut estimer E_a au moyen de l'équation $\hat{E}_a = \sum_{j\epsilon s_a} e_j/\pi_j$, où π_j est la probabilité de sélection pour l'élément j de l'échantillon. L'estimateur $\hat{E}_a$ est non biaisé, peu importe si le modèle utilisé est valide ou non. En fait, la procédure est conçue de telle manière qu'au lieu d'estimer directement Y_a, on estime E_a et on ajoute à cette valeur une constante connue $\tilde{Y}_a$. Pour que la procédure soit efficace, la variance des e_i pour le domaine doit être inférieure à celle des Y_i. Il n'est pas nécessaire que $E_a = 0$. Le raisonnement est le même dans le cas plus habituel où les valeurs de B sont estimées au moyen de l'échantillon. Dans ces circonstances, l'estimation de Y_a est convergente selon le plan, peu importe le modèle utilisé (Särndal 1984). En outre, il est possible d'estimer les valeurs de B à l'aide des données de l'échantillon relatives au domaine étudié seulement, ce qui donne ce que SGM appellent un estimateur direct, ou à l'aide des données de l'échantillon global, ce qui donne un estimateur direct modifié. Un facteur déterminant dans le choix entre l'estimateur direct et l'estimateur direct modifié dans ces circonstances est de savoir si les valeurs de B pour l'échantillon global valent aussi pour le domaine. Si ce n'est pas le cas, il faut ajouter dans le modèle pour l'échantillon global des termes d'interaction entre X et les indicateurs de domaine. Avec un ensemble complet de termes d'interaction, l'estimateur direct modifié se ramène en fait à l'estimateur direct ordinaire.

On doit recourir à une approche dépendante d'un modèle lorsque les estimations fondées sur un plan ne sont pas assez précises, malgré qu'on ait utilisé le plus efficacement possible l'information supplémentaire qui était disponible. Il arrive en effet qu'on ne soit pas capable de calculer une estimation directe à cause de l'absence d'éléments de l'échantillon dans la petite région. Dans ces circonstances, on doit recourir à un modèle statistique afin d'"emprunter de l'information" à d'autres régions. Ces modèles reposent sur des hypothèses (ex.: $E_a = 0$ dans l'exemple ci-dessus) et la qualité des estimations obtenues à l'aide de ces modèles dépend de la validité des hypothèses. Comme une hypothèse est toujours un peu inexacte, les estimations régionales sont entachées d'un biais, et comme il s'agit d'estimations indirectes biaisées, on se sert largement de l'erreur quadratique moyenne (EQM) de plan pour en mesurer la qualité (notons que EQM $= V' + B^2$ où V' est la variance et B est le biais de l'estimation).

La manière la plus courante de comparer la qualité d'une estimation directe avec celle d'une estimation indirecte est de comparer la variance (V) de la première avec l'EQM de la seconde. Or, en lisant l'article de SGM, j'en suis venu à me demander si l'EQM était la bonne mesure pour évaluer la qualité d'un estimateur indirect. Dans la pratique, il est possible d'estimer la variance V d'un estimateur direct, ce qui n'est pas le cas pour l'EQM de plan d'un estimateur indirect. Compte tenu de cette observation, on préférera nettement l'estimateur direct si V = EQM. De fait, on optera le plus souvent pour l'estimateur direct si celui-ci offre un degré de précision acceptable, peu importe le rapport probable des valeurs V et EQM. Dans d'autres cas, si B est le biais prévu, on pourra préférer l'estimateur direct à l'estimateur indirect, à moins que $V > V' + kB^2$, où k est un facteur plus grand que 1 qui tient compte de la possibilité que le biais inconnu soit plus élevé que prévu.

On peut appliquer le même raisonnement aux estimateurs combinés (ou composites), qui sont la moyenne pondérée d'un estimateur direct et d'un estimateur indirect. On considère souvent que la meilleure manière de déterminer les poids est de minimiser l'erreur quadratique moyenne de l'estimateur combiné, ce qui donne des poids inversement proportionnels à V et à EQM, selon qu'il s'agit de

l'estimateur direct ou de l'estimateur indirect. Or, d'après le raisonnement ci-dessus, on pourrait tout aussi bien minimiser le poids de l'estimateur indirect, à la condition que l'estimateur combiné soit suffisamment précis. On pourrait aussi déterminer les poids selon une valeur probable maximum de l'EQM, plutôt que selon l'EQM prévue, afin de réduire la probabilité d'un biais important dans l'estimateur combiné.

Je ne comprends pas très bien l'explication des estimateurs tributaires de la taille d'échantillon définis par les équations (6.11) et (6.12) de SGM mais suivant certaines hypothèses, ces estimateurs pourraient cadrer dans l'argumentation présentée ci-dessus. Étant donné un plan d'échantillonnage avec probabilités égales et une valeur $\delta = 1$, ces estimateurs se ramènent à l'estimateur direct lorsque la taille effective de l'échantillon est égale ou supérieure à la taille prévue. Si on suppose que la taille prévue de l'échantillon implique un degré de précision acceptable pour une petite région, l'argumentation ci-dessus est respectée. Si la taille effective de l'échantillon est plus petite que prévu, l'estimateur tributaire de la taille d'échantillon devient la moyenne pondérée d'un estimateur direct et d'un estimateur indirect. Si on suppose que la taille prévue de l'échantillon représente l'effectif minimum nécessaire pour obtenir le degré de précision voulu, l'argumentation ci-dessus est de nouveau respectée. Si les observations précédentes constituent le fondement des estimateurs tributaires de la taille d'échantillon, il semblerait utile d'étendre ces estimateurs aux situations où la taille prévue de l'échantillon ne correspond pas à celle permettant tout juste d'atteindre le degré de précision voulue.

Comme je l'ai mentionné plus haut, l'information supplémentaire compte pour beaucoup dans la production d'estimations régionales précises. Cette information peut servir à accroître la précision des estimations fondées sur un plan ou elle peut être utilisée dans les modèles auxquels on recourt par l'approche dépendante de modèles. Idéalement, il faut de l'information supplémentaire qui soit étroitement liée aux variables étudiées dans l'enquête. En extrayant régulièrement de sources administratives ou autres des données auxiliaires récentes concernant les petites régions, on peut contribuer à améliorer les données régionales.

Bien que les auteurs évoquent la question plus générale des petits domaines, ils concentrent leur attention sur les petites régions. D'ailleurs, la question des petites régions est celle qui est le plus souvent traitée dans la littérature statistique et qui est en rapport avec l'application de méthodes d'estimation indirecte. Cela peut s'expliquer d'une part, par le fait que les domaines socio-économiques et les autres petits domaines pertinents (ex.: groupes d'âge-sexe) sont habituellement assez peu nombreux comparativement aux petites régions de sorte qu'on peut concevoir pour ces domaines un plan de sondage qui produise des estimations d'une précision acceptable pour chacun d'eux et d'autre part, par le fait que les domaines socio-économiques et démographiques sont souvent définis en fonction de la possibilité d'obtenir des estimations de plan d'une précision acceptable pour ces domaines (par ex., utiliser des groupes d'âge plus étendus pour certains domaines); toutefois, en ce qui concerne les domaines spatiaux, les territoires sont définis à l'avance et aucun regroupement n'est possible. Il en est ainsi d'une part, parce que les données auxiliaires sont en nombre insuffisant pour être utilisées dans les modèles statistiques s'appliquant aux domaines de ce genre et d'autre part, parce que l'analyse des domaines socio-économiques a souvent pour but d'établir des comparaisons entre les domaines. Or, ces comparaisons sont faussées si l'estimation relative à un domaine est "renforcée" par de l'information provenant d'autres domaines (voir, par exemple, Schaible 1992). Cela nous amène à dire qu'on ne devrait pas utiliser les estimations indirectes sans faire preuve d'esprit critique.

En conclusion, je suis d'accord avec l'approche générale présentée dans cet article. Lorsque c'est possible, les échantillons devraient être conçus de manière à produire des estimations directes pour petite région d'une précision acceptable, et les plans de sondage devraient être élaborés en ce sens. De même, lorsque c'est possible, on devrait se servir de données auxiliaires pour accroître la précision des estimations directes pour petite région. Lorsqu'il faut recourir à des estimations indirectes, on doit procéder avec prudence. Il faut élaborer soigneusement les modèles, chercher des estimateurs qui résistent bien aux défaillances d'un modèle et faire des analyses pour évaluer la pertinence des estimations indirectes. En l'absence de bons indices pour mesurer la qualité des estimations indirectes, il est essentiel de distinguer clairement ces estimations des estimations fondées sur un plan. Comme les estimations indirectes ne conviennent pas à tous les usages, l'utilisateur doit s'assurer que les estimations indirectes dont il veut se servir répondront à ses besoins.

BIBLIOGRAPHIE

SÄRNDAL, C.-E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association*, 79, 624-631.

SCHAIBLE, W.L. (1992). Use of small area estimators in U.S. Federal programs. *Proceedings of the International Scientific Conference on Small Area Statistics and Survey Design*, (Vol. 1), 95-114, Central Statistical Office of Poland, Warsaw.

RÉPONSE DES AUTEURS

Nous remercions Wayne Fuller et Graham Kalton pour leurs commentaires stimulants qui complètent très bien les idées que nous avons exprimées dans l'article. Plusieurs de ces commentaires apportent des éclaircissements sur certains points et renforcent les arguments exposés. Encouragés par cet appui, nous aimerions approfondir certains points touchant le plan de sondage et, en même temps, répondre aux principaux commentaires des critiques.

Il est clair que le concepteur de plans de sondage s'efforce, malgré des contraintes opérationnelles, d'optimiser le plan dans le but de respecter les objectifs d'une enquête. Les grandes enquêtes poursuivent habituellement plusieurs objectifs, et il est fort probable que le concepteur a peu de prise sur l'établissement d'un ordre de priorité pour ces objectifs. Néanmoins, c'est justement à l'étape de l'établissement d'un ordre de priorité qu'il faut faire ressortir avec vigueur les besoins en données régionales, particulièrement en ce qui concerne les grandes enquêtes permanentes.

Dans les années soixante et soixante-dix, la plupart des pays ont mis l'accent sur les estimations infranationales (État, province) et modifié dans une certaine mesure les plans de sondage antérieurs qui, eux, visaient à optimiser les estimations nationales. Par exemple, on a utilisé des fractions de sondage différentes pour être sûr que les petits États ou les petites provinces contiendraient au moins tel nombre d'unités de l'échantillon. Vu la demande croissante de données au niveau infraprovincial ou infra-État (comté, district et municipalité), il faut modifier davantage les plans antérieurs, qui, eux, prévoient une répartition optimale de l'échantillon au niveau national. Cette modification peut consister à utiliser des fractions de sondage différentes pour les divisions administratives d'un État ou d'une province. Par exemple, si l'on souhaite produire des estimations infraprovinciales de qualité comparable, on attribuera sans doute à chaque province un échantillon qui sera à peu près proportionnel au nombre de régions infraprovinciales que contient la province. Cela n'équivaut peut-être pas à la taille relative de la population des provinces. Comme nous l'avons mentionné dans la section 5.4 de notre article, il faudrait privilégier une approche ascendante pour la répartition de l'échantillon. Les pertes enregistrées aux niveaux d'agrégation supérieurs, comme les gains aux niveaux inférieurs, varieraient selon les enquêtes, mais il est probable que, dans beaucoup de cas, une légère augmentation du coefficient de variation au niveau national se traduise par des gains appréciables au niveau des petites régions.

Kalton souligne l'importance de limiter la grosseur des grappes pour l'estimation de la variance; il est avantageux d'accroître le nombre de degrés de liberté en formant un grand nombre de petites grappes au lieu d'un petit nombre de grosses grappes. Précisons que l'agglomération d'unités d'échantillonnage présente un autre inconvénient du point de vue de l'estimation, en particulier de l'estimation pour petites régions; en effet, un plan qui prévoit la formation de grosses grappes produira des effets de plan prononcés, même pour de petits domaines planifiés. En règle générale, on recourt au plan avec formation de grappes pour réduire les coûts de l'enquête. Compte tenu de l'évolution actuelle des méthodes de collecte de données – par exemple l'usage moins fréquent de l'interview sur place au profit de l'interview assistée par ordinateur – il est devenu nécessaire de revoir périodiquement les modèles coût-variance qui conditionnent les décisions concernant la formation de grappes.

Une autre question que nous n'avons pas traitée dans l'article est celle de l'effet du renouvellement de l'échantillon dans les enquêtes permanentes. Pour une période donnée, il se peut que certains petits domaines n'aient pas un effectif suffisant pour qu'on puisse établir des estimations fiables. Cependant, à mesure que les unités de l'échantillon se succèdent par renouvellement, la taille cumulée ou effective de l'échantillon dans les domaines augmente et on peut ainsi calculer, par domaine, des estimations fiables, quoique biaisées chronologiquement. Par un choix judicieux de plans de renouvellement, les concepteurs d'enquêtes peuvent maximiser la taille cumulée de l'échantillon dans une période donnée. Par exemple, pour des estimations trimestrielles d'une enquête mensuelle, le plan de renouvellement optimal est $[1(2)]^k$, c'est-à-dire la répétition de k fois la séquence "un mois dans l'échantillon, deux mois hors de l'échantillon". Ce raisonnement s'apparente à celui de Leslie Kish sur le cumul d'échantillons dans le temps.

Kalton nuance la mise en garde faite sur l'emploi des estimateurs indirects en proposant une erreur quadratique moyenne pondérée, par laquelle on associe un poids plus grand que 1 au terme du biais pour tenir compte de ce que le biais de l'estimateur indirect peut être plus élevé que prévu. Deux raisons font que le biais peut être plus élevé que le laisse croire le modèle pour les effets de petite région: la variation aléatoire à l'intérieur du modèle et la défaillance du modèle. Il est utile de se rappeler ici la suggestion de Fay et Herriot (1979), qui proposent qu'il y ait un écart de moins d'une erreur type entre une estimation combinée et une estimation de plan; cette mesure permet de concevoir la possibilité que l'estimateur de modèle renferme un biais élevé quelqu'en soit la raison. Kalton affirme aussi, comme nous, qu'étant donné un estimateur direct de qualité acceptable, on pourra choisir en pratique d'utiliser cet estimateur même si l'erreur quadratique moyenne estimée correspondante est plus grande que celle d'estimateurs fondés sur un modèle. Compte tenu de la possibilité toujours réelle d'une défaillance du modèle, on favorisera cette approche du "mieux vaut prévenir que guérir", du moins tant qu'on n'aura pas expérimenté certains estimateurs indirects dans des conditions particulières. Et cela n'empêche pas qu'il y ait des situations où il faut faire preuve d'un peu d'audace.

Dans ses observations sur l'estimateur tributaire de la taille de l'échantillon, Kalton laisse entendre que l'attribution d'un poids nul à la composante synthétique peut poser un problème si la taille réelle de l'échantillon pour un petit domaine excède la taille prévue, car celle-ci peut n'être pas assez élevée pour qu'on puisse produire des estimations directes acceptables. Une solution possible serait d'utiliser

une valeur n_{min}, soit l'effectif qui permet de produire des estimations directes tout juste acceptables. Notons toutefois que n_{min}, tel qu'il est défini ici, est dépendant des caractéristiques.

Dans ses commentaires, Fuller décrit sommairement une méthode d'estimation pour petites régions qui exploite un modèle fondé sur les composantes de la variance et qui, pourtant, utilise des poids fixes pour garantir la cohérence interne des estimations relatives à diverses caractéristiques. Outre la cohérence interne des estimations régionales, il faut parfois que la somme des totaux estimés des petites régions qui forment une grande région corresponde à l'estimation directe officielle pour la grande région. Une façon d'assurer cette concordance est d'ajuster les estimations régionales à l'estimation directe pour la grande région au moyen, par exemple, d'un simple coefficient de correction; cependant, si les coefficients de correction dépendent des caractéristiques, la cohérence interne mentionnée précédemment n'existera plus. On peut garantir simultanément les deux types de cohérence à la condition que l'estimateur direct pour la grande région soit un estimateur par régression généralisé, $\hat{Y}_e + (X - \hat{X}_e)\hat{\beta}$, et qu'on utilise l'estimateur direct modifié $\hat{Y}_{sreg,a} = \hat{Y}_{e,a} + (X_a - \hat{X}_{e,a})\hat{\beta}$ (section 6.1 de notre article) pour les petites régions.

Comme le mentionne Fuller, il est possible d'estimer la valeur moyenne du biais au carré d'un estimateur pour un sous-ensemble quelconque de petites régions. Nous aimerions préciser de nouveau ici que la valeur moyenne du biais pour un ensemble de petites régions n'a pas de rapport direct avec une petite région en particulier. C'est pourquoi nous préférons utiliser, quand c'est possible, des estimateurs qui sont approximativement non biaisés selon le plan. Si nous n'avons d'autre choix que d'utiliser un estimateur fondé sur un modèle, nous devons nous efforcer de trouver des covariables pertinentes pour lesquelles il existe de l'information supplémentaire fiable afin de réduire au minimum le biais résiduel de l'estimateur.

Peut-être à cause des délais qu'implique inévitablement la production des données du recensement, aucun des deux critiques ne considère le recensement comme une source de données pour de petits domaines. À cet égard, il convient de mentionner qu'on envisage actuellement une formule qui prévoirait une sorte de grande enquête postcensitaire continue en remplacement du questionnaire détaillé utilisé dans le recensement décennal. Cette formule, appelée méthode des échantillons avec renouvellement complet, est décrite dans Kish (1990); une formule semblable, appelée méthode de dénombrement continu, est décrite dans Alexander (1994). Cette méthode comporte un certain nombre de fonctions qui méritent d'être examinées dans la mesure où elles peuvent devenir autant de moyens de produire à meilleur marché et dans de courts délais des données pour petits domaines.

Enfin, nous tenons à préciser qu'en soulignant l'importance de considérer l'estimation pour domaine à l'étape de la conception du plan de sondage, surtout en ce qui a trait aux domaines de taille moyenne, nous ne voulons aucunement minimiser le rôle des modèles dans l'estimation qui s'applique à de très petits domaines.

Nous espérons que l'approche générale présentée dans l'article, de même que les commentaires perspicaces des critiques, en particulier les conseils et les mises en garde formulés par Kalton dans le dernier paragraphe de ses commentaires, permettront aux concepteurs de plans de sondage et aux chercheurs de trouver les solutions qui conviennent pour les problèmes qui les intéressent.

SOURCES ADDITIONNELLES

ALEXANDER, C.H. (1994). A prototype continuous measurement system for the U.S. Census of Population and Housing. Document pour présentation à la réunion annuelle de la Population Association of America, Miami, Florida, 5 mai, 1994.

KISH, L. (1990). Recensement par étapes et échantillons avec renouvellement complet. *Techniques d'enquête*, 16, 67-77.

Techniques d'enquête, juin 1994
Vol. 20, n° 1, pp. 25–33
Statistique Canada

Estimation pour petits domaines dans des plans de sondage avec probabilités inégales

D. HOLT et D.J. HOLMES[1]

RÉSUMÉ

L'estimation de totaux et de moyennes pour un domaine à l'aide de données d'échantillon est un problème courant. Lorsque le domaine est grand, l'échantillon observé est généralement suffisamment grand pour que des estimateurs directs, fondés sur un plan, soient assez précis. Mais lorsque le domaine est petit, la taille de l'échantillon observé est faible et les estimateurs directs ne conviennent plus. L'estimation pour petite région est un cas typique et des solutions de rechange comme l'estimation synthétique et l'estimation fondée sur un modèle ont été élaborées pour contourner le problème. Ces deux méthodes ont ceci de caractéristique, que de l'information est "empruntée" à d'autres petits domaines (ou petites régions) de manière à obtenir des estimateurs de paramètres plus précis, lesquels sont ensuite combinés avec de l'information supplémentaire, comme des moyennes ou des totaux pour population, tirée de chaque petite région afin d'établir une estimation plus précise de la moyenne ou du total pour un domaine (ou une région). Dans cet article, nous étudions un cas d'échantillonnage avec probabilités inégales où il n'existe pas d'information supplémentaire et où l'emprunt d'information n'est pas permis; pourtant, des estimateurs simples fondés sur un modèle sont élaborés dans ces conditions et semblent offrir des gains d'efficacité appréciables. L'exemple que nous utilisons ici est une étude de marché mais les domaines d'application de ces méthodes sont nombreux.

MOTS CLÉS: Estimation synthétique; estimation fondée sur un plan; estimation pour petites régions; estimation fondée sur un modèle; parts de marché.

1. INTRODUCTION

Dans cet article, nous nous intéressons à l'estimation de totaux et de moyennes pour un domaine à partir d'un échantillon réparti de façon non proportionnelle. Certains domaines peuvent être grands, auquel cas la taille effective de l'échantillon peut aussi être élevée et des estimateurs fondés sur un plan (ou estimateurs directs) conviendront. D'autres domaines peuvent être petits, auquel cas la taille effective de l'échantillon sera également faible et les estimateurs fondés sur un plan (ou estimateurs directs) seront trop imprécis pour être utilisés convenablement. Nous allons illustrer les méthodes proposées à l'aide d'un exemple où sont estimées des ventes, des parts de marché et des taux de pénétration pour des produits dans une étude de marché. Les domaines sont constitués de fabricants d'automobiles ou de modèles de voiture. Cependant, la méthode générale peut s'appliquer à d'autres échantillons d'entreprises ou d'institutions répartis de façon non proportionnelle.

Le problème étudié ressemble au problème de l'utilisation de l'estimation synthétique pour de petites régions (Gonzales 1973; Gonzales et Hoza 1978; Platek et coll. 1987). L'estimation synthétique dépend normalement de deux facteurs: i) l'emploi combiné de variables auxiliaires et de moyennes ou de totaux de population pour chaque petite région (ou petit domaine) en vue d'améliorer les estimations par la stratification a posteriori ou l'estimation par régression, et ii) l'amélioration des estimations par le regroupement de données de plusieurs petites régions (ou petits domaines). Dans le cas qui nous occupe, il n'existe pas de moyennes ou de totaux de population et comme l'objectif essentiel est de comparer des domaines (c.-à-d. fabricants et produits particuliers), l'emprunt d'information de l'un à l'autre est inconcevable. Nous proposons une catégorie d'estimateurs synthétiques qui n'utilisent aucune de ces deux méthodes et qui, pourtant, sont préférés aux estimateurs directs. Les estimateurs proposés ont une structure simple et une interprétation intéressante et peuvent être justifiés suivant un ensemble d'hypothèses de modèle qui peuvent être testées selon l'hypothèse générale des plan de sondage non informatifs.

2. EXEMPLE: ÉTUDE DE MARCHÉ

Les responsables d'études de marché estiment souvent le volume total des ventes et la part de marché pour chaque fabricant d'un produit. Nous examinons le cas des automobiles achetées au cours d'une année en vue d'être utilisées comme voitures d'entreprise. Il faut des estimations de totaux et de parts de marché pour chaque fabricant d'automobiles et pour les modèles qui constituent le plus souvent les parcs de véhicules.

Les termes "parc" et "entreprise" ont chacun une définition assez large. Une voiture d'entreprise est définie comme une automobile qui est achetée à des fins commerciales plutôt que pour un usage privé et qui est associée à une entreprise au sens le plus large. Cette définition

[1] D. Holt et D.J. Holmes, Department of Social Statistics, University of Southampton, Highfield, Southampton, UK, SO95NH.

comprend les voitures achetées à l'intention des représentants, lesquelles sont en grand nombre. Elle comprend aussi les voitures de luxe achetées à l'intention des dirigeants et des autres membres de la haute direction de grandes entreprises, de même que les voitures achetées par de petites sociétés, comme des groupes de médecins, ou des travailleurs indépendants, comme des propriétaires d'usine. La population d'entreprises acheteuses - appelées consommateurs - comprend donc un grand nombre de petites entreprises qui n'achètent qu'une ou deux automobiles à quelques années d'intervalle.

Nous définissons Y_{ki} comme le nombre d'automobiles de type k achetées par le consommateur i dans la période de référence d'un an. Le type de produit k (c.-à-d. le domaine) peut désigner un modèle particulier ou tous les modèles d'un fabricant. Par conséquent, $Y_k = \sum_i Y_{ki}$ est le nombre total d'automobiles de type k achetées par l'ensemble des consommateurs. Soit Z_i le nombre total d'automobiles de toutes sortes achetées par le consommateur i et $Z = \sum_i Z_i$, le nombre total d'automobiles vendues. La part de marché pour le produit de type k est définie par l'équation $R_k = Y_k/Z$.

Nous posons aussi les relations suivantes

$$Y'_{ki} = 1 \quad \text{si} \quad Y_{ki} > 0$$
$$= 0 \quad \text{si} \quad Y_{ki} = 0$$

et

$$Z'_i = 1 \quad \text{si} \quad Z_i > 0$$
$$= 0 \quad \text{si} \quad Z_i = 0.$$

Ainsi, Y'_{ki} et Z'_i sont des variables indicatrices pour les consommateurs qui achètent respectivement des produits de type k et au moins une automobile de n'importe quelle sorte dans la période de référence. Le nombre de consommateurs qui achètent des produits de type k est donc défini par l'équation $Y'_k = \sum_i Y'_{ki}$ et le nombre total de consommateurs qui achètent au moins une voiture de n'importe quel type est défini par l'équation $Z' = \sum_i Z'_i$. Le taux de pénétration du marché pour le produit de type k - c'est-à-dire la proportion de consommateurs qui, parmi ceux qui achètent une voiture de n'importe quel type dans la période de référence, achètent une voiture de type k - est défini par l'équation $R'_k = Y'_k/Z'$.

Les quatre paramètres Y_k, R_k, Y'_k et R'_k peuvent tous faire l'objet d'inférence dans une étude de marché et sont définis comme des paramètres de population finie, notamment des totaux de domaine ou des rapports de totaux de domaine.

3. PLAN DE SONDAGE ET ESTIMATEURS DIRECTS

La plan de sondage repose sur deux bases s'excluant mutuellement et peut être considéré comme un plan d'échantillonnage simple stratifié comportant dix strates. La première base de sondage consiste dans un registre (Dun et Bradstreet) de 35,000 compagnies réparties dans huit strates selon le nombre d'employés et la nature des activités de l'entreprise (fabrication ou distribution). La deuxième base de sondage consiste dans un vaste registre de 1.4 million d'abonnés au service commercial de British Telecom, répartis dans deux strates: lignes privées et lignes commerciales. Notons qu'il s'agit d'abonnés au service commercial dans les deux cas; une ligne commerciale est attribuée si l'abonné dispose de plusieurs locaux.

À l'aide de données d'enquête antérieures, on a réparti l'échantillon de façon optimale au moyen de la formule de Neyman afin de réduire au maximum la variance de l'estimateur du nombre total d'automobiles achetées (Z). On a recueilli des données sur les achats d'automobiles immédiatement après la fin de l'année de référence. Les tailles de strate $\{N_h\}$ et les tailles d'échantillon $\{n_h\}$ pour les strates $h = 1, \ldots, 10$ figurent dans le tableau 1.

Tableau 1

Base de sondage: effectif et poids par strate

Strate (h)	Taille de la strate N_h	Effectif n_h	Poids $\pi_h^{-1} = N_h/n_h$
British Telecom:			
Ligne privée	389,445	1,150	338.65
Ligne commerciale	1,007,399	7,406	136.02
Dun et Bradstreet:			
Fabrication			
50-99 employés	6,646	235	28.28
100-499	6,826	1,113	6.13
500-999	992	520	1.91
1,000+	1,110	849	1.31
Distribution			
50-99 employés	8,703	472	18.44
100-499	7,625	1,437	5.31
500-999	1,133	484	2.34
1,000+	1,523	1,117	1.36
Total	1,431,402	14,783	96.83

L'échantillon est un échantillon simple stratifié réparti de façon non proportionnelle et les estimateurs directs de même que les variances correspondantes sont connus avec exactitude. La stratification donne des poids d'échantillonnage très différents (de 1.31 à 338.65); elle est utile mais loin d'être idéale. Beaucoup de consommateurs n'achètent aucune automobile dans l'année de référence, de sorte que chaque strate renferme un mélange de valeurs nulles et de valeurs non nulles. La proportion de valeurs nulles dans chaque strate est évidemment élevée pour un produit de type k quelconque.

Le tableau 2 contient les estimations directes d'enquête, les erreurs types estimées (voir Holt et Holmes (1993) pour la façon dont elles ont été calculées) et les coefficients de variation pour certains produits de fabricants d'automobiles. Les produits A et B représentent tous les modèles de deux grands fabricants. Le produit C représente un seul

Tableau 2

Estimations directes d'enquête, erreurs types et coefficients de variation pour certains produits

Produit (k)	Estimation du nombre de consommateurs		Estimation du nombre d'automobiles	
	Total $\hat{Y}'_k$	Taux de pénétration $\hat{R}'_k$	Total $\hat{Y}_k$	Part $\hat{R}_k$
A	59,890	.3843	270,051	.3781
	(2,651)	(.0144)	(35,704)	(.0315)
	(.044)	(.037)	(.132)	(.083)
B	34,282	.2200	153,518	.2149
	(1,960)	(.0117)	(8,653)	(.0131)
	(.057)	(.053)	(.056)	(.061)
C	23,363	.1499	81,381	.1139
	(1,602)	(.0098)	(17,559)	(.0194)
	(.069)	(.065)	(.216)	(.170)
D	13,857	.0889	25,312	.0354
	(1,311)	(.0081)	(2,906)	(.0039)
	(.095)	(.091)	(.115)	(.110)
E	9,025	.0579	24,370	.0341
	(1,146)	(.0072)	(7,336)	(.0101)
	(.127)	(.124)	(.301)	(.296)
F	5,125	.0329	13,724	.0192
	(676)	(.0043)	(2,369)	(.0030)
	(.132)	(.131)	(.173)	(.156)
G	7,518	.0482	11,031	.0154
	(1,015)	(.0064)	(1,456)	(.0022)
	(.135)	(.133)	(.132)	(.143)

1ière ligne: estimation 2ième ligne: e.t. 3ième ligne: c.v.

modèle, produit par le fabricant A, qui occupe une part appréciable du marché des parcs de voitures. Les autres produits du tableau 2 occupent des petites parts de marché. Les produits F et G sont plutôt destinés au marché des voitures de fonction. La liste du tableau 2 est incomplète, ce qui explique que la somme des parts de marché n'est pas égale à un. Il convient aussi de souligner que les catégories de produits ne s'excluent pas mutuellement. De façon générale, on juge que l'enquête donne des résultats satisfaisants mais on s'aperçoit que les estimations relatives aux fabricants ou aux modèles qui ont de faibles parts de marché sont instables sur plusieurs années. Cela peut se remarquer par le coefficient de variation, qui est supérieur à 0.1 pour des produits qui occupent une part de marché modeste et qui peut dépasser 0.15 ou 0.2 dans certains cas. Cette instabilité se répercute sur les estimations de la variance de même que les estimations des ventes totales ou des parts de marché pour les produits.

4. MÉTHODE FONDÉE SUR UN MODÈLE

Étant donné le plan de sondage, on ne peut s'attendre à améliorer l'efficacité des estimateurs directs dans le cadre classique des enquêtes par sondage. La façon habituelle de procéder est d'utiliser de l'information supplémentaire pour la stratification a posteriori, l'estimation par quotient ou l'estimation par régression mais dans tous ces cas, il faut connaître les moyennes ou les totaux de population. Or, cette information n'existe pas. C'est pourquoi nous recourons plutôt à une méthode fondée sur un modèle pour obtenir d'autres estimateurs pour toute la gamme de produits.

4.1 Estimation de Y'_k: le nombre de consommateurs qui achètent un produit de type k

Nous considérons en premier lieu le nombre de consommateurs qui achètent un produit de type k. Nous donnons à Y'_{ki} une nouvelle extension, Y'_{khi}, dans le but évident de définir la variable aléatoire indicatrice d'achat pour le produit k et le consommateur i dans la strate h. Nous considérons chaque décision de consommateur comme le résultat d'un tirage bernoullien. Soit $P_{k|h}$ la probabilité qu'un consommateur dans la strate h achète une automobile de type $k [P_{k|h} = \text{Prob}(Y'_{khi} = 1)]$. Nous définissons l'équivalent de Y'_k pour un modèle, c.-à-d. le nombre total de consommateurs du produit de type k, par la formule

$$\Theta'_k = \sum_h N_h P_{k|h}. \qquad (1)$$

En supposant que chaque décision de consommateur est indépendante, on peut exprimer la vraisemblance comme le produit habituel de termes binomiaux. Les estimateurs du maximum de vraisemblance sont définis par l'équation $\hat{P}_{k|h} = n_{kh}/n_h$, et l'estimateur du maximum de vraisemblance de Θ'_k est l'estimateur de plan de sondage stratifié bien connu

$$\hat{\Theta}'_k(1) = \sum_h \frac{N_h}{n_h} n_{kh} = \sum_h N_h \bar{y}'_{kh}, \qquad (2)$$

où n_{kh} est le nombre de consommateurs échantillonnés de la strate h qui achètent un produit de type k, n_h, est l'effectif de la strate pour l'échantillon et $\bar{y}'_{kh} = n_{kh}/n_h$ est la moyenne d'échantillon pour les consommateurs de la strate h (c.-à-d. la proportion empirique de consommateurs de la strate h qui achètent un produit de type k). L'estimateur ci-dessus est peu satisfaisant en règle générale lorsqu'un trop petit nombre de consommateurs achètent le produit de type k.

Supposons que nous introduisions une autre variable de condition, par laquelle chaque consommateur peut être classé dans une catégorie f, $f = 1, \ldots, F$, et que nous donnions à la variable aléatoire indicatrice une nouvelle extension, Y'_{khfi}. Les catégories f recouperont les strates h et il s'agit de définir f de telle sorte que, pour n'importe quelle catégorie donnée, la décision d'un consommateur d'acheter le produit de type k ne dépende pas de la strate à laquelle appartient ce consommateur. En ce qui concerne les automobiles destinées à des parcs de véhicules, nous définissons un classement basé sur le nombre total d'automobiles que possède et utilise chaque consommateur (c.-à-d. la taille du parc d'automobiles). Dans la section 5, nous examinons plus en détail le choix de f.

Si N_{hf} – le nombre total de consommateurs inclus dans la strate h et la catégorie f (taille du parc d'automobiles) – est connu, on peut élargir l'équation (1) de la manière habituelle et exprimer le paramètre étudié sous cette nouvelle forme:

$$\Theta'_k = \sum_h \sum_f N_{hf} P_{k|hf}. \tag{3}$$

L'équation ci-dessus équivaut à la stratification a posteriori lorsque $\{N_{hf}\}$ est connu et si tel est le cas, l'information supplémentaire entraînera un gain d'efficacité (Holt et Smith 1979). Si $\{N_{hf}\}$ est inconnu, nous pouvons reformuler le modèle selon deux ensembles de probabilités:

$Q_{f|h}$ = Prob {consommateur possède un parc de taille f | strate h},

$P_{k|hf}$ = Prob {consommateur achète un produit de type k | strate h et taille f du parc de véhicules}.

On peut alors exprimer le paramètre étudié par l'équation

$$\Theta'_k = \sum_h \sum_f N_h Q_{f|h} P_{k|hf}. \tag{4}$$

Nous pouvons obtenir un autre estimateur fondé sur un modèle en faisant d'autres hypothèses sur les paramètres de modèle. Supposons maintenant que

$$P_{k|hf} = P_{k|f} \quad \text{pour tous} \quad h. \tag{5}$$

Cette équation implique qu'étant donné la catégorie f (la taille du parc d'automobiles qu'exploite un consommateur), la probabilité que le consommateur achète un produit de type k est *indépendante* de la strate (h) à laquelle appartient ce consommateur. Du point de vue algébrique, l'hypothèse est semblable à celle utilisée dans l'estimation synthétique pour petites régions sauf que dans ce dernier cas, on regroupe des données de plusieurs régions, ce qui est inacceptable dans le cas qui nous occupe. Nous allons plutôt regrouper des données de plusieurs strates dans le domaine étudié. Il s'agit en fait de choisir une variable de condition qui tienne compte de la relation marginale entre la décision d'un consommateur d'acheter un produit de type k et la strate à laquelle appartient ce consommateur.

En se servant de l'hypothèse (5) et en élargissant la notation de la façon habituelle ($n_{kf} = \sum_h n_{khf}$, *etc.*), on peut montrer que

$$\hat{Q}_{f|h} = \frac{n_{hf}}{n_h}, \qquad \hat{P}_{k|f} = \frac{n_{kf}}{n_f}$$

et l'estimateur du maximum de vraisemblance de Θ'_k devient

$$\hat{\Theta}'_k(2) = \sum_h \sum_f N_h \frac{n_{hf}}{n_h} \frac{n_{kf}}{n_f} = \sum_f \hat{N}_f \frac{n_{kf}}{n_f}$$

$$= \sum_h \hat{N}_f \bar{y}'_{kf}, \tag{6}$$

où $\hat{N}_f = \sum_h N_h n_{hf}/n_h$ et $\bar{y}'_{kf} = n_{kf}/n_f$ est la moyenne d'échantillon non pondérée pour les consommateurs de la catégorie f (c.-à-d. la proportion empirique de consommateurs de la catégorie f qui achètent un produit de type k).

Par conséquent, l'équation (6) a la forme d'un estimateur de stratification basé sur la classification f, sauf que l'effectif de population des strates, $\{N_f\}$, est inconnu. Notons qu'on aurait eu naturellement un estimateur du même type, à la différence que $\{N_f\}$ serait connu, si un échantillon stratifié basé sur f avait été prélevé. En fait, ce n'est pas le cas: les membres de la catégorie f qui sont échantillonnés ne le sont pas avec la même probabilité. Cependant, à cause des hypothèses paramétriques, on considère l'échantillon dans chaque catégorie f comme un échantillon à probabilités égales puisque suivant l'hypothèse (5), les poids de sondage sont non informatifs et amènent simplement une perte d'efficacité lorsqu'on estime $P_{k|f}$. Par conséquent, même si les fractions de sondage n_h/N_h servent à estimer $\{N_f\}$, elles ne sont pas utilisées explicitement dans $\hat{P}_{k|f} = n_{kf}/n_f = \bar{y}'_{kf}$. Notons que l'estimateur regroupe de l'information provenant des strates h à l'intérieur du domaine k mais non dans plusieurs domaines à la fois (c.-à-d. les produits).

Il convient de noter que si n_h/N_h est constant, l'équation (6) se ramène à l'estimateur par facteur d'extension habituel défini en (2) et l'hypothèse (5) ne donne pas de nouvel estimateur. Si l'échantillon est réparti de façon non proportionnelle, l'hypothèse amène l'utilisation des poids de sondage pour $\hat{N}_f$ (où ils sont nécessaires) mais non pour l'estimation de $P_{k|f}$ (où ils sont non informatifs étant donné f et l'hypothèse (5)).

L'équation (5) représente un ensemble d'hypothèses contraignantes, qui exigent que les valeurs $P_{k|hf}$ soient égales à une valeur unique $P_{k|f}$ pour tous h. En pratique, on peut introduire des hypothèses aléatoires telles que $P_{k|hf} = P_{k|f} + \epsilon_{k|hf}$, où $E[\epsilon_{k|hf}] = 0$ et $V[\epsilon_{k|hf}] = \sigma^2_\epsilon$. Ces hypothèses conduiront à une analyse hiérarchique de Bayes ou à une analyse empirique de Bayes, comme le décrivent Ghosh et Rao (1994) ou Fay et Herriot (1979). Nous n'examinons pas ici ces méthodes car cela évacuerait la forme élémentaire de l'estimateur fondé sur un modèle de même que l'éclairage qu'elle nous apporte. Dans le même esprit, on peut appliquer la méthode de Särndal et Hidiroglou (1989) ou celle de Drew, Singh et Choudhry (1982) pour obtenir des estimateurs dépendants de la taille d'échantillon sans contrevenir à la règle voulant que l'estimateur regroupe de l'information qui provient d'un seul domaine et non de plusieurs domaines à la fois (produits).

Nous pouvons comparer les estimateurs (2) et (6) lorsque l'hypothèse (5) se vérifie puisqu'on peut montrer que

$$V_\xi(\hat{\Theta}'_k(1)) = \sum_h \frac{N_h^2}{n_h} P_{k|h}(1 - P_{k|h})$$

$$= \sum_h \sum_f \frac{N_h^2}{n_h} Q_{f|h} P_{k|f}$$

$$- \sum_h \sum_f \sum_{f'} \frac{N_h^2}{n_h} Q_{f|h} Q_{f'|h} P_{k|f} P_{k|f'}, \quad (7)$$

où le symbole $V_\xi(\cdot)$ veut souligner le fait que la variance est évaluée par rapport à la distribution basée sur un modèle.

On peut aussi montrer que suivant l'hypothèse (5),

$$V_\xi(\hat{\Theta}'_k(2)) = \sum_h \sum_f \frac{N_h^2}{n_h} P_{k|f}^2 Q_{f|h} (1 - Q_{f|h})$$

$$- \sum_h \sum_{\substack{f \\ f \neq f'}} \sum_{f'} \frac{N_h^2}{n_h} P_{k|f} P_{k|f'} Q_{f|h} Q_{f'|h}$$

$$+ \sum_h \sum_f \frac{N_h^2}{n_h} \frac{P_{k|f}(1 - P_{k|f}) Q_{f|h}}{\sum_h n_h Q_{f|h}}$$

$$\left\{(1 - Q_{f|h}) + n_h Q_{f|h} + \frac{[1 + (2n_h - 3)Q_{f|h} - 2(n_h - 1)Q_{f|h}^2]}{\sum_h n_h Q_{f|h}}\right\} \quad (8)$$

et que $V_\xi(\hat{\Theta}'_k(1)) - V_\xi(\hat{\Theta}'_k(2)) \geq 0$.

Par conséquent, suivant les hypothèses de modèle additionnelles, $\hat{\Theta}'_k(2)$ a une variance plus petite que prévu. Les expressions ci-dessus représentent des variances fondées sur un modèle et n'impliquent pas de facteur de correction pour population finie, contrairement à ce qui se produirait si on utilisait une méthode prédictive pour les éléments non observés de chaque strate formée a posteriori.

L'estimateur du maximum de vraisemblance du taux de pénétration du marché pour le produit de type k, R'_k, suivant l'hypothèse (5), est défini simplement

$$\hat{\Omega}'_k(2) = \frac{\sum_f \hat{N}_f \frac{n_{kf}}{n_f}}{\sum_f \hat{N}_f \frac{n_{af}}{n_f}} = \frac{\sum_f \hat{N}_f \bar{y}'_{kf}}{\sum_f \hat{N}_f \bar{z}'_f}, \quad (9)$$

où n_{af} est le nombre de consommateurs échantillonnés de la catégorie f qui achètent une automobile d'un modèle quelconque et $\bar{z}'_f = n_{af}/n_f$ est la proportion empirique de consommateurs de la catégorie f qui achètent une automobile d'un modèle quelconque.

4.2 Efficacité de l'estimateur fondé sur un modèle de Y'_k

Afin d'analyser la différence d'efficacité entre $\hat{\Theta}'_k(2)$ et $\hat{\Theta}'_k(1)$, nous examinons l'efficacité de l'estimateur fondé sur un modèle, définie par l'équation

$$e[\hat{\Theta}'_k(2)] = \frac{V_\xi(\hat{\Theta}'_k(1)) - V_\xi(\hat{\Theta}'_k(2))}{V_\xi(\hat{\Theta}'_k(1))}, \quad (10)$$

pour diverses structures de population pour lesquelles l'hypothèse (5) se vérifie.

Nous considérons une population divisée en strates $\{h\}$, avec des tailles de strate $\{N_h\}$ et des tailles d'échantillon pour les strates $\{n_h\}$ comme dans le tableau 1, ainsi qu'une variable de condition (taille du parc d'automobiles) stratifiée en dix catégories f ($f = 1, \ldots, 10$) selon un ordre croissant. Nous calculons le facteur d'efficacité $e[\hat{\Theta}'_k(2)]$ pour diverses combinaisons de valeurs paramétriques de $\{Q_{f|h}\}$ et de $\{P_{k|f}\}$.

Nous considérons cinq structures différentes pour $\{Q_{f|h}\}$:

a) $Q_{f|h} = \begin{cases} 1 \text{ si } f = h \\ 0 \text{ si } f \neq h \end{cases}$ pour $h = 1, \ldots, 10$.

b) $Q_{f|h} = \begin{cases} 0.95 & f = h & \text{pour} & h = 1, \ldots, 10 \\ 0.025 & f = h - 1 & \text{pour} & h = 2, \ldots, 10 \\ 0.025 & f = h + 1 & \text{pour} & h = 1, \ldots, 9 \\ 0.05 & h = 1, f = 2 \text{ et} & & h = 10, f = 9 \\ 0 & \text{autrement} & & \end{cases}$

= matrice bande (0.025, 0.95, 0.025).

c) $Q_{f|h}$ = matrice bande (0.05, 0.90, 0.05).

d) $Q_{f|h}$ = matrice bande (0.05, 0.10, 0.70, 0.10, 0.05).

e) $Q_{f|h} = 0.1$ pour $h = 1, \ldots, 10$ et $f = 1, \ldots, 10$.

Nous considérons quatre structures différentes pour $\{P_{k|f}\}$:

i) $P_{k|f} = \begin{cases} 0.1 & f = 1, 2 \\ 0 & \text{autrement.} \end{cases}$

ii) $P_{k|f} = 0.1 - 0.01\,(f - 1)$ pour $f = 1, \ldots, 10$.

iii) $P_{k|f} = 0.1f$ pour $f = 1, \ldots, 10$.

iv) $P_{k|f} = 0.5$ pour $f = 1, \ldots, 10$.

Dans le cas de la structure (a), les catégories f coïncident avec les strates. En ce qui concerne les structures (b), (c) et (d), pour n'importe quelle strate h, la majorité des consommateurs se trouvent dans la même catégorie

($f = h$) et le reste dans les catégories voisines (pour (b) et (c) par exemple, $f = h - 1, h + 1$). Enfin, la structure (e) suppose que, pour n'importe quelle strate h, les consommateurs ont autant de chances d'appartenir à l'une ou l'autre des catégories $f = 1, \ldots, 10$.

En ce qui regarde $P_{k|f}$, la structure (i) suppose un modèle de voiture qui est acheté peu fréquemment par les consommateurs qui ont parc d'automobiles réduit (c.-à-d. ceux qui appartiennent à la catégorie 1 ou 2), et qui n'est pas du tout acheté par les consommateurs qui ont de grands parcs d'automobiles. La structure (ii) suppose un modèle de voiture qui est acheté selon une probabilité qui est inversement proportionnelle à la taille du parc d'automobiles, tandis que la structure (iii) suppose un modèle de voiture qui est acheté selon une probabilité directement proportionnelle à la taille du parc d'automobiles. En ce qui concerne la structure (iv), un modèle est acheté dans une probabilité de 0.5, peu importe la taille du parc d'automobiles du consommateur.

Le tableau 3 donne les valeurs du facteur d'efficacité exprimé par l'équation (10) pour chaque combinaison de structures de $Q_{f|h}$ et de $P_{k|f}$ selon la formule de répartition non proportionnelle définie dans le tableau 1. La colonne (a) du tableau correspond au cas dans lequel les catégories f coïncident avec les strates, et les deux estimateurs, $\hat{\Theta}_k'(1)$ et $\hat{\Theta}_k'(2)$, sont identiques. Le tableau montre qu'on peut réaliser de forts gains d'efficacité (par ex., 70%) pour certaines combinaisons de paramètres: moins il y a de rapport entre f et h, plus le gain d'efficacité est élevé. Même dans le cas des structures (c) et (d), où f et h sont étroitement liés, on peut observer des gains d'efficacité appréciables. Les gains d'efficacité dépendent beaucoup plus de la structure de $Q_{f|h}$ que de celle de $P_{k|f}$.

Tableau 3

Facteurs d'efficacité, $e[\hat{\Theta}_k'(2)]$, pour diverses combinaisons de structures de $Q_{f|h}$ et de $P_{k|f}$

		Structure de $Q_{f\|h}$				
		(a)	(b)	(c)	(d)	(e)
Structure de $P_{k\|f}$	(i)	0	0.108	0.196	0.355	0.648
	(ii)	0	0.116	0.206	0.391	0.695
	(iii)	0	0.103	0.181	0.387	0.695
	(iv)	0	0.115	0.203	0.391	0.706

Dans le cas de la structure (e), où $Q_{f|h}$ a la même valeur pour tous f et h, il est possible de montrer que le facteur d'efficacité peut être exprimé par l'équation

$$e[\hat{\Theta}_k'(2)] = \left(1 - \frac{\delta^2}{\bar{P}_{k|f}\,(1 - \bar{P}_{k|f})}\right) \frac{\sum_h \tau_h N_h^2/n_h}{\sum_h N_h^2/n_h}, \quad (11)$$

où

$$\bar{P}_{k|f} = \frac{1}{F}\sum_{f=1}^{F} P_{k|f} \quad \text{and} \quad \delta^2 = \frac{1}{F}\sum_{f=1}^{F} (P_{k|f} - \bar{P}_{k|f})^2$$

sont, respectivement, la moyenne et la variance de $\{P_{k|f}\}$ pour l'ensemble des catégories f et $\tau_h = 1 - n_h/n + O(n^{-1})$. Le terme entre parenthèses dans l'équation (11) peut prendre des valeurs de 0 à 1, selon la distribution des valeurs de $\{P_{k|f}\}$ pour les catégories f. Dans le cas (iv), $P_{k|f}$ est constant et le terme entre parenthèses est donc égal à un. Le second terme du membre de droite de l'équation (11) dépend uniquement du plan de sondage et sa valeur pour le mode de répartition de l'échantillon défini dans le tableau 1 est 0.706.

4.3 Estimation de Y_k: le nombre d'automobiles de type k achetées

On peut se servir de la méthode exposée dans la section 4.1 pour estimer le nombre d'automobiles achetées. Nous ajoutons une variable de condition pour représenter le nombre total d'automobiles achetées, m, tous modèles confondus, et nous donnons à la variable aléatoire une nouvelle extension, Y_{khfmi}, c'est-à-dire le nombre d'automobiles de type k achetées par le consommateur i de la strate h et de la catégorie f qui achète m voitures de toutes sortes. Ainsi, le nombre d'automobiles de type k achetées variera en fonction du nombre total d'automobiles achetées. Posons

$S_{m|hf}$ = Prob{consommateur achète m automobiles de toute sorte | h,f}, $m = 0, 1, 2, \ldots$,

$T_{\ell|hfm}$ = Prob{consommateur achète ℓ automobiles de type k | h,f,m}, $\ell = 0, 1, \ldots, m$.

On peut alors exprimer le paramètre à l'étude basé sur un modèle, qui est l'équivalent du nombre total de produits de type k achetés (Y_k), par l'équation

$$\Theta_k = \sum_h \sum_f \sum_m \sum_\ell N_h Q_{f|h} S_{m|hf} T_{\ell|hfm}\, \ell, \quad (12)$$

qui est une extension de l'équation (4).

Nous considérons deux ensembles d'hypothèses additionnels, dont le premier est

$$T_{\ell|hfm} = T_{\ell|fm} \quad \text{pour tous} \quad h. \quad (13)$$

Ces hypothèses signifient qu'étant donné la catégorie f et le nombre total d'automobiles neuves achetées, m, la distribution du nombre d'automobiles de type k achetées est indépendante de la strate (h).

L'estimateur du maximum de vraisemblance de Θ_k, suivant les hypothèses (13), est

$$\hat{\Theta}_k(2) = \sum_f \sum_m \hat{N}_{fm}\, \bar{y}_{kfm}, \quad (14)$$

où $\hat{N}_{fm} = \sum_h N_h n_{hfm}/n_h$, et $\bar{y}_{kfm} = \sum_\ell \ell\, n_{fm\ell}/n_{fm}$ est le nombre moyen non pondéré (pour l'échantillon) d'automobiles de type k achetées par des consommateurs de la catégorie f qui ont acheté au total m automobiles de toute sorte.

Les probabilités de sélection servent ici à définir un estimateur pondéré de N_{fm}, c'est-à-dire le nombre total de consommateurs de la catégorie f qui achètent m voitures de toute sorte. Cet estimateur a une forme semblable à celle de l'estimateur défini en (6). Suivant les hypothèses de modèle (13), on peut montrer que

$$V_\xi(\hat{\Theta}_k(2)) = \sum_h \sum_f \sum_m \frac{N_h^2}{n_h} \mu_{fm}^2 Q_{fm|h}(1 - Q_{fm|h}) - \sum_h \sum_f \sum_m \sum_{f'} \sum_{m'} \frac{N_h^2}{n_h} \mu_{fm}\mu_{f'm'} Q_{fm|h} Q_{f'm'|h} \quad (f,m) \neq (f',m') + \sum_h \sum_f \sum_m \frac{N_h^2}{n_h} \frac{\sigma_{fm}^2 Q_{fm|h}}{\sum_h n_h Q_{fm|h}} \left\{(1 - Q_{fm|h}) + n_h Q_{fm|h} + \frac{[1 + (2n_h - 3)Q_{fm|h} - 2(n_h - 1)Q_{fm|h}^2]}{\sum_h n_h Q_{fm|h}}\right\}, \quad (15)$$

où $Q_{fm|h} = Q_{f|h} S_{m|hf}$, $\mu_{fm} = E_\xi\{Y_{khfmi}\}$, et $\sigma_{fm}^2 = V_\xi\{Y_{khfmi}\}$.

Dans la pratique, $\bar{y}_{kfm}$ reposera sur un très petit nombre d'observations si seulement quelques consommateurs de la catégorie f achètent exactement m voitures. Pour obtenir une plus grande stabilité, on pourrait définir m comme une variable ordinale en créant un petit nombre de classes pour le total d'automobiles achetées. De cette manière, l'hypothèse (13) signifierait que la distribution du nombre d'automobiles de type k achetées est la même à l'intérieur d'une catégorie f et d'une classe m. En outre, on pourrait considérer ℓ comme une variable aléatoire continue et faire des hypothèses sur sa distribution, ce qui nous amènerait à des estimateurs par quotient ou à des estimateurs par régression.

Un second ensemble d'hypothèses paramétriques, plus contraignantes encore, est le suivant:

$$T_{\ell|hfm} = T_{\ell|fm} \quad \text{pour tous} \quad h,$$
$$S_{m|hf} = S_{m|f} \quad \text{pour tous} \quad h. \quad (16)$$

Ces hypothèses signifient qu'étant donné la catégorie f, la distribution conjointe du nombre d'automobiles de type k achetées et du nombre total d'automobiles achetées, tous modèles confondus (m) est indépendante de la strate (h). Dans ces conditions, l'estimateur du maximum de vraisemblance de Θ_k est défini par l'équation

$$\hat{\Theta}_k(3) = \sum_f \hat{N}_f \bar{y}_{kf}, \quad (17)$$

où $\bar{y}_{kf} = \sum_\ell \ell\, n_{f\ell}/n_f$ est le nombre moyen non pondéré (pour l'échantillon) d'automobiles de type k achetées par des consommateurs de la catégorie f, peu importe le nombre total d'automobiles achetées, et $\hat{N}_f = \sum_h N_h n_{hf}/n_h$ est l'estimateur pondéré du nombre total de consommateurs inclus dans la catégorie f. On peut montrer que, suivant les hypothèses (16),

$$V_\xi(\hat{\Theta}_k(3)) = \sum_h \sum_f \frac{N_h^2}{n_h} \mu_f^2 Q_{f|h}(1 - Q_{f|h}) - \sum_h \sum_f \sum_{f'} \frac{N_h^2}{n_h} \mu_f \mu_{f'} Q_{f|h} Q_{f'|h} \quad f \neq f' + \sum_h \sum_f \frac{N_h^2}{n_h} \frac{\sigma_f^2 Q_{f|h}}{\sum_h n_h Q_{f|h}} \left\{(1 - Q_{f|h}) + n_h Q_{f|h} + \left[\frac{1 + (2n_h - 3)Q_{f|h} - 2(n_h - 1)Q_{f|h}^2}{\sum_h n_h Q_{f|h}}\right]\right\}. \quad (18)$$

Si les hypothèses (16) étaient plausibles, $\bar{y}_{kf}$ reposerait sur un plus gros échantillon que $\bar{y}_{kfm}$ (équ. 14) et $\hat{\Theta}_k(3)$ serait donc plus stable.

L'estimateur du maximum de vraisemblance de la part de marché pour le produit de type k, R_k, suivant l'hypothèse (16), est défini

$$\hat{\Omega}_k(3) = \frac{\sum_f \hat{N}_f \bar{y}_{kf}}{\sum_f \hat{N}_f \bar{z}_f}, \quad (19)$$

où $\bar{z}_f$, défini comme $\bar{y}_{kf}$, est le nombre moyen non pondéré (pour l'échantillon) d'automobiles de toutes sortes achetées par des consommateurs de la catégorie f.

5. RÉSULTATS EMPIRIQUES

5.1 Estimation du nombre de consommateurs

Dans la section 4.2, nous avons étudié l'efficacité de $\hat{\Theta}_k'(2)$ pour diverses structures de population pour lesquelles l'hypothèse (5) se vérifiait. Cet exercice peut paraître douteux aux yeux du lecteur puisque l'hypothèse (5) ne se vérifie pas dans la pratique. Nous allons maintenant nous servir des données réelles d'enquête pour calculer $\hat{\Theta}_k'(2)$ pour une classification spéciale basée simultanément sur deux critères: la taille du parc d'automobiles *et* un indicateur permettant de savoir si le consommateur a acheté ou non des automobiles d'un type quelconque à des fins commerciales (voir tableau 4). Schaible, Brock et Schnack (1977) et Drew, Singh et Choudhry (1982) ont fait des analyses empiriques d'estimateurs synthétiques pour différentes situations.

Pour chacun des types de produit énumérés dans le tableau 2 (A à G), nous avons soumis à un test χ^2 l'hypothèse qu'étant donné la catégorie de la variable de condition (f), la décision d'un consommateur d'acheter le produit ne dépend pas de la strate (h). Notons que pour notre exemple, nous utilisons un plan d'échantillonnage aléatoire stratifié et les hypothèses de distribution multinomiale habituelles s'appliquent. Dans le cas de plans à plusieurs degrés, il faudrait modifier l'analyse chi carré ordinaire en utilisant par exemple des facteurs de correction de Rao-Scott. En pratique, il est difficile de trouver une classification de f qui soit telle que les hypothèses d'indépendance conditionnelle (5) se vérifient pour chaque type de produit. Néanmoins, pour la classification définie dans le tableau 4, nous avons observé que la variation de la probabilité qu'on achète un produit d'un type donné était expliquée en majeure partie par la catégorie de la variable de condition (f) et qu'une infime partie seulement de la variation des résiduels était attribuable à des différences entre les strates.

Tableau 4

Définition des catégories (f) de la variable de condition

Catégories f	Définition de f	
	Taille du parc	Achat de véhicules
1	Néant	0
2	1-4	> 0
3	5-8	> 0
4	9-15	> 0
5	16-25	> 0
6	26-50	> 0
7	51-100	> 0
8	101-200	> 0
9	201-550	> 0
10	> 550	> 0

Les estimations basées sur un modèle calculées à l'aide des équations (6) et (9) – c'est-à-dire les estimations relatives aux consommateurs $\hat{\Theta}_k'(2)$ et $\hat{\Omega}_k'(2)$ – sont reproduites dans le tableau 5. Les variances basées sur un modèle risquent d'exagérer la précision des estimateurs étant donné qu'elles reposent sur les hypothèses d'indépendance conditionnelle du modèle et que celles-ci peuvent être fausses dans la pratique. En revanche, il est toujours possible de calculer l'estimation empirique de la variance en p de l'estimateur fondé sur un modèle (voir Holt et Holmes 1993). Ce calcul ne nécessite aucune hypothèse de distribution ou d'indépendance conditionnelle et son résultat peut être considéré comme une mesure plus objective. Les estimations de l'erreur type calculées de cette manière figurent dans le tableau 5. Comme il s'agit d'estimations fondées sur un plan, elles renferment des corrections pour population finie. [Notons ici que les erreurs types basées sur un modèle pour $\hat{\Theta}_k'(2)$ (qui ne figurent pas dans le tableau 5) étaient environ 10% moins élevées que les erreur types fondées sur un plan pour tous les cas.]

Tableau 5

Estimations basées sur un modèle et erreurs types en p correspondantes pour certains types de produits

Produit (k)	Estimation du nombre de consommateurs		Estimation du nombre d'automobiles	
	Total $\hat{\Theta}_k'(2)$	Pénétration $\hat{\Omega}_k'(2)$	Total $\hat{\Theta}_k(3)$	Part de marché $\hat{\Omega}_k(3)$
A	63,433	.4070	263,511	.3722
	(2,230)	(.0105)	(13,007)	(.0048)
B	39,673	.2546	177,067	.2501
	(1,587)	(.0086)	(9,530)	(.0046)
C	21,930	.1407	65,357	.0923
	(1,142)	(.0066)	(3,836)	(.0027)
D	13,422	.0861	22,146	.0313
	(868)	(.0052)	(1,351)	(.0016)
E	7,366	.0473	15,798	.0223
	(675)	(.0041)	(1,223)	(.0014)
F	5,826	.0374	14,398	.0203
	(492)	(.0031)	(1,113)	(.0012)
G	7,686	.0493	11,207	.0158
	(633)	(.0039)	(813)	(.0011)

1ière ligne: estimation 2ième ligne: e.t. en p

Si nous comparons ces résultats à ceux du tableau 2, nous observons que les erreurs types des totaux estimés sont beaucoup moins élevées: de 30 à 40% moins élevées pour tous les produits sauf A et B (les principaux modèles), pour lesquels la différence est plutôt de 15 à 20%. Ces observations étaient prévisibles puisque le plan de sondage de l'enquête est optimal pour le total des ventes d'automobiles et donc, relativement efficace pour des produits qui occupent une grande part de marché. Par conséquent, la méthode fondée sur un modèle devrait être utile surtout pour des produits dont la part de marché est plus modeste.

En ce qui concerne l'estimation du taux de pénétration du marché, on observe aussi dans ce cas-là des erreurs types moins élevées pour tous les produits (des écarts de 30 à 40% sauf pour les produits A et B, pour lesquels les écarts sont un peu moins prononcés).

5.2 Estimation du nombre d'automobiles

Le tableau 5 contient aussi les estimations (basées sur un modèle) du nombre total d'automobiles de type k achetées, $\hat{\Theta}_k(3)$, et de la part de marché correspondante, $\hat{\Omega}_k(3)$, calculées à l'aide des équations (17) et (19) respectivement, pour la classification définie dans le tableau 4. Ces estimations sont accompagnées des erreurs types en p correspondantes.

Si nous comparons ces résultats aux estimations du tableau 2, nous observons une forte réduction des erreurs types des totaux estimés (de 40 à 80%) sauf pour le produit de catégorie B. De la même manière, les erreurs types des parts de marché estimées sont beaucoup moins élevées avec un modèle.

6. ANALYSE

L'utilisation d'hypothèses d'indépendance conditionnelle dans l'élaboration d'estimateurs fondés sur un modèle permet de considérer deux problèmes d'estimation: dans le premier cas (estimation de N_f: nombre de consommateurs appartenant à la catégorie f), on a recours à des probabilités de sélection inégales tandis que dans le second cas (estimation de la proportion de consommateurs de la catégorie f qui achètent un produit de type k, ou nombre moyen d'automobiles de type k qui sont achetées par des consommateurs de la catégorie f), on n'a pas recours à de telles probabilités, ce qui peut se traduire par un gain d'efficacité appréciable.

Si les hypothèses d'indépendance conditionnelle ne se vérifient pas, les estimateurs comporteront un biais résiduel au sens normal du plan, mais ce peut être un risque acceptable pour garantir la stabilité des estimateurs pour l'ensemble des produits. En ce qui concerne les résultats numériques observés dans les sections précédentes, seules les estimations basées sur un modèle relatives au produit B sont en dehors de l'intervalle de confiance à 95% basé sur l'estimateur direct. Les hypothèses d'indépendance conditionnelle dépendent du choix de la catégorie f et elles peuvent être testées à l'aide de tests chi carré pour tableaux de contingence.

Bien que les données du tableau 5 montrent que les erreurs types de plan des estimations basées sur un modèle sont généralement moins élevées que les erreurs types des estimations directes (tableau 2), on peut prétendre que les estimateurs fondés sur un modèle sont susceptibles d'être biaisés et que par conséquent, leur erreur quadratique moyenne (EQM) n'est pas moins élevée que celle des autres estimateurs. Le biais des estimateurs fondés sur un modèle viendra de l'inadéquation des hypothèses d'indépendance conditionnelle (par ex., équation (5)). On ne peut vérifier cette affirmation mais en comparant les données des tableaux 2 et 5, on peut avoir une idée de la taille du biais qui serait nécessaire pour que l'estimateur direct et l'estimateur fondé sur un modèle aient la même EQM. Prenons l'estimation du nombre total de consommateurs pour le produit de type E; cette estimation est fortement influencée par la procédure et donc, probablement très susceptible d'être entachée d'un biais. La variance (et donc l'EQM) de l'estimateur direct est $1{,}146^2 = 1{,}313{,}316$, tandis que celle de l'estimateur fondé sur un modèle est $675^2 = 455{,}625$. Par conséquent, il faudrait que la valeur estimée de 7,366 (selon un modèle) soit entachée d'un biais de 926 pour que l'EQM soit la même pour les deux estimateurs.

REMERCIEMENTS

Les auteurs tiennent à remercier les arbitres pour leurs commentaires utiles.

BIBLIOGRAPHIE

DREW, J.D., SINGH, M.P., et CHOUDHRY, G.H. (1982). Évaluation des techniques d'estimation pour les petites régions dans l'enquête sur la population active au Canada. *Techniques d'enquête*, 8, 19-52.

FAY, R.E., et HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

GHOSH, M., et RAO, J.N.K. (1994). Small area estimation: An appraisal. À paraître dans *Statistical Science*.

GONZALEZ, M.E. (1973). Use and evaluation of synthetic estimators. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.

GONZALEZ, M.E., et HOZA, C. (1978). Small area estimation with application to Unemployment and Housing Estimates. *Journal of the American Statistical Association*, 73, 7-15.

HOLT, D., et HOLMES, D.J. (1993). Small domain estimation for unequal probability survey designs. Working Paper Series, No. 2, Department of Social Statistics, University of Southampton, UK.

HOLT, D., et SMITH, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society*, Sér. A, 142, 33-46.

PLATEK, R., RAO, J.N.K., SÄRNDAL, C.-E., et SINGH, M.P. (1987). *Small Area Statistics*. New York: John Wiley and Sons.

SÄRNDAL, C.-E., et HIDIRIGLOU, M.A. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

SCHAIBLE, W.L., BROCK, D.B., et SCHNACK, G.A. (1977). An empirical comparison of the simple inflation, synthetic and composite estimators for small area statistics. *Proceedings of the Social Statistics Section, American Statistical Association*, 1017-1021.

Techniques d'enquête, juin 1994
Vol. 20, n° 1, pp. 35–46
Statistique Canada

MPLSE à données chronologiques pour petites régions évalués à l'aide de données d'enquête

A.C. SINGH, H.J. MANTEL et B.W. THOMAS[1]

RÉSUMÉ

Lorsqu'on effectue des estimations pour petites régions, il est courant de renforcer les estimations en "empruntant" de l'information à d'autres petites régions, car les estimations directes tirées de l'enquête présentent souvent une variabilité d'échantillonnage élevée. Il existe, pour résoudre ce problème de forte variabilité, un ensemble de méthodes, dites d'estimation composite, qui utilisent la combinaison linéaire d'un estimateur direct et d'un estimateur synthétique. La composante synthétique se fonde sur un modèle qui relie les moyennes des petites régions transversalement (entre les régions) et/ou par rapport au temps. Le meilleur prédicteur linéaire sans biais empirique (MPLSE) transversal est un estimateur composite fondé sur un modèle de régression linéaire comportant des effets de petite région. Dans cet article, nous examinons trois modèles qui généralisent le MPLSE transversal et permettent d'utiliser les données de plusieurs points dans le temps. Dans le premier modèle, les paramètres de régression sont aléatoires et présentent une dépendance sériale, mais les effets de petite région sont supposés indépendants dans le temps. Dans le deuxième modèle, les paramètres de régression ne sont pas aléatoires et peuvent prendre des valeurs communes dans le temps, mais il y a une dépendance sériale dans les effets de petite région. Le troisième modèle est plus général en ce sens qu'on suppose une dépendance sériale dans les paramètres de régression et dans les effets de petite région. Les estimateurs résultants, ainsi que certains estimateurs transversaux, sont évalués à l'aide des données bisannuelles de l'Enquête nationale sur les fermes et de l'Enquête sur les fermes de janvier de Statistique Canada.

MOTS CLÉS: Estimation composite; modèles d'espace d'états; filtre de Kalman; estimateur de Fay-Herriot.

1. INTRODUCTION

Un grand nombre d'études ont porté sur la façon de produire des estimations pour petites régions au moyen de données d'enquête transversales, en utilisant comme complément des données du recensement ou de sources administratives. On trouve dans Platek, Rao, Särndal et Singh (1987) un ensemble d'articles instructifs à ce sujet. Les méthodes d'estimation pour petites régions utilisées dans les programmes statistiques fédéraux des États-Unis font l'objet d'une évaluation par le Federal Committee on Statistical Methodology (1993). L'idée de base qui sous-tend toutes les méthodes d'estimation pour petites régions consiste à renforcer les estimations en empruntant de l'information à d'autres petites régions, c'est-à-dire en supposant que différentes régions sont reliées conformément à un modèle contenant des variables auxiliaires tirées des données complémentaires. Par ailleurs, comme il existe de nombreuses enquêtes à passages répétés, il serait également important d'emprunter de l'information à d'autres passages. Des méthodes utilisant des séries chronologiques ont récemment été mises au point dans le but d'améliorer les estimateurs pour petites régions; voir Pfeffermann et Burck (1990) et Rao et Yu (1992). Il est intéressant de signaler qu'après les travaux de Scott et Smith (1974) sur le renforcement des données d'enquête par le recours à des séries chronologiques, on n'a constaté que récemment un regain d'intérêt pour l'estimation d'agrégats tirés d'enquêtes complexes répétées à intervalles réguliers; voir par exemple Bell et Hillmer (1987), Binder et Dick (1989), Pfeffermann (1991) et Tiller (1992).

Dans cet article, nous examinons certaines généralisations naturelles du meilleur prédicteur linéaire sans biais (MPLS) pour petites régions quand on dispose d'une série chronologique d'estimations directes pour les petites régions. L'estimateur de Fay-Herriot (FH), fondé sur le lissage des estimateurs directs par une modélisation transversale des totaux des petites régions, est un important exemple de MPLS pour petites régions. Les estimateurs résultants sont des estimateurs composites (c.-à-d. des combinaisons convexes d'estimateurs directs et synthétiques), et ils sont appelés MPLS empiriques (MPLSE) lorsque des estimations de certaines composantes de variance sont insérées dans l'expression du MPLS. Les travaux de Fay et Herriot (1979) ont marqué une étape importante dans le domaine de l'estimation pour petites régions, car c'est probablement la première fois que celle-ci a été utilisée à grande échelle par des organismes gouvernementaux à des fins d'analyse des politiques. En nous servant de modèles structurels, nous obtenons des MPLSE à données chronologiques, qui comportent à la fois des données transversales et des données longitudinales. Le choix des modèles sous-jacents a été fait d'après des considérations heuristiques générales plutôt qu'en fonction de tests formels des modèles. Les tests formels de ce genre de modèles à l'aide de données d'enquête sont très difficiles et il n'existe pas beaucoup de telles données. Nous commençons plutôt par un modèle de régression qui est

[1] A.C. Singh et H.J. Mantel, Division des méthodes d'enquêtes sociales; B.W. Thomas, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6.

raisonnable pour la grande région, et dans lequel des effets aléatoires de petite région peuvent expliquer tout écart local par rapport au modèle global. Les paramètres de régression et les effets aléatoires de petite région peuvent évoluer dans le temps selon un modèle d'espace d'états également formulé de façon heuristique. Nous n'avons pas examiné ici le problème de l'estimation de l'erreur quadratique moyenne (EQM) de nos estimateurs. Des EQM correspondant aux modèles sous-jacents pourraient être définies et estimées pour bon nombre des estimateurs, mais notre article concerne principalement la performance des estimateurs dans un contexte d'échantillonnage répété. L'estimation de l'EQM est un problème important et difficile, et la disponibilité d'estimateurs fiables de l'EQM pourrait être un facteur important dans le choix des estimateurs.

Le but principal de cet article est de comparer des MPLSE utilisant des données chronologiques avec des estimateurs transversaux (p. ex. avec des estimateurs pour domaines avec stratification a posteriori, synthétique, FH et lié à la taille de l'échantillon). Dans la modélisation chronologique des estimations directes des petites régions, nous supposons que les erreurs de l'enquête ne sont pas corrélées dans le temps. Quand les erreurs de l'enquête sont corrélées dans le temps et peuvent être décrites par un modèle raisonnable (p. ex. ARMM), on peut utiliser la méthode de Pfeffermann (1991) pour obtenir des MPLSE à données chronologiques, au moyen du filtre de Kalman. Rao et Yu (1992) obtiennent des MPLSE pour un modèle auquel le filtre de Kalman ne peut s'appliquer, avec des erreurs de l'enquête présentant une structure de corrélation arbitraire dans le temps, mais non corrélées entre les régions. Ils produisent également des approximations du second ordre, ainsi que l'estimation, de l'erreur quadratique moyenne en vertu de leur modèle. Quand il est difficile de formuler un modèle pour les erreurs de l'enquête corrélées, il est possible, au moyen d'un filtre de Kalman convenablement modifié, d'obtenir de bons estimateurs sous-optimaux (Singh et Mantel 1991).

Dans cet article, nous donnons les résultats d'une étude empirique de l'efficacité de MPLSE à données chronologiques. L'étude utilise des simulations de Monte Carlo fondées sur des données de séries chronologiques réelles tirées des enquêtes bisannuelles sur les fermes réalisées par Statistique Canada. Voici les principales conclusions de l'étude:

(i) Les MPLSE à données chronologiques peuvent produire des gains d'efficacité raisonnables par rapport aux estimateurs transversaux.

(ii) Dans la classe des méthodes avec données chronologiques examinées dans cet article, il est avantageux de permettre qu'il existe une dépendance sériale dans les effets aléatoires de petite région.

(iii) Il est prévisible que toute version lissée de l'estimateur direct pour petites régions sera biaisée, mais les MPLSE à données chronologiques affichent un biais moindre que les méthodes de lissage transversal.

La section 2 présente une description de diverses méthodes transversales servant à l'estimation pour petites régions. Les MPLSE à données chronologiques sont décrits à la section 3, tandis que les détails et les résultats de l'étude comparative de Monte Carlo sont donnés à la section 4. Enfin, la section 5 présente quelques commentaires en guise de conclusion.

2. MÉTHODES FONDÉES SUR DES DONNÉES TRANSVERSALES

Dans cette section, nous décrivons certaines méthodes bien connues d'estimation pour petites régions fondées uniquement sur des données transversales. Ghosh et Rao (1994) font un survol intéressant de divers estimateurs pour petites régions.

Soit $\underset{\sim}{\Theta}$ le vecteur des totaux Θ_k, $k = 1, \ldots, K$ de la population pour de petites régions. Dans la présente section, qui traite des méthodes fondées sur des données transversales, nous ne tenons pas compte du lien entre $\underset{\sim}{\Theta}$ et le temps t, à des fins de simplicité.

2.1 Méthode 1 (estimateur d'expansion pour domaines)

Cet estimateur s'exprime ainsi

$$g_{1k} = \sum_{j \epsilon s_k} d_j y_j,$$

où d_j est le poids de l'enquête pour l'unité j de l'échantillon. Dans le cas de l'échantillonnage aléatoire simple stratifié, dont nous nous servons dans notre étude de simulation à la section 4, nous avons:

$$g_{1k} = \sum_h (N_h/n_h) \sum_{j \epsilon s_{hk}} y_{hj}, \tag{2.1}$$

où y_{hj} est la $j$ième observation dans la $h$ième strate, s_{hk} désigne l'ensemble des n_{hk} unités de l'échantillon se trouvant dans la $k$ième petite région et dans la $h$ième strate, et n_h, N_h désignent respectivement les tailles de l'échantillon et de la population pour la $h$ième strate. Cet estimateur est souvent peu fiable, car n_{hk}, la taille de l'échantillon aléatoire dans la petite région, peut avoir une valeur espérée faible et présenter une variabilité élevée. Avec comme condition la taille d'échantillon réalisée n_{hk}, g_{1k} est biaisé. Toutefois, en l'absence de condition, cet estimateur est non biaisé pour Θ_k.

2.2 Méthode 2 (estimateur pour domaines avec stratification a posteriori)

Nous appellerons aussi cet estimateur l'estimateur direct pour petites régions. Si la taille de la population N_{lk} est connue pour certaines strates a posteriori représentées par l'indice l, l'efficacité de l'estimateur g_{1k} pourrait être améliorée par une stratification a posteriori. Nous définissons

$$g_{2k} = \sum_l N_{lk} \sum_{j \epsilon s_{lk}} d_j y_j \Big/ \sum_{j \epsilon s_{lk}} d_j = \sum_l N_{lk} \bar{y}_{lk}.$$

Dans nos simulations, les strates a posteriori sont les intersections des strates du plan avec les petites régions, ce qui donne

$$g_{2k} = \sum_h (N_{hk}/n_{hk}) \sum_{j \in s_{hk}} y_{hj} = \sum_h N_{hk} \bar{y}_{hk}. \quad (2.2)$$

Il se peut que cet estimateur ne soit pas lui non plus suffisamment fiable, vu la possibilité que les n_{hk} aient une valeur espérée faible. Si $n_{hk} = 0$, l'estimateur ci-dessus n'est pas défini. Il est d'usage de remplacer $\bar{y}_{hk}$ par 0 lorsque $n_{hk} = 0$. Dans l'étude empirique présentée dans cet article, nous avons remplacé $\bar{y}_{hk}$ par l'estimation synthétique $(\bar{X}_{hk}/\bar{X}_h)\bar{y}_h$, où X est une covariable appropriée, dans les cas où $n_{hk} = 0$.

L'estimateur g_{2k} dans (2.2) est conditionnellement sans biais (étant donné $n_{hk} > 0$) et approximativement sans biais en l'absence de condition. L'annexe A.1 donne les détails de l'estimation de l'erreur quadratique moyenne conditionnelle, v_k, de g_{2k}.

2.3 Méthode 3 (estimateur synthétique)

Il est possible de définir un estimateur plus efficace en supposant un modèle qui permette de renforcer les estimations par l'"emprunt" d'information à d'autres petites régions. On obtient ainsi des estimateurs synthétiques; voir par exemple Gonzalez (1973) et Ericksen (1974). Supposons que différents totaux de petites régions soient reliés par une variable auxiliaire X_k selon un modèle linéaire comme

$$\Theta_k = \beta_1 + \beta_2 X_k, \; k = 1, \ldots, K, \quad (2.3a)$$

ou, en notation matricielle,

$$\underset{\sim}{\Theta} = F\underset{\sim}{\beta}, \quad (2.3b)$$

où $F = (F_1, F_2, \ldots, F_K)'$, $F_k = (1, X_k)'$. Considérons maintenant un modèle applicable aux estimateurs directs pour petites régions g_{2k}, de la forme suivante

$$\underset{\sim}{g}_2 = F\underset{\sim}{\beta} + \underset{\sim}{\epsilon},$$

où $\underset{\sim}{g}_2 = (g_{21}, \ldots, g_{2K})'$, $\underset{\sim}{\epsilon} = (\epsilon_1, \ldots, \epsilon_K)'$, ϵ_k, les ϵ_k sont des erreurs de l'enquête non corrélées avec moyenne 0 et variance v_k. Notons que les g_{2k} ne sont pas corrélés entre les régions puisqu'ils sont conditionnellement (étant donné n_{hk}) sans biais et que les échantillons de petites régions différentes sont conditionnellement indépendants.

Si l'on désigne par $\hat{\underset{\sim}{\beta}}$ l'estimation de $\underset{\sim}{\beta}$ selon les moindres carrés pondérés (MCP), nous obtenons l'estimateur synthétique de régression de Θ_k en vertu du modèle supposé:

$$\underset{\sim}{g}_3 = F\hat{\underset{\sim}{\beta}}.$$

L'estimateur ci-dessus pourrait être fortement biaisé à moins que le modèle (2.3) ne soit raisonnablement vérifié. Ce modèle pourrait ne pas être réaliste, car il ne laisse aucune place à des fluctuations aléatoires ou à un effet aléatoire de petite région (disons (a_k)).

2.4 Méthode 4 (estimateur de Fay-Herriot ou MPLSE)

En adoptant l'approche bayesienne empirique de Fay et Herriot (1979) ou l'approche plus générale du meilleur prédicteur linéaire sans biais (voir par exemple Battese, Harter et Fuller (1988), et Pfeffermann et Barnard (1991)), on peut réduire considérablement le biais de l'estimateur synthétique par l'utilisation d'un estimateur composite. Pour une des premières communications touchant l'estimation composite, voir Schaible (1978). L'estimateur composite est obtenu sous forme d'une combinaison convexe de $\underset{\sim}{g}_2$ et d'un $\underset{\sim}{g}_3$ modifié. On suppose à cette fin que

$$\underset{\sim}{\Theta} = F\underset{\sim}{\beta} + \underset{\sim}{a}, \quad (2.4)$$

où les a_k sont des effets aléatoires non corrélés de petite région avec moyenne 0 et variance w_k connus jusqu'à une constante. Dans notre étude empirique décrite plus loin, nous prenons $w_k = w$. Notre modèle pour $\underset{\sim}{g}_2$ est donc

$$\underset{\sim}{g}_2 = F\underset{\sim}{\beta} + \underset{\sim}{a} + \underset{\sim}{\epsilon}. \quad (2.5)$$

Ici, on suppose également que $\underset{\sim}{a}$ est non corrélé avec $\underset{\sim}{\epsilon}$. Le MPLS de $\underset{\sim}{\Theta}$ en vertu du modèle défini en (2.4) et (2.5) est

$$\begin{aligned} \underset{\sim}{g}_4 &= \underset{\sim}{g}_3^* + \Lambda(\underset{\sim}{g}_2 - \underset{\sim}{g}_3^*) \\ &= \Lambda\underset{\sim}{g}_2 + (I - \Lambda)\underset{\sim}{g}_3^*, \end{aligned} \quad (2.6)$$

où

$$\Lambda = (V^{-1} + W^{-1})^{-1}V^{-1} = WU^{-1}, \; U \equiv V + W,$$

$$V = \text{diag}(v_1, \ldots, v_K), \quad W = \text{diag}(w_1, \ldots, w_K),$$

et $\underset{\sim}{g}_3^* = F\underset{\sim}{\beta}^*$, $\underset{\sim}{\beta}^*$ est l'estimation MCP de $\underset{\sim}{\beta}$ en vertu du modèle (2.5). On suppose que les matrices de covariance V et W sont toutes deux connues pour le calcul du MPLS.

L'expression (2.6) découle des résultats généraux relatifs aux modèles linéaires avec effets aléatoires; voir par exemple Rao (1973, p. 267) et Harville (1976). Le MPLS ou le MELS (meilleur estimateur linéaire sans biais) de $F\underset{\sim}{\beta}$ est $\underset{\sim}{g}_3^*$, et le MPLS de $\underset{\sim}{a}$ est $\Lambda(\underset{\sim}{g}_2 - \underset{\sim}{g}_3^*)$. Il est intéressant de signaler que la structure du MPLS ne change pas peu importe que $\underset{\sim}{\beta}$ soit connu ou non. Toutefois, comme il faut s'y attendre, son EQM change, en raison de l'estimation de $\underset{\sim}{\beta}$.

Si V et W sont remplacées par des estimations, l'estimateur $\underset{\sim}{g}_4$ est appelé MPLSE. Notons que le modèle (2.4) est plus réaliste que le modèle (2.3) et que, par conséquent, la performance de $\underset{\sim}{g}_4$ devrait être très bonne. L'estimateur $\underset{\sim}{g}_4$ s'approche de $\underset{\sim}{g}_2$ quand les v_k deviennent petits, c.-à-d. quand les n_{hk} deviennent grands. Toutefois, il demeure biaisé, en général, pour $\underset{\sim}{\Theta}$ donné, le biais tendant vers 0 à mesure que les v_k deviennent petits.

2.5 Méthode 5 (estimateur lié à la taille de l'échantillon)

Un autre estimateur composite possible est l'estimateur lié à la taille de l'échantillon de Drew, Singh et Choudhry (1982). Il est défini ainsi

$$\underset{\sim}{g}_5 = \Delta \underset{\sim}{g}_2 + (I - \Delta)\underset{\sim}{g}_3,$$

où $\Delta = \text{diag}(\delta_1, \ldots, \delta_K)$,

$$\delta_k = \begin{cases} 1 & \text{si } \sum_{j \in s_k} d_j \geq \lambda N_k, \\ \sum_{j \in s_k} d_j / \lambda N_k & \text{sinon} \end{cases} \tag{2.7}$$

et le paramètre λ est choisi subjectivement comme moyen de contrôler la contribution de la composante synthétique. L'estimateur ci-dessus tient compte des tailles d'échantillon réalisées n_{hk}, et si ces dernières sont jugées suffisamment élevées selon la condition énoncée en (2.7), il ne dépend pas de l'estimateur synthétique. Cette propriété s'apparente à celle de $\underset{\sim}{g}_4$; toutefois, contrairement à $\underset{\sim}{g}_4$, l'estimateur ci-dessus ne tient pas compte des tailles relatives de la variation intra-région et de la variation inter-régions. Rao et Choudhry (1993) ont démontré de façon empirique comment les MPLSE peuvent parfois se révéler supérieurs aux estimateurs liés à la taille de l'échantillon, notamment quand la variation inter-régions est peu élevée par rapport à la variation intra-région. Särndal et Hidiroglou (1989) ont également proposé des estimateurs semblables à l'estimateur ci-dessus lié à la taille de l'échantillon.

3. MÉTHODES FONDÉES SUR DES DONNÉES TRANSVERSALES ET CHRONOLOGIQUES COMBINÉES

Supposons qu'on dispose des données recueillies à plusieurs moments différents, $t = 1, \ldots, T$, sous forme des estimateurs directs pour petites régions $\underset{\sim}{g}_{2t}$, où $\underset{\sim}{g}_{2t}$ est le vecteur des estimations g_{2k} de (2.2) d'après les données dans le temps t, ainsi que des totaux de la population pour la variable auxiliaire dans les petites régions. Nous allons maintenant présenter quelques estimateurs qui généralisent l'estimateur de Fay-Herriot $\underset{\sim}{g}_{4T}$ de différentes façons, en tenant compte de la dépendance sériale des estimation directes $\{\underset{\sim}{g}_{2t} : t = 1, \ldots, T\}$. Rappelons que dans le cas de l'estimateur de Fay-Herriot, le modèle formulé pour $\underset{\sim}{\Theta}_T$ a deux composantes: la composante structurelle $F_T \underset{\sim}{\beta}_T$ et la composante des régions $\underset{\sim}{a}_T$. L'estimateur $\underset{\sim}{g}_{4T}$ emprunte de l'information aux régions pour le moment courant T et est donné par la somme de deux composantes, chacune étant le MPLSE (MELS) pour l'effet aléatoire (fixe) correspondant, c.-à-d.

$$\underset{\sim}{g}_{4T} = F_T \underset{\sim}{\beta}_T^* + \underset{\sim}{a}_T^*. \tag{3.1}$$

Les méthodes fondées sur des données chronologiques pourraient toutefois emprunter également de l'information recueillie à d'autres moments. Nous présentons trois estimateurs qui sont justifiés par des modèles structurels particuliers de la dépendance sériale. Ces estimateurs sont tous trois optimaux en vertu de cas spéciaux différents d'un modèle chronologique structurel pour les estimations directes de petites régions $\{\underset{\sim}{g}_{2t} : t = 1, \ldots, T\}$ définies par le modèle d'espace d'états suivant. Supposons que $\underset{\sim}{\alpha}_t$ désigne $(\underset{\sim}{\beta}_t', \underset{\sim}{a}_t')'$ et que H_t désigne (F_t, I). On a

$$\begin{aligned} \underset{\sim}{g}_{2t} &= \underset{\sim}{\Theta}_t + \underset{\sim}{\xi}_t, \\ \underset{\sim}{\Theta}_t &= F_t \underset{\sim}{\beta}_t + \underset{\sim}{a}_t \equiv H_t \underset{\sim}{\alpha}_t \end{aligned} \tag{3.2a}$$

et

$$\underset{\sim}{\alpha}_t = G_t \underset{\sim}{\alpha}_{t-1} + \underset{\sim}{\zeta}_t, \tag{3.2b}$$

où

$$G_t = \begin{pmatrix} G_t^{(1)} & 0 \\ 0 & G_t^{(2)} \end{pmatrix}, \quad \underset{\sim}{\zeta}_t = \begin{pmatrix} \underset{\sim}{\xi}_t \\ \underset{\sim}{\eta}_t \end{pmatrix}, \tag{3.2c}$$

et selon les hypothèses habituelles relatives aux erreurs aléatoires, c.-à-d. que $\underset{\sim}{\xi}_t$, $\underset{\sim}{\zeta}_t$ sont non corrélés, $\underset{\sim}{\zeta}_t$ n'est pas corrélé avec $\underset{\sim}{\alpha}_s$ pour $s < t$, et que $\underset{\sim}{\xi}_t \sim (\underset{\sim}{0}, V_t)$, $\underset{\sim}{\zeta}_t \sim (\underset{\sim}{0}, \Gamma_t)$ où Γ_t est la matrice diagonale par blocs $\{B_t, Q_t\}$. Les matrices de covariance V_t, B_t, et Q_t sont généralement diagonales. Si $G_t^{(1)} = I$ et $G_t^{(2)} = I$, alors $\underset{\sim}{\beta}_t$ et $\underset{\sim}{a}_t$ évoluent selon une marche aléatoire.

Ce modèle appartient à la classe générale définie par Pfeffermann et Burck (1991) au moyen de modèles chronologiques structurels. Le but principal de leur étude était de démontrer comment la prise en compte des corrélations transversales entre petites régions voisines (en sus des corrélations sériales) et l'inclusion de certaines modifications visant à assurer la robustesse (pour se prémunir contre les défaillances du modèle) pouvaient améliorer la performance des estimateurs fondés sur des modèles chronologiques. Ils ont par ailleurs utilisé la méthode du maximum de vraisemblance dans des conditions de normalité pour estimer les paramètres du modèle. L'objet principal du présent article, en revanche, est l'évaluation de Monte Carlo d'une classe spéciale d'estimateurs à données chronologiques (reliés à l'estimateur de Fay-Herriot) choisis d'après des considérations heuristiques plutôt que par un ajustement de modèle. Les méthodes examinées pourraient donc être considérées comme des méthodes assistées par un modèle, dont la performance sera évaluée dans un cadre fondé sur le plan (c.-à-d. un échantillonnage répété) au moyen d'une simulation de Monte Carlo. De plus, comme nous le verrons plus loin, pour les types de dépendance sériale visés, les paramètres du modèle peuvent être estimés avec une relative simplicité par la méthode des moments, sans qu'il soit nécessaire de faire aucune hypothèse quant à la distribution (p. ex. distribution normale).

Pour trouver l'estimateur optimal (MPLS) de $\underset{\sim}{\Theta}_T$ dans (3.2) à partir de toutes les estimations directes jusqu'au temps T, nous avons d'abord trouvé le MPLS $\tilde{\underset{\sim}{\alpha}}_T$ de $\underset{\sim}{\alpha}_T$, d'où nous avons pu obtenir le MPLS de $\underset{\sim}{\Theta}_T$ sous la forme $H_T\tilde{\underset{\sim}{\alpha}}_T$. Il est possible, bien que peu pratique, d'obtenir $\tilde{\underset{\sim}{\alpha}}_T$ directement à partir des données complètes en recourant à la théorie des modèles linéaires avec effets aléatoires. Toutefois, comme les $\underset{\sim}{\alpha}_T$ sont reliés dans le temps selon l'équation de transition (3.2b), il est plus commode de le calculer récursivement au moyen du filtre de Kalman. On a généralement considéré le filtre de Kalman comme une technique bayesienne dans laquelle à chaque temps t, on met à jour la distribution postérieure de $\underset{\sim}{\alpha}_t$ pour les données jusqu'au temps $t-1$ afin d'obtenir la distribution postérieure de $\underset{\sim}{\alpha}_t$ pour les données jusqu'au temps t. Bien qu'il soit instructif de percevoir le filtre de Kalman de cette façon, ce n'est pas nécessaire dans le cas de modèles linéaires mixtes. Supposons que $\tilde{\underset{\sim}{\alpha}}_{T|s}$ dénote le MPLS de $\underset{\sim}{\alpha}_T$ d'après les données jusqu'au temps s, $s < T$. C'est une fait connu (voir Duncan et Horn 1972) que pour la structure spéciale de dépendance sérielle considérée ici, le MPLS $\tilde{\underset{\sim}{\alpha}}_T$ de $\underset{\sim}{\alpha}_T$ d'après les données jusqu'au temps T est le même que le MPLS de $\underset{\sim}{\alpha}_T$ d'après $\tilde{\underset{\sim}{\alpha}}_{T|s}$ et les $T-s$ dernières observations. En d'autres termes, l'information contenue dans les données précédentes peut être condensée en un MPLS approprié avant que soient ajoutées des données de points dans le temps plus récents. On trouve dans le chapitre 3 de Harvey (1989) une bonne description du filtre de Kalman.

3.1 Méthode 6 (MPLSE-I à données chronologiques)

Pour le premier estimateur, nous laissons $\underset{\sim}{\beta}_t$ évoluer dans le temps (selon une marche aléatoire), mais nous supposons que $\underset{\sim}{a}_t$ est sérialement indépendant. Les équations du modèle d'espace d'états sont dans ce cas semblables à celles de (3.2), sauf que l'indépendance sérielle des $\underset{\sim}{a}_t$ fait en sorte que $G_t^{(2)} = 0$. Il en résulte un estimateur composite

$$\underset{\sim}{g}_{6T} = F_T\tilde{\underset{\sim}{\beta}}_T + \tilde{\underset{\sim}{a}}_T. \tag{3.3}$$

Notons que $\tilde{\underset{\sim}{\beta}}_T$ dans (3.3) se fonderait maintenant sur toutes les estimations des petites régions jusqu'au temps T et serait donc différent de $\underset{\sim}{\beta}_T^*$ dans (3.1), qui dépend uniquement des estimations directes au temps T. L'estimateur $\tilde{\underset{\sim}{a}}_T$, par conséquent, serait lui aussi différent de la composante correspondante $\underset{\sim}{a}_T^*$ dans (3.1).

Dans l'étude de simulation décrite plus loin, nous prenons $G_t^{(1)} = I$, $B_t = \text{diag}(\gamma_1^2, \gamma_2^2)$, ce qui correspond à un modèle de marche aléatoire, et $Q_t = \tau^2 I$. L'annexe A.2 montre comment les paramètres γ_1^2, γ_2^2, et τ^2 sont estimés par la méthode des moments. Le filtre de Kalman peut ensuite être appliqué, avec les valeurs initiales de $\tilde{\underset{\sim}{\alpha}}_1$ et de son EQM obtenues à partir de l'estimateur FH au temps $t = 1$, ce qui donne le MPLSE de $\tilde{\underset{\sim}{\alpha}}_T$. Alors, $H_T\tilde{\underset{\sim}{\alpha}}_T$ est l'estimateur MPLSE-I à données chronologiques g_{6T} au temps T.

Comme l'a signalé un arbitre, quand le nombre de petites régions est très élevé, ou quand la variation de $\underset{\sim}{\beta}_t$ par rapport à t est relativement grande, il y a peu de différence entre g_{6T} et g_{4T}. En fait, il y a peu de différence entre les performances de ces deux estimateurs dans notre étude de simulation décrite à la section 4.

3.2 Méthode 7 (MPLSE-II à données chronologiques)

Pour le deuxième estimateur, nous supposons que $\underset{\sim}{\beta}_t$ est fixe (il peut ou non être commun à différents points dans le temps) et que les effets des régions $\underset{\sim}{a}_t$ présentent une dépendance sérielle conforme, par exemple, à une marche aléatoire. Cette généralisation à des données chronologiques est comparable au modèle proposé par Rao et Yu (1992). L'estimateur composite résultant aura la même forme qu'en (3.1), c.-à-d.

$$\underset{\sim}{g}_{7T} = F_T\tilde{\underset{\sim}{\beta}}_T + \tilde{\underset{\sim}{a}}_T, \tag{3.4}$$

mais les estimations des composantes $\tilde{\underset{\sim}{\beta}}_T$ et $\tilde{\underset{\sim}{a}}_T$ seront différentes. Deux cas doivent être examinés.

3.2.1 Cas 1: Supposons que les $\underset{\sim}{\beta}_t$ sont fixes et invariables dans le temps, mais que les $\underset{\sim}{a}_t$ présentent une dépendance sérielle. Alors, dans (3.2), $G_t^{(1)} = I$ et $B_t = 0$. Si l'on pose Q_t égal à $\tau^2 I$, le seul paramètre inconnu τ^2 peut être estimé par la méthode des moments; voir l'annexe A.2. Nous désignerons par g_{7T} le MPLSE qui est alors obtenu après substitution de l'estimation du paramètre.

3.2.2 Cas 2: Nous supposons ici que les $\underset{\sim}{\beta}_t$ sont fixes, mais qu'ils ne sont pas les mêmes à différents points dans le temps. Les effets des régions $\underset{\sim}{\alpha}_t$ évoluent dans le temps comme dans le cas 1. Dans (3.2), nous avons $G_t^{(1)} = 0$ et $B_t = mI$, où m est un nombre élevé. Les expressions de $\tilde{\underset{\sim}{\alpha}}_T$ et de son EQM obtenues par le filtre de Kalman dans ce cas donnent les formules correctes quand $m \to \infty$ (voir Sallas et Harville 1981). Les équations de mise à jour du filtre de Kalman pour $\tilde{\underset{\sim}{a}}_t$ ont alors cette forme spéciale

$$\tilde{\underset{\sim}{\beta}}_t = (F_t' A_t^{-1} F_t)^{-1} F_t' A_t^{-1} (\underset{\sim}{g}_{2t} - G_t^{(2)} \tilde{\underset{\sim}{a}}_{t-1});$$

$$\tilde{\underset{\sim}{a}}_t = G_t^{(2)} \tilde{\underset{\sim}{a}}_{t-1} + P_{t|t-1} A_t^{-1} (\underset{\sim}{g}_{2t} - G_t^{(2)} \tilde{\underset{\sim}{a}}_{t-1} - F_t \tilde{\underset{\sim}{\beta}}_t);$$

$$P_t = P_{t|t-1} - P_{t|t-1} A_t^{-1} (A_t - F_t (F_t' A_t^{-1} F_t)^{-1} F_t') A_t^{-1} P_{t|t-1},$$

où $A_t = P_{t|t-1} + V_t$, P_t est l'EQM de $\tilde{\underset{\sim}{a}}_t$ autour de $\underset{\sim}{a}_t$, et $P_{t|t-1} = G_t^{(2)} P_{t-1} \{G_t^{(2)}\}' + Q_t$ est l'EQM de $G_t^{(2)} \tilde{\underset{\sim}{a}}_{t-1}$ en tant qu'estimateur de $\underset{\sim}{a}_t$. Le MPLSE à données chronologiques, dans ce cas, est désigné par $\underset{\sim}{g}_{7T}^*$.

3.3 Méthode 8 (MPLSE-III à données chronologiques)

Dans le troisième estimateur, $\underset{\sim}{\beta}_t$ et $\underset{\sim}{a}_t$ peuvent tous deux évoluer dans le temps. Il en résulte une dépendance sérielle plus complexe que dans (3.3) ou (3.4). Cet estimateur aura une forme semblable à celle de (3.1) et peut s'écrire ainsi

$$\underset{\sim}{g}_{8T} = F_T\tilde{\tilde{\underset{\sim}{\beta}}}_T + \tilde{\tilde{\underset{\sim}{a}}}_T. \tag{3.5}$$

Comme auparavant, si $B_t = \text{diag}[\gamma_1^2, \gamma_2^2]$ et $Q_t = \tau^2 I$, les paramètres du modèle τ^2, γ_1^2, γ_2^2 peuvent être estimés par la méthode des moments décrite à l'annexe A.2. Le MPLSE résultant pour $\underset{\sim}{\Theta}_T$ sera désigné par $\underset{\sim}{g}_{8T}$.

Il est intéressant de noter que bon nombre des estimateurs examinés jusqu'ici sont optimaux en vertu de cas spéciaux du modèle qui sous-tend $\underset{\sim}{g}_{8T}$. Comme nous l'avons vu, on obtient les MPLSE à données chronologiques des méthodes 6 et 7 en imposant des restrictions aux matrices G_t and Γ_t. Dans le cas des estimateurs de Fay-Herriot transversaux de la section 2.4, les données sont limitées à un seul point dans le temps. Les estimateurs synthétiques de la section 2.3 sont des cas spéciaux des estimateurs de Fay-Herriot avec variance zéro pour les effets aléatoires de petite région, et l'estimateur direct (de stratification a posteriori) est obtenu à la limite, quand la variance des effets de petite région tend vers l'infini.

Une autre généralisation pourrait être utile: permettre des corrélations entre les effets de petites régions voisines. Il suffit pour cela que la matrice Q_t de (3.2) puisse être non diagonale; toutefois, il n'est pas évident de savoir quelle serait la structure de corrélation appropriée dans Q_t.

4. ÉTUDE DE MONTE CARLO

Les méthodes transversales et les méthodes avec données chronologiques ont été comparées de façon empirique au moyen d'une simulation de Monte Carlo. Nous avons utilisé à cette fin une série chronologique réelle tirée des enquêtes bisannuelles sur les fermes de Statistique Canada, soit l'Enquête nationale sur les fermes (en juin) et l'Enquête sur les fermes de janvier. En raison de la refonte qui a suivi le recensement de l'agriculture de 1986, nous avons utilisé les données des six points dans le temps qui étaient disponibles à partir de l'été 1988 pour créer une pseudo-population à des fins de simulation. Nous avons ajouté à cet ensemble des données de l'année du recensement (1986). Nous disposions ainsi de données pour un point additionnel dans le temps, mais il y avait en conséquence trois points manquants dans notre série. Ces vides peuvent toutefois être comblés facilement par des méthodes propres aux séries chronologiques. Notons que la série ainsi obtenue, bien qu'elle soit courte, est jugée adéquate à des fins d'illustration. Le paramètre d'intérêt choisi est le nombre total de veaux et bovins pour chaque district agricole (défini comme la petite région) à chaque point dans le temps. Pour plus de simplicité, des échantillons aléatoires stratifiés indépendants ont été tirés de la pseudo-population pour chaque point dans le temps, bien que les enquêtes sur les fermes utilisent des panels avec renouvellement. Pour modéliser la dépendance des estimations directes des petites régions dans le temps, on a supposé que les totaux globaux sous-jacents des petites régions étaient reliés en vertu d'un quelconque processus aléatoire. La variable auxiliaire utilisée dans le modèle était la valeur rajustée par quotient du total de veaux et bovins du recensement de 1986 dans chaque petite région. Cette variable présentait de fortes corrélations avec la variable correspondante dans le temps au niveau des fermes. Les détails de l'étude empirique sont présentés dans les paragraphes qui suivent.

4.1 Conception de l'étude de simulation

Il nous faut d'abord construire une pseudo-population à partir des données de l'enquête pour six points dans le temps (juin 1988, janvier 1989, . . . , janvier 1991). Le plan réel comporte deux bases (liste et base aréolaire), avec un échantillonnage stratifié à un seul degré pour la liste et un échantillonnage stratifié à deux degrés pour la base aréolaire; pour plus de détails, voir Julien et Maranda (1990). Nous avons décidé de nous limiter aux données provenant de la liste, parce que cette dernière contient les fermes qui existaient au moment du recensement de 1986 et que la variable auxiliaire choisie pour construire le modèle se fondait sur de l'information du recensement de 1986. En outre, nous avons choisi d'utiliser les données de la province de Québec, parce que l'échantillon aréolaire n'y constitue qu'une faible fraction de l'échantillon total et que la variation estimée des coefficients pour les douze districts agricoles (c.-à-d. les petites régions visées) de cette province représentait un vaste intervalle pour les variables relatives au bétail. Pour éviter d'avoir à tenir compte de la variabilité attribuable à l'évolution de la population sous-jacente, nous n'avons conservé que les fermes qui avaient répondu à chacune des six occasions. Par ailleurs, les unités agricoles appartenant à des exploitations multiples à l'une ou l'autre des sept occasions (y compris le recensement) ont été exclues à cause de la difficulté d'extraire les données de chaque ferme à partir des données sommaires des exploitations multiples, et des changements apportés à leurs modalités de déclaration au fil du temps.

Les diverses exclusions ci-dessus étaient justifiées par la volonté d'établir des comparaisons plus nettes entre les estimateurs pour petites régions. Le nombre total d'unités agricoles après les exclusions était de 1,160, sur un total de plus de 40,000 fermes figurant dans la liste. Pour la pseudo-population, nous avons reproduit les 1,160 unités agricoles proportionnellement à leur poids d'échantillonnage, de sorte que la taille totale N de la pseudo-population était de 10,362 unités, une quantité se prêtant à une simulation sur micro-ordinateur.

La pseudo-population a été stratifiée en quatre strates à tirage partiel et une strate à tirage complet, la variable de stratification étant les totaux de veaux et bovins du recensement de 1986. Bien que nous n'ayons pas examiné d'autres stratifications ou tailles d'échantillon dans notre étude de simulation, rien ne porte à croire que nos conclusions seraient sensiblement modifiées si nous le faisions. La règle de l'écart sigma (Julien et Maranda 1990) a servi à définir la strate à tirage complet. Pour appliquer la règle de l'écart sigma, nous cherchons la plus petite valeur de la population qui soit supérieure à la médiane et pour laquelle la distance jusqu'à la valeur de la population suivante, en ordre de taille, est d'au moins un écart-type de la population; toutes les unités situées au-delà de ce point sont placées dans la strate à tirage complet. L'algorithme de Sethi (1963) a servi à déterminer les limites de

stratification optimales pour les strates à tirage partiel. La répartition optimale de Neyman a été utilisée pour déterminer les tailles des échantillons des strates, de façon à optimiser la précision de l'estimation provinciale de la quantité totale. Pour une taille globale d'échantillon de 207 (taux d'échantillonnage de 2%), il en est résulté des tailles de 51, 62, 48 et 35 pour des strates à tirage partiel comprenant respectivement 5,001, 3,188, 1,850 et 312 fermes, tandis que la strate à tirage complet comprenait 11 unités. Le nombre espéré de fermes prélevées dans chaque petite région variait de 4.6 dans la région 9 à 27.5 dans la région 6, la moyenne étant de 17.3. Le nombre espéré de fermes prélevées comptant des veaux et bovins variait de 3.6 dans la région 9 à 18.8 dans la région 3, et la moyenne pour l'ensemble des petites régions était de 11.7. Un total de 30,000 simulations ont été exécutées. A chaque simulation, les échantillons étaient tirés indépendamment pour chaque point dans le temps, au moyen d'un échantillonnage aléatoire simple stratifié sans remplacement. Les 30,000 simulations ont été réalisées sous forme de 15,000 ensembles de 2 simulations, chaque ensemble correspondant à un vecteur différent de tailles d'échantillon réalisées dans les douze petites régions formant chaque strate. Il fallait procéder de la sorte pour calculer certaines mesures d'évaluation conditionnelle, décrites dans la sous-section qui suit; voir aussi Särndal et Hidiroglou (1989).

4.2 Mesures d'évaluation

Supposons que nous effectuions m simulations dans lesquelles m_1 ensembles de vecteurs différents de tailles d'échantillon réalisées dans les domaines (h,k) sont répétés m_2 fois. Les mesures suivantes peuvent servir à comparer la performance de divers estimateurs au temps T. Supposons que i varie de 1 à m_1 et que j varie de 1 à m_2.

(i) Biais relatif absolu pour la région k:

$$\mathrm{BRA}_k = \left| m^{-1} \sum_i \sum_j (\mathrm{est}_{ijk} - \mathrm{vraie}_k)/\mathrm{vraie}_k \right|. \quad (4.1)$$

La moyenne des BRA_k pour l'ensemble des régions k sera désignée par MBRA. Nous prenons le biais relatif absolu puisque nous nous intéressons principalement dans cette étude à une mesure globale comme MBRA; toutefois, dans d'autres contextes, les biais réels des petites régions individuelles peuvent aussi être d'un intérêt considérable.

La mesure suivante est justifiée par le désir d'évaluer la performance conditionnelle des estimateurs étant donné les vecteurs des tailles d'échantillon réalisées dans les domaines. Il est d'usage de mesurer la performance avec, comme condition, les tailles d'échantillon fixes des domaines; ici, nous considérons l'écart-type du biais conditionnel, B_{ik}, comme une simple mesure sommaire. Si cet écart-type est faible, la méthode est robuste par rapport aux variations des tailles d'échantillon réalisées. Notons que la valeur espérée de B_{ik} est simplement le biais inconditionnel, estimé par BRA_k. Désignons par B_k^2 la valeur espérée inconditionnelle de B_{ik}^2. Nous définissons la mesure de Monte Carlo suivante:

(ii) Écart-type du biais relatif conditionnel pour la région k:

$$\mathrm{ETBRC}_k = \left\{ m_1^{-1} \sum_i (\hat{B}_{ik}^2 - \hat{C}_{ik}) \Big/ \mathrm{vraie}_k - \mathrm{BRA}_k^2 \right\}^{1/2};$$

$$\hat{B}_{ik} = m_2^{-1} \sum_j \mathrm{est}_{ijk} - \mathrm{vraie}_k, \quad (4.2)$$

$$\hat{C}_{ik} = m_2^{-1}(m_2 - 1)^{-1} \left(\sum_j \mathrm{est}_{ijk}^2 - \left(\sum_j \mathrm{est}_{ijk} \right)^2 \Big/ m_2 \right).$$

Le terme de correction $\hat{C}_{ik}$ compense le biais de $\hat{B}_{ik}^2$ en tant qu'estimation de B_{ik}^2, du fait que m_2 est fini. $\hat{B}_{ik}^2 - \hat{C}_{ik}$ est conditionnellement sans biais pour B_{ik}^2; il est aussi inconditionnellement sans biais pour B_k^2. La moyenne de Monte Carlo $m_1^{-1} \sum_i (\hat{B}_{ik}^2 - \hat{C}_{ik})$ converge vers B_k^2 avec probabilité 1 quand $m_1 \to \infty$. $\hat{B}_{ik}^2 - \hat{C}_{ik}$ peut être négatif pour certains i, du fait que m_2 est fini. Si m_1 est grand, la moyenne sur les i est généralement très voisine de B_k^2; si la moyenne est inférieure à BRA_k^2, nous posons ETBRC_k égal à 0. METBRC désignera la moyenne des ETBRC_k pour l'ensemble des régions k.

(iii) L'erreur relative absolue moyenne pour la région k est

$$\mathrm{ERAM}_k = m^{-1} \sum_i \sum_j | \mathrm{est}_{ijk} - \mathrm{vraie}_k | / \mathrm{vraie}_k \quad (4.3)$$

et MERAM désigne la moyenne des ERAM_k pour l'ensemble des régions.

(iv) L'erreur quadratique moyenne pour la région k est:

$$\mathrm{EQM}_k = m^{-1} \sum_i \sum_j (\mathrm{est}_{ijk} - \mathrm{vraie}_k)^2 \quad (4.4)$$

et MEQM, de la même façon, désigne la moyenne pour l'ensemble des régions.

(v) La racine carrée de l'erreur quadratique moyenne relative pour la région k est

$$\mathrm{REQMR}_k = \{\mathrm{EQM}_k\}^{1/2} / \mathrm{vraie}_k \quad (4.5)$$

et, de la même façon, la moyenne pour l'ensemble des régions est dénotée par MREQMR.

La précision (c.-à-d. l'erreur-type de Monte Carlo) de chaque mesure dépend de m_1, m_2. Pour toutes les mesures sauf (ii), le choix optimal de m_1, m_2, avec comme restriction $m_2 > 1$ est $m_1 = m/2$, $m_2 = 2$, car ce choix réduit au minimum l'erreur-type de Monte Carlo. Pour s'en convaincre, posons A comme la moyenne d'une mesure d'évaluation fondée sur m_2 échantillons ayant tous la

même configuration d'échantillon (ensemble des tailles des échantillons aléatoires des domaines), que nous appellerons C. La valeur espérée de A étant donné C est une fonction de C, disons $E(C)$, et la variance conditionnelle de A est proportionnelle à m_2^{-1}, disons $V(C)/m_2$. La variance inconditionnelle de A est alors $V\{E(C)\} + E\{V(C)\}/m_2$, et la variance globale de Monte Carlo d'une mesure d'évaluation fondée sur m_1 configurations d'échantillon répétées m_2 fois est $V\{E(C)\}/m_1 + E\{V(C)\}/m_1m_2$, dont on trouve le minimum, du fait que $m = m_1m_2$ est fixe, en prenant m_1 aussi grand que possible. Pour la deuxième mesure, le choix approprié de m_1, m_2 est moins direct. Dans l'étude de simulation, nous avons fixé m à 30,000 et les valeurs correspondantes de m_1, m_2 étaient de 15,000 et 2.

4.3 Estimateurs utilisés dans l'étude comparative

L'étude a porté sur neuf estimateurs, c'est-à-dire g_1 à g_8, ainsi que g_7^*, et tous ont été évalués au temps $T = 10$. Nous avons utilisé un modèle de régression linéaire simple pour la composante synthétique, avec variable auxiliaire ainsi définie

$$X_{kt} = (\hat{\Theta}_t/\Theta_1)\Theta_{k1}, \tag{4.6}$$

où Θ_{k1}, Θ_1 désignent respectivement les totaux de la population pour la petite région k et la province au temps $t = 1$, c.-à-d. au moment du recensement de 1986. L'estimateur $\hat{\Theta}_t$ dénote l'estimateur de stratification a posteriori de Θ_t d'après les données de l'enquête sur les fermes au temps t, à l'échelle de la province. Ainsi, X_{kt} est simplement une variable synthétique de rajustement par quotient. Les variances des composantes de l'erreur dans le modèle de régression ont été supposées constantes pour l'ensemble des régions. Pour les modèles chronologiques, nous avons supposé que la dépendance sériale était conforme à une marche aléatoire. Des hypothèses de ce genre ont été utilisées avec succès dans de nombreuses applications, et nous les avons choisies principalement pour favoriser la simplicité. Nous espérions, toutefois, que les modèles choisis conviennent bien au but recherché et puissent illustrer les variations entre les gains obtenus avec différents types d'estimateurs pour petites régions assistés par un modèle, reposant aussi bien sur des méthodes de lissage transversal que sur des méthodes de lissage chronologique.

Puisque les données du recensement de 1986 ont été incluses dans la série chronologique, l'estimation directe g_{21} correspond aux résultats du recensement de 1986, et donc l'erreur de l'enquête $\underset{\sim}{\epsilon}_1$ correspondrait à $\underset{\sim}{0}$. En outre, d'après la définition de X_{kt}, un choix raisonnable pour (β_{11}, β_{21}) serait (0,1), de sorte que $\underset{\sim}{a}_1$ doit être $\underset{\sim}{0}$. Ainsi, les matrices de covariance B_t et W_t au temps $t = 1$ sont nulles, de sorte que la distribution de $\underset{\sim}{\alpha}_t$ à $t = 1$ n'exige pas d'estimation. Cette modification dans la distribution initiale de $\underset{\sim}{\alpha}_t$ est naturelle, compte tenu de l'information additionnelle fournie par le recensement. Par ailleurs, puisque les estimations directes g_{2t} n'étaient pas disponibles pour $t = 2, 3, 4$, les équations d'estimation des composantes de variance du modèle de l'annexe A.2 ont été modifiées en conséquence.

Pour la méthode 7 (cas 1), on a supposé que $\underset{\sim}{\beta}_t$ avait une valeur fixe commune seulement pour $t \geq 2$, car à $t = 1$, $\underset{\sim}{\beta}_t = (0,1)'$. Pour l'estimateur lié à la taille de l'échantillon g_5, le paramètre λ a été considéré égal à 1.

4.4 Résultats empiriques

Les principales conclusions ont été énoncées à la section 1. Nous présentons ici quelques comparaisons détaillées, accompagnées de certaines explications possibles. Nous n'indiquons pas de résultats distincts pour g_7^*, dont les résultats sont un peu inférieurs, quoique dans l'ensemble semblables, à ceux de g_7. Ces estimateurs sont présentés au tableau 1. Les figures 1 à 3, ainsi que les tableaux 2 à 4, présentent certains résultats empiriques. Nous n'avons pas indiqué les erreurs-types de Monte Carlo, mais toutes étaient largement négligeables.

Tableau 1

Sommaire des estimateurs

g_1 - Expansion	g_6 - MPLSE-I à données chronologiques, les β évoluent dans le temps, les a sont indépendants dans le temps
g_2 - Stratification a posteriori	
g_3 - Synthétique	g_7 - MPLSE-II à données chronologiques, les a évoluent dans le temps, β commun et fixe
g_4 - Fay-Herriot	
g_5 - Lié à la taille de l'échantillon	g_8 - MPLSE-III à données chronologiques, les β et les a évoluent dans le temps

Le tableau 2 présente les moyennes pour l'ensemble des petites régions des cinq mesures d'évaluation, et la figure 1 présente sous forme graphique les mesures d'évaluation moyennes par rapport à la valeur pour l'estimateur de Fay-Herriot (g_4). Certaines tendances ressortent clairement des résultats des diverses mesures selon les différents estimateurs. L'estimateur direct g_2 donne un très bon résultat dans le cas de la mesure du biais (MBRA), mais un résultat plutôt médiocre dans le cas des autres mesures. La méthode de lissage transversal g_3 (synthétique) est très peu performante en ce qui a trait aux mesures du biais. La méthode de Fay-Herriot g_4 se révèle un peu meilleure en moyenne que la méthode de stratification a posteriori en ce qui touche la mesure EQM, mais est bien inférieure dans le cas du biais. La méthode liée à la taille de l'échantillon g_5 donne des résultats très semblables à ceux de g_2; elle se révèle un peu inférieure dans le cas des mesures du biais et un peu supérieure dans le cas des autres mesures. Les méthodes avec données chronologiques g_7 et g_8 ont une très bonne performance dans l'ensemble, bien qu'elles soient un peu inférieures à g_2 dans le cas du biais. La performance de l'estimateur à données chronologiques g_6

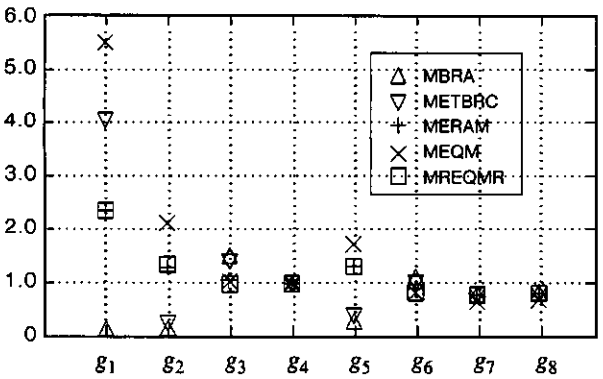


Figure 1. Mesures d'évalution par rapport à l'estimateur de Fay-Herriot

Nota: METBRC relatif pour g_1 (= 18.98) non indiqué.

Tableau 2

Moyennes des mesures d'évaluation

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
MBRA	.001	.007	.097	.065	.018	.070	.053	.053
METBRC	.282	.016	.016	.015	.023	.010	.010	.010
MERAM	.269	.147	.115	.108	.136	.097	.087	.088
MREQMR	.339	.192	.137	.137	.176	.120	.109	.111
MEQM (000)	72,979	27,596	13,382	12,898	22,760	10,603	8,610	8,829

se situe généralement entre celle de l'estimateur de Fay-Herriot et celles des estimateurs à données chronologiques g_7 et g_8. Pour tous les estimateurs (y compris l'estimateur synthétique g_3), l'écart-type du biais relatif conditionnel (METBRC) est appréciable; toutefois, il est le moins élevé dans le cas des méthodes avec données chronologiques. Comme prévu, l'estimateur d'expansion g_1 donne de bons résultats dans le cas de la mesure du biais inconditionnel (MBRA), mais est très peu performant dans le cas de la mesure du biais conditionnel (METBRC).

La figure 2 présente les moyennes des mesures REQMR_k pour trois groupes de taille, soit les petites régions de faible, moyenne et grande taille, d'après le classement de leurs totaux réels pour la population au temps T. Ces trois regroupements ont été faits parce qu'on s'attendait à ce que les erreurs relatives d'estimation soient supérieures pour les totaux les plus petits, ce que le graphique ne contredit pas. Encore une fois, les méthodes avec données chronologiques g_7 et g_8 donnent les meilleurs résultats. Notons que la méthode avec données chronologiques g_6, qui suppose que les effets de petite région sont indépendants dans le temps, n'est pas aussi performante. Le tableau 3 présente les mesures REQMR_k sans calcul de moyennes. La valeur REQMR_9 est relativement élevée du fait que le nombre total de veaux et bovins dans la région 9 est inférieur à la moitié de celui de toute autre petite région. Parmi les petites régions de taille moyenne, les régions 6 et 9 se distinguent en étant celles où l'estimation par les méthodes de lissage est la plus difficile. Cela vient du fait que, par rapport à la baisse globale d'environ 16% du nombre total de veaux et bovins dans la pseudo-population entre juin 1986 et janvier 1991, les régions 6 et 8 sont celles qui se situent le plus loin de la moyenne, à 33% et à 1% respectivement, de sorte que la covariable de rajustement par quotient est moins appropriée pour ces régions. Néanmoins, les méthodes avec données chronologiques g_7 et g_8 ont une performance sensiblement meilleure que celle de l'estimateur de stratification a posteriori pour les régions 6 et 8. Cela s'explique par le fait que le modèle de la marche aléatoire, appliqué aux effets de petite région, peut rendre compte de petites régions qui, comme les régions 6 et 8, dévient progressivement du modèle.

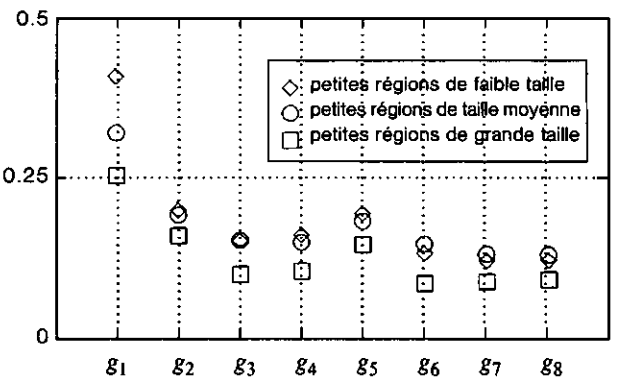


Figure 2. Racines carrées des erreurs quadratiques moyennes relatives: moyennes pour des groupes de tailles

Tableau 3

Racines carrées des erreurs quadratiques moyennes relatives et totaux vrais de veaux et bovins pour de petites régions

	Région	Valeurs vraies	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
Petite taille	9	8,502	.580	.277	.342	.275	.277	.199	.160	.174
	10	18,990	.360	.196	.078	.113	.175	.097	.103	.104
	11	18,776	.339	.122	.122	.103	.112	.096	.086	.087
	12	19,819	.409	.237	.076	.152	.212	.123	.117	.117
	Moyenne	16,522	.422	.208	.154	.161	.194	.129	.116	.120
Taille moyenne	1	27,595	.312	.206	.117	.130	.185	.120	.100	.102
	6	29,012	.306	.241	.256	.216	.224	.224	.168	.172
	7	23,600	.341	.121	.107	.094	.110	.088	.092	.092
	8	23,627	.383	.250	.155	.165	.219	.155	.146	.144
	Moyenne	25,959	.336	.205	.159	.151	.185	.147	.126	.127
Grande taille	2	35,592	.268	.171	.113	.110	.156	.096	.089	.088
	3	40,582	.241	.151	.087	.090	.137	.070	.072	.073
	4	42,396	.256	.160	.099	.103	.144	.080	.088	.089
	5	35,996	.270	.176	.091	.097	.160	.088	.085	.088
	Moyenne	38,642	.259	.164	.098	.100	.149	.083	.083	.084

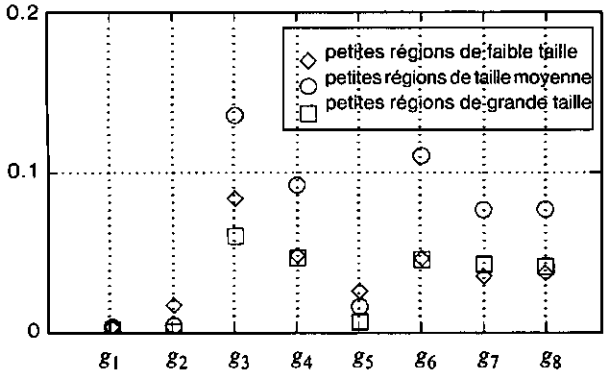


Figure 3. Biais relatifs absolus: moyennes pour des groupes de tailles

Tableau 4

Biais relatifs absolus et totaux vrais de veaux et bovins pour de petites régions

	Région	Valeurs vraies	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
Petite taille	9	8,502	.002	.047	.232	.139	.085	.099	.061	.069
	10	18,990	.002	.002	.006	.007	.003	.015	.026	.025
	11	18,776	.002	.009	.090	.052	.021	.062	.039	.037
	12	19,819	.000	.007	.019	.011	.007	.023	.024	.023
	Moyenne	16,522	.001	.016	.087	.052	.029	.050	.037	.039
Taille moyenne	1	27,595	.001	.003	.093	.063	.007	.078	.044	.045
	6	29,012	.000	.001	.239	.157	.023	.195	.120	.123
	7	23,600	.000	.005	.088	.053	.014	.058	.062	.061
	8	23,627	.002	.008	.143	.106	.024	.124	.093	.091
	Moyenne	25,959	.001	.004	.141	.095	.017	.114	.080	.080
Grande taille	2	35,592	.000	.000	.095	.071	.009	.068	.049	.047
	3	40,582	.000	.001	.047	.041	.005	.029	.026	.025
	4	42,396	.001	.002	.066	.056	.008	.044	.057	.056
	5	35,996	.000	.000	.045	.029	.005	.048	.035	.039
	Moyenne	38,642	.000	.001	.063	.049	.006	.047	.042	.042

La figure 3 et le tableau 4 sont de format identique à la figure 2 et au tableau 3, mais présentent les biais relatifs plutôt que les racines carrées des erreurs quadratiques moyennes relatives. Tant pour l'estimateur d'expansion g_1 que pour l'estimateur de stratification a posteriori g_2, les biais sont négligeables. Dans le cas des méthodes de lissage, les biais relatifs absolus moyens pour les petites régions de taille moyenne sont relativement élevés, et sont principalement attribuables aux régions 6 et 8, pour lesquelles la covariable est moins appropriée. Parmi les méthodes de lissage, la méthode liée à la taille de l'échantillon g_5 présente le biais le plus faible, car elle donne généralement des résultats très voisins de ceux de l'estimateur direct g_2; toutefois, elle n'est que très peu supérieure à g_2 dans le cas de l'erreur quadratique moyenne. Parmi les autres méthodes de lissage, les estimateurs avec données chronologiques g_7 et g_8, qui avaient l'erreur quadratique moyenne la plus faible, présentent également le biais le plus faible. Néanmoins, le biais relatif de ces méthodes peut être très élevé, comme dans les régions 6 et 8. Il serait impossible en pratique d'estimer ces biais; toutefois, la taille possible du biais pourrait être évaluée au moyen d'un échantillonnage simulé fait à partir de diverses populations plausibles.

5. CONCLUSION

Une étude de simulation nous a permis de constater que les méthodes d'estimation pour petites régions reposant à la fois sur des données transversales et des données chronologiques peuvent être plus performantes que celles fondées uniquement sur des données transversales, tant du point de vue du biais que de celui de l'erreur quadratique moyenne. Toutefois, le coût en termes de biais pourrait encore être élevé. Il importe évidemment de se demander s'il est possible en pratique de déterminer si les gains découlant de toute forme de lissage l'emporteront sur les coûts, et comment faire cette évaluation.

Les modèles visés par l'étude de simulation ont été choisis à partir de considérations générales. Toutefois, en pratique, des diagnostics appropriés semblables à ceux employés dans Pfeffermann et Barnard (1991) devraient être mis au point pour les données d'enquête avant la recommandation de toute méthode assistée par un modèle. Notons également que les estimateurs pour petites régions pourraient être modifiés et rendus robustes à l'égard d'une formulation erronée du modèle sous-jacent, comme il est suggéré dans Pfeffermann et Burck (1990); voir aussi Mantel, Singh et Bureau (1993). Enfin, il y aurait lieu d'étudier, dans de futurs travaux, la modification et l'extension des méthodes présentées dans cet article pour les adapter au cas plus réaliste d'erreurs d'échantillonnage corrélées.

REMERCIEMENTS

Nous tenons à remercier Jon Rao, Danny Pfeffermann et M.P. Singh, avec qui nous avons eu des discussions fructueuses et qui nous ont fourni des commentaires utiles relatifs à des versions précédentes de cet article. Les commentaires et suggestions d'un arbitre anonyme, ainsi que d'un rédacteur associé, ont également été très appréciés. La recherche du premier auteur a été rendue possible en partie par une subvention du Conseil de recherches en sciences naturelles et en génie du Canada accordée à l'Université Carleton.

ANNEXE

A.1 Estimation de la variance de g_{2kt}

Soit v_{kt} la variance conditionnelle (étant donné n_{hkt}) de g_{2kt} dans (2.2). On peut alors exprimer v_{kt} (lorsque $n_{hkt} > 0$ pour tout h) de la façon suivante

$$v_{kt} = \sum_h N_{hkt}^2 \left(n_{hkt}^{-1} - N_{hkt}^{-1} \right) \sigma_{hkt}^2, \qquad \text{(A.1)}$$

où σ^2_{hkt} est la variance de la population pour l'intersection de la $h$ième strate et de la $k$ième petite région. La variance σ^2_{hkt} peut être évaluée par l'estimateur habituel s^2_{hkt} pour $n_{hkt} \geq 2$. Notons que l'estimation de la variance conditionnelle v_{kt} fournit aussi une estimation de la variance inconditionnelle de g_{2kt}.

Si $n_{hkt} = 1$, nous pouvons utiliser une valeur synthétique comme estimation de σ^2_{hkt}, qu'on peut écrire sous la forme $\sum (n_{hkt} - 1)s^2_{hkt} / \sum (n_{hkt} - 1)$, la sommation étant faite sur tous les k pour lesquels $n_{hkt} \geq 2$ dans chaque strate (h,t). Si $n_{hkt} = 0$, v_{ht} dans (A.1) n'est évidemment pas défini. La valeur synthétique de $\bar{y}_{hkt}$ étant utilisée dans ce cas, il nous faut une valeur synthétique de son erreur quadratique moyenne. Pour chaque (h,t), celle-ci peut être ainsi définie

$$(\bar{X}_{hkt}/\bar{X}_{ht})^2(n_{ht}^{-1} - N_{ht}^{-1})s^2_{ht} + (\widehat{\text{biais}})^2,$$

où $(\widehat{\text{biais}})^2$ sera défini comme

$$\sum_{n_{hlt}>0} ((\bar{X}_{hlt}/\bar{X}_{ht})\bar{y}_{ht} - \bar{y}_{hlt})^2/m_{ht},$$

où m_{ht} est le nombre de petites régions avec échantillon dans la strate h.

A.2 Estimation des composantes de la variance

Selon la notation de (3.2), nous montrons ici l'application de la méthode des moments à l'estimation des composantes de la variance pour le modèle de la section 3.1, dans le cas spécial où il y a une seule variable auxiliaire X_{ht}, $Q_t = \tau^2 I$ et $\underset{\sim}{\beta}_t$ suit une marche aléatoire, c.-à-d. $G_t^{(1)} = I$. Soit $F_t = (F_{1t}, \ldots, F_{Kt})'$, $F_{kt} = (1, X_{kt})'$, $\underset{\sim}{\beta}_t = (\beta_{1t}, \beta_{2t})'$, et $B_t = \text{diag}(\gamma_1^2, \gamma_2^2)$. Le paramètre τ^2 est estimé par la solution de

$$\sum_{t=1}^{T} \sum_{k=1}^{K} (g_{2kt} - F'_{kt}\,\hat{\underset{\sim}{\beta}}_t)^2/(v_{kt} + \tau^2) = T(K-2).$$

S'il n'y a pas de solution positive, nous posons $\hat{\tau}^2 = 0$. Ici, $\hat{\underset{\sim}{\beta}}_t$ désigne l'estimation selon les moindres carrés pondérés de $\underset{\sim}{\beta}_t$, fondée uniquement sur les données transversales au temps t. Cette façon de procéder est semblable à la méthode utilisée dans Fay et Herriot (1979) pour des données transversales. On peut obtenir une estimation de γ_i^2 en résolvant (pour $i = 1, 2$)

$$\sum_{t=2}^{T} (\hat{\beta}_{i,t} - \hat{\beta}_{i,t-1})^2/(\gamma_i^2 + d_{ii}^{(t)}) = T - 1,$$

où $d_{ii}^{(t)}$ est le (i,i)ième élément de $(F'_{t-1}U_{t-1}^{-1}F_{t-1})^{-1} + (F'_t U_t^{-1} F_t)^{-1}$.

BIBLIOGRAPHIE

BATTESE, G.E., HARTER, R.M., et FULLER, W.A. (1988). An error-components model for prediction of country crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

BELL, W.R., et HILLMER, S.C. (1987). Time series methods for survey estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 83-92.

BINDER, D.A., et DICK, J.P. (1989). Enquêtes répétées modélisation et estimation. *Techniques d'enquête*, 15, 31-48.

DUNCAN, D.B., et HORN, S.D. (1972). Linear dynamic recursive estimation from the viewpoint of regression analysis. *Journal of the American Statistical Association*, 67, 815-821.

ERICKSEN, E.P. (1974). A regression method for estimating populations of local areas. *Journal of the American Statistical Association*, 69, 867-875.

FAY, R.E., et HERRIOT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY (1993). Indirect Estimators in Federal Programs. Statistical Policy Working Paper 21, U.S. Office of Management and Budget.

GHOSH, M., et RAO, J.N.K. (1994). Small Area Estimation: an Appraisal. *Statistical Science*, 9, à paraître.

GONZALEZ, M.E. (1973). Use and evaluation of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.

HARVEY, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: University Press.

JULIEN, C., et MARANDA, F. (1990). Sample design of the 1988 national farm survey. *Survey Methodology*, 16, 117-129.

MANTEL, H.J., SINGH, A.C., et BUREAU, M. (1993). Benchmarking of small area estimators. *Proceedings of the International Conference on Establishment Surveys, Buffalo, June 1993*, 920-925.

PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economics Statistics*, 9, 163-175.

PFEFFERMANN, D., et BARNARD, C.H. (1991). Some new estimators for small area means with application to the assessment of farmland values. *Journal of Business and Economics Statistics*, 9, 73-84.

PFEFFERMANN, D., et BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.

PLATEK, R., RAO, J.N.K., SÄRNDAL, C.-E., et SINGH, M.P. Éds. (1987). *Small Area Statistics: An International Symposium*; New York; John Wiley and Sons.

RAO, J.N.K., et CHOUDHRY, G.H. (1993). Small area estimation: Overview and empirical study. Présenté à l'International Conference on Establishment Surveys, Buffalo, June 1993, à paraître.

RAO, J.N.K., et YU, M. (1992). Small Area Estimation by Combining Time Series and Cross-sectional Data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1-9.

SALLAS, W.M., et HARVILLE, D.A. (1981). Best linear recursive estimation for mixed linear models. *Journal of the American Statistical Association*, 76, 860-869.

SÄRNDAL, C.-E., et HIDIROGLOU, M.A. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

SCHAIBLE, W.L. (1978). Choosing weights for composite estimation for small area statistics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 741-746.

SCOTT, A.J., et SMITH, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.

SETHI, V.K. (1963). A note on optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5, 20-33.

SINGH, A.C., et MANTEL, H.J. (1991) Estimation composite par espace d'états pour les petites régions. *Recueil: Symposium 91: Conception et analyse des enquêtes longitudinals*, Statistique Canada, Ottawa, Novembre 1991, 17-25.

TILLER, R. (1992). Time series modelling of sample survey data from the U.S. Current Population Survey. *Journal of Official Statistics*, 8, 149-166.

Techniques d'enquête, juin 1994
Vol. 20, n° 1, pp. 47–55
Statistique Canada

Méthode du jackknife pour l'estimation de la variance en présence de données imputées

J.G. KOVAR et E.J. CHEN[1]

RÉSUMÉ

L'imputation est une méthode dont se servent couramment les organismes d'enquête afin de corriger le problème posé par la non-réponse à des questions particulières. Bien que dans la plupart des cas, les ensembles de données ainsi complétés offrent de bonnes estimations des moyennes et des totaux, les variances correspondantes, souvent, sont largement sous-estimées. Plusieurs méthodes permettent de remédier à ce problème, mais la plupart dépendent du plan d'échantillonnage et de la méthode d'imputation. Récemment, Rao (1992) et Rao et Shao (1992) ont proposé une méthode jackknife unifiée pour l'estimation de la variance d'ensembles de données ayant fait l'objet d'une imputation. Le présent article évalue cette technique de manière empirique, au moyen d'une population réelle d'entreprises, et selon un plan d'échantillonnage aléatoire simple et un mécanisme de non-réponse uniforme. La possibilité d'étendre cette méthode à des plans d'échantillonnage stratifié à plusieurs degrés est examinée, et l'on se penche brièvement sur la performance de l'estimateur de la variance proposé dans le cas de mécanismes de réponse qui ne sont pas uniformes.

MOTS CLÉS: Non-réponse à des questions particulières; imputation hot-deck; imputation du plus proche voisin; non-réponse non aléatoire; plan d'échantillonnage complexe.

1. INTRODUCTION

Le problème de la non-réponse touche toutes les enquêtes par sondage, à des degrés divers. Dans le cas de la non-réponse totale d'une unité de l'échantillon, cette lacune est souvent corrigée par un rajustement approprié de la pondération de l'enquête, mais si la non-réponse ne touche que des questions particulières, la plupart des organismes d'enquête recourent à l'imputation. Ainsi, des valeurs plausibles sont insérées à la place de valeurs manquantes ou incohérentes, ce qui simplifie l'estimation des moyennes et des totaux à tous les niveaux d'agrégation. Dès les années 1950, toutefois, Hansen, Hurwitz et Madow (1953) ont reconnu que le fait de traiter les valeurs imputées comme des valeurs observées peut entraîner une sous-estimation des variances des estimateurs si les formules habituelles sont utilisées, sous-estimation qui s'amplifie à mesure que la proportion des réponses imputées s'accroît.

Plusieurs solutions ont été suggérées pour résoudre ce problème. En particulier, Rubin (1987) a proposé de recourir à l'imputation multiple pour estimer la variance attribuable à l'imputation, en répétant le processus plusieurs fois et en estimant la variation entre les divers ensembles de données ainsi obtenus. Plus récemment, Särndal (1990) a présenté un certain nombre d'estimateurs de la variance fondés sur un modèle, tandis que Rao et Shao (1992) ont proposé une technique de rajustement des valeurs imputées qui permet de corriger l'estimateur de la variance jackknife habituel (ou simpliste). Les méthodes de Särndal et de Rao et Shao sont attrayantes du fait que seul le fichier soumis à l'imputation (dans lequel les valeurs imputées sont marquées) est nécessaire à l'estimation de la variance. Aucun fichier auxiliaire n'est requis. La méthode de Särndal fondée sur un modèle donne des estimateurs sans biais de la variance pourvu que le modèle se vérifie (Lee, Rancourt et Särndal 1991). La méthode jackknife rajustée de Rao et Shao donne une estimation convergente selon le plan et non biaisée selon le modèle (Rao 1992). Toutefois, la méthode fondée sur un modèle exige des estimateurs de la variance différents pour chaque méthode d'imputation, tandis que la méthode jackknife rajustée offre une approche unifiée exigeant la mise en oeuvre d'un seul estimateur, l'estimateur jackknife, pourvu que les valeurs imputées soient convenablement rajustées à l'étape de l'estimation de la variance.

Dans le présent article, nous décrivons une étude de simulation qui permet d'évaluer l'estimateur de la variance jackknife rajusté de Rao et Shao (1992). Dans la section 2, nous faisons valoir le bien-fondé de la présente étude empirique en montrant les caractéristiques de l'estimateur de la variance simpliste selon quatre méthodes d'imputation, dans le cas d'un échantillonnage aléatoire simple. Dans la section 3, nous décrivons brièvement la méthode de rajustement de Rao et Shao et nous présentons les résultats empiriques. Des extensions à des plans plus complexes et à des expériences comportant des mécanismes de non-réponse non aléatoires sont examinées à la section 4. Enfin, la section 5 contient des conclusions et des recommandations, et propose notamment des voies futures de recherche.

[1] J.G. Kovar, Division des méthodes d'enquêtes-entreprises; E.J. Chen, Division des méthodes d'enquêtes sociales, Statistique Canada, Immeuble R.H. Coats, Parc Tunney, Ottawa, Ontario, K1A 0T6.

2. CONTEXTE

Suivant la notation de Rao (1992), nous supposons que dans un échantillon s de taille n, m unités répondent à la question y, tandis que $n - m$ unités ne le font pas. Désignons par y_i^* la valeur imputée pour l'unité i, $i \in s\text{-}s_r$, où s_r est l'ensemble des unités qui ont répondu. L'estimateur habituel de la moyenne $\bar{Y}$ basé sur le fichier soumis à l'imputation, dans le cas d'un échantillonnage aléatoire simple, est donné par

$$\bar{y}_I = \frac{1}{n}\left(\sum_{i \in s_r} y_i + \sum_{i \in s-s_r} y_i^* \right). \quad (1)$$

2.1 Méthodes d'imputation

Dans la présente étude de simulation, nous examinons quatre méthodes d'imputation simple: moyenne des répondants, quotient, plus proche voisin et hot-deck. Le lecteur est invité à lire l'article de Kalton et Kasprzyk (1986) pour un examen approfondi de la question de l'imputation. La méthode d'imputation la plus simple et la plus intuitive quand on veut estimer la moyenne des réponses à une question y consiste à imputer la moyenne des unités répondantes observées aux unités non répondantes. La valeur imputée y_i^* de l'unité i est donc la suivante, dans le cas de la méthode d'imputation de la moyenne:

$$y_i^* = \bar{y}_m = \sum_{j \in s_r} y_j / m. \quad (2)$$

L'estimateur de la moyenne de la population $\bar{Y}$ donné en (1) se réduit alors à l'estimateur $\bar{y}_I = \bar{y}_m$. Comme cette méthode a l'inconvénient de causer une distorsion des distributions, elle n'est généralement utilisée qu'en dernier recours. Nous la présentons ici à des fins d'illustration.

En second lieu, nous examinons une méthode d'imputation par quotient reposant sur l'hypothèse qu'une variable auxiliaire corrélée x est disponible et que le quotient $\bar{y}/\bar{x}$ est le même dans les ensembles s_r et $s - s_r$, comme ce serait le cas si la non-réponse survenait au hasard, par exemple. Selon cette méthode du quotient, nous imputons la valeur suivante à la place de la valeur manquante y_i:

$$y_i^* = \frac{\bar{y}_m}{\bar{x}_m} x_i, \quad (3)$$

où $\bar{x}_m$ est la moyenne des valeurs x de l'ensemble des répondants s_r. L'estimateur de la moyenne de la population $\bar{Y}$ donné en (1) se réduit à l'estimateur d'échantillonnage double $\bar{y}_I = (\bar{y}_m/\bar{x}_m)\bar{x}$, si on considère les répondants comme l'échantillon de deuxième phase.

La troisième méthode d'imputation examinée est celle du plus proche voisin (PPV). Cette méthode consiste à imputer à une donnée manquante la valeur observée d'une autre unité de l'ensemble s_r, dont la distance par rapport à l'unité non répondante est minimale. En pratique, les fonctions de distance utilisées sont habituellement les normes ζ_1, ζ_2, ou ζ_∞ de Minkowski fondées sur les variables auxiliaires x, qu'on suppose observées pour toutes les unités de s. Ainsi

$$y_i^* = y_j, j \in s_r, \quad \text{tel que} \quad \| x_i - x_j \| \quad \text{est minimisé}, \quad (4)$$

où $\| \cdot \|$ désigne l'une des normes ci-dessus.

Les trois méthodes qui viennent d'être décrites sont souvent qualifiées de déterministes, car pour un échantillon de répondants donné, les valeurs imputées sont déterminées de façon unique. La quatrième méthode d'imputation examinée dans cette étude, la méthode du hot-deck (HD), est non déterministe, puisque les valeurs imputées sont choisies au hasard dans l'échantillon de répondants. Bien qu'en pratique des classes d'imputation soient souvent créées et qu'un processus séquentiel quelconque soit généralement mis en oeuvre, nous examinons ici le hot-deck pur, dans lequel le donneur (j) est choisi au hasard, avec remplacement, dans l'ensemble s_r complet, c.-à-d.

$$y_i^* = y_j, \; j \in s_r. \quad (5)$$

2.2 Variance due à l'imputation

Si l'on traite les valeurs imputées comme des valeurs observées, on a l'estimateur incorrect de la variance

$$v_{naive} = (1 - f)s_I^2/n, \quad (6)$$

où s_I^2 est la variance de l'échantillon complet (valeurs des répondants et valeurs imputées) et $(1 - f)$ est le facteur de correction pour population finie $(f = n/N)$. Il est facile de montrer que la variance vraie de l'estimateur $\bar{y}_I$ dans (1), $V(\bar{y}_I)$, peut s'écrire ainsi (Särndal 1990):

$$V(\bar{y}_I) = V_{sam} + V_{imp} + V_{mix}, \quad (7)$$

où V_{sam} est la variance d'échantillonnage, V_{imp} est la variance introduite par la méthode d'imputation en cause et V_{mix} est un terme de covariance entre V_{sam} et V_{imp} qui, dans la plupart des cas, est négligeable ou nul. On pourrait obtenir une estimation de V_{sam} en ajoutant à v_{naive} un terme de correction tenant compte du fait que la formule habituelle sous-estime la variance d'échantillonnage quand l'ensemble de données contient des valeurs imputées. Pour estimer $V(\bar{y}_I)$, toutefois, il est nécessaire d'estimer une composante de variance additionnelle, V_{imp}, attribuable au mécanisme d'imputation. Cela peut se faire explicitement, comme dans l'imputation multiple de Rubin (1987), ou encore en modifiant des formules de variance commune comme dans Särndal (1990) et dans Rao et Shao (1992). Notons que l'intérêt réside dans l'estimation de la variance de l'estimateur en cause, soit $V(\bar{y}_I)$, et non de la variance d'un estimateur qui aurait été obtenu en l'absence de non-réponse.

2.3 Sous-estimation de la variance

Pour montrer l'ampleur de la sous-estimation de $V(\bar{y}_I)$ par v_{naive}, et le lien entre le degré de sous-estimation et la méthode d'imputation, nous décrivons d'abord l'étude de simulation que nous avons réalisée à cette fin. Nous examinons un ensemble de données de 5,620 unités auxquelles sont associées deux variables: une variable auxiliaire x, le revenu d'exploitation brut, connu pour toutes les unités et pouvant servir de mesure de la taille, et une variable connexe relative aux achats, y. La corrélation entre x et y dans cet ensemble de données particulier est d'environ 0.92. Des échantillons aléatoires simples de taille 200 étaient prélevés sans remplacement. Une proportion fixe d'unités étaient désignées au hasard comme des non-répondants; leur valeur y était supprimée, puis imputée selon l'une des quatre méthodes décrites ci-dessus. Divers taux de non-réponse ont été produits, mais pour l'essentiel, nous ne présentons que les résultats obtenus avec les taux de non-réponse de 5% et de 30%.

Pour évaluer la performance des estimateurs de la variance proposés, nous calculons le biais relatif en pourcentage de l'estimateur de la variance $v.$, donné par

$$\text{Biais rel.}(v.) = \sum_{k=1}^{K} \frac{(v_k - V(\bar{y}_I))/K}{V(\bar{y}_I)} \times 100, \qquad (8)$$

où $V(\bar{y}_I)$ est obtenue par simulation, et v_k est la $k$ième réalisation des K estimations de la variance simulées en question. De même, la stabilité relative en pourcentage des estimateurs de la variance est donnée par

$$\text{Stab. rel.}(v.) = \sum_{k=1}^{K} \frac{\sqrt{(v_k - V(\bar{y}_I))^2/K}}{V(\bar{y}_I)} \times 100. \qquad (9)$$

Toutes les simulations ont été faites sur un PC d'IBM avec le logiciel FORTRAN 77, version 5.0, de Microsoft. Dans le cas de l'échantillonnage aléatoire simple, les résultats se fondent sur des moyennes de 100,000 répétitions (K = 100,000). Avec un tel nombre de répétitions, nous avons observé que les biais relatifs présentés ne variaient pas de plus d'un point de pourcentage. Les résultats sont résumés ci-dessous au tableau 1 pour les taux de non-réponse de 5% et de 30%.

L'examen du tableau 1 nous indique d'abord que l'estimateur simpliste sous-estime la variance vraie de $\bar{y}_I$ de 10.7% dans le cas de l'imputation de la moyenne, à un niveau de non-réponse de 5%. Cette sous-estimation est attribuable environ pour moitié au fait que v_{naive} sous-estime V_{sam}, et pour le reste au fait que v_{naive} ne tient pas compte de la composante V_{imp}. Särndal (1990) obtient des résultats très semblables pour ce qui est de la ventilation de la sous-estimation dans le cas de l'imputation de la moyenne. Nous observons ensuite, dans la première ligne du tableau 1, que la variance vraie de $\bar{y}_I$ est plus élevée dans le cas de l'imputation hot-deck que dans celui de l'imputation de la moyenne, en raison de la variabilité

Tableau 1

Sous-estimation de la variance de $\bar{y}_I$ par l'estimateur simpliste (v_{naive}) selon quatre méthodes d'imputation, à des taux de non-réponse de 5% et de 30%

Taux de non-réponse	Estimateur de la variance	Méthode d'imputation			
		Moyenne	HD	Quotient	PPV
5%	$V(\bar{y}_I)$	9.9	10.3	9.5	9.5
	v_{naive}	8.9	9.4	9.2	9.3
	Biais rel. (v_{naive})	−10.7%	−9.4%	−2.5%	−2.2%
30%	$V(\bar{y}_I)$	13.5	16.5	10.1	10.3
	v_{naive}	6.5	9.4	8.5	9.0
	Biais rel. (v_{naive})	−51.4%	−43.4%	−15.3%	−12.8%

inhérente au hot-deck (c.-à-d. que la composante V_{imp} est plus grande). Par contre, $V(\bar{y}_I)$ est légèrement plus faible dans le cas des méthodes du quotient et du plus proche voisin, puisque V_{imp} diminue avec la capacité de la méthode d'imputation d'estimer les valeurs manquantes vraies (Särndal 1990), comme c'est le cas dans la présente étude en raison de la corrélation relativement élevée entre les variables x et y. Un autre fait que révèle le tableau 1, c'est que $V(\bar{y}_I)$ augmente, tandis que v_{naive} diminue, avec l'élévation du taux de non-réponse. Ainsi, la sous-estimation de $V(\bar{y}_I)$, quand les valeurs imputées sont traitées comme des valeurs observées, s'amplifie à mesure que s'accroît la proportion de valeurs manquantes. Le problème est plus prononcé dans le cas des méthodes d'imputation de la moyenne et du hot-deck, qui n'utilisent pas d'information auxiliaire. Signalons qu'une sous-estimation de la variance de l'ordre de 50% comme celle observée ici peut donner des intervalles de confiance trop étroits d'environ 30% et amener à déclarer comme significatifs des résultats qui ne le sont pas. Il est intéressant, par ailleurs, de noter le comportement semblable des méthodes du quotient et du plus proche voisin, caractéristique dont nous nous servirons plus loin.

3. ESTIMATEUR DE LA VARIANCE JACKKNIFE

Soit $\bar{y}_I(j)$ l'estimateur imputé de $\bar{Y}$ obtenu lorsque la $j$ième unité est retirée de l'échantillon. Dans le cas de l'échantillonnage aléatoire simple, un estimateur de la variance jackknife simpliste de $\bar{y}_I$ est alors donné par

$$v_J = \frac{n-1}{n} \sum_{j=1}^{n} [\bar{y}_I(j) - \bar{y}_I]^2, \qquad (10)$$

qui se réduit à v_{naive}, comme cela a été démontré (Rao 1992).

3.1 Rajustement des valeurs imputées

Afin de produire la version "appropriée" (Rao 1990) de l'estimateur de la variance jackknife, Rao (1992) a proposé de rajuster les valeurs imputées de la façon décrite ci-dessous. Intuitivement, un tel rajustement est nécessaire dès qu'une unité répondante est retirée d'une répétition jackknife, car pour la plupart des méthodes d'imputation, toutes les valeurs imputées dépendent directement ou indirectement de la valeur observée qui a été supprimée. Cela est clair dans le cas de l'imputation de la moyenne et de l'imputation par quotient, car tous les répondants contribuent directement à la moyenne $\bar{y}_m$, mais est moins évident pour les méthodes du plus proche voisin et du hot-deck, dans lesquelles l'unité supprimée contribue au processus d'imputation uniquement en ce sens qu'elle n'est pas disponible pour être choisie comme donneur. Ainsi, dès qu'une unité répondante est supprimée, *toutes* les valeurs imputées de l'échantillon doivent être rajustées avant que l'estimateur imputé "à une suppression" de la moyenne soit calculé. Le rajustement, de toute évidence, doit être fonction de la méthode d'imputation utilisée. Dans le cas des méthodes d'imputation de la moyenne et du hot-deck, on peut montrer que le rajustement suivant est approprié (Rao 1992; Rao et Shao 1992). Soit $z_i^*(j)$ la valeur rajustée de la $i$ième unité imputée y_i^*, quand la $j$ième unité a été supprimée. Alors, $z_i^*(j)$ est donnée par

$$z_i^*(j) = \begin{cases} y_i^* + [\bar{y}_m(j) - \bar{y}_m] & \text{si } j \in s_r \\ y_i^* & \text{si } j \in s\text{-}s_r. \end{cases} \quad (11)$$

Autrement dit, aucun rajustement n'est nécessaire si l'unité supprimée (j) a elle-même été imputée, c'est-à-dire si l'unité j est un non-répondant. Dans le cas de l'imputation de la moyenne, par exemple, quand $j \in s_r$, la valeur rajustée se réduit à $\bar{y}_m(j)$, la moyenne des $m - 1$ répondants restants, comme souhaité.

On évalue l'estimateur de la variance jackknife d'abord en calculant l'estimateur imputé rajusté $\bar{y}_I^a(j)$ suivant

$$\bar{y}_I^a(j) = \sum_{\substack{i \in s \\ i \neq j}} z_i^*(j)/(n-1), \quad (12)$$

puis en posant

$$v_J(\bar{y}_I) = \frac{n-1}{n} \sum_{j=1}^{n} [\bar{y}_I^a(j) - \bar{y}_I]^2. \quad (13)$$

Il peut être démontré que l'estimateur de la variance jackknife rajusté se réduit à l'estimateur de la variance approprié dans le cas de l'imputation de la moyenne (Rao 1990) et offre une estimation convergente dans le cas de l'imputation hot-deck (Rao et Shao 1992).

Pour ce qui est de l'imputation par quotient, les valeurs rajustées sont données par

$$z_i^*(j) = \begin{cases} y_i^* + \left[\dfrac{\bar{y}_m(j)}{\bar{x}_m(j)} x_i - \dfrac{\bar{y}_m}{\bar{x}_m} x_i\right] & \text{si } j \in s_r \\ y_i^* & \text{si } j \in s\text{-}s_r, \end{cases} \quad (14)$$

où $\bar{x}_m(j)$ est la moyenne des $m - 1$ valeurs de x pour les unités répondantes lorsque l'unité j est supprimée. L'estimateur de la variance jackknife $v_J(\bar{y}_I)$ est alors calculé comme en (13) ci-dessus, ce qui donne l'estimateur de la variance approprié. En outre, Rao (1992) montre non seulement que l'estimateur de la variance jackknife rajusté est convergent selon le plan (p-convergent) dans le cas d'une non-réponse uniforme et indépendamment du modèle, mais aussi qu'il est non biaisé selon le plan et le modèle (pm-non biaisé), en vertu du modèle (15) et de tout mécanisme de non-réponse qui ne dépend pas des valeurs y.

$$E_m(y_i) = \beta x_i, \quad V_m(y_i) = \sigma^2 x_i, \quad \operatorname{cov}_m(y_i, y_j) = 0 \quad i \neq j \in s. \quad (15)$$

Puisque l'estimateur de la variance simpliste en vertu de l'imputation du plus proche voisin a affiché un comportement très semblable à celui de l'estimateur de la variance simpliste en vertu de l'imputation par quotient, le rajustement relatif à l'imputation par quotient donné en (14) a été appliqué à l'imputation du plus proche voisin. Nous avons aussi examiné un autre rajustement, qui consistait à faire une nouvelle imputation de l'unité i au moyen de la méthode du plus proche voisin, lorsque l'unité supprimée (j) avait servi à l'imputation de l'unité i. Autrement dit, le rajustement a lieu seulement si l'unité supprimée est un répondant (comme ci-dessus), et seulement les non-répondants de la $j$ième répétition jackknife à qui l'unité j a été imputée font l'objet d'une nouvelle imputation pour recevoir la valeur de l'un des $m - 1$ donneurs restants. (Cela équivaut à l'imputation du deuxième voisin le plus proche pour ces unités.) Il convient de signaler qu'il n'existe pas de justification théorique pour l'un ou l'autre de ces rajustements. Puisque le dernier rajustement a donné un rendement inférieur à celui du rajustement de la méthode du quotient dans nos exemples, et qu'il serait lourd à appliquer en pratique, nous ne l'avons pas analysé davantage, bien qu'il ait toujours produit des estimations prudentes.

Il importe de souligner que pour toutes les méthodes d'imputation, les rajustements ne servent qu'à l'estimation de la variance et peuvent par conséquent n'être appliqués que temporairement, au moment de l'exécution du programme d'estimation de la variance. Aucun rajustement permanent du fichier imputé n'est nécessaire pour l'estimation des moyennes et des totaux, bien qu'il faille marquer de façon appropriée les champs ayant reçu une valeur imputée.

3.2 Résultats empiriques

L'estimateur de la variance jackknife, avec rajustements correspondant aux quatre méthodes d'estimation décrites ci-dessus, a été calculé en sus de v_{naive} dans l'étude de simulation décrite à la section 2. Des taux de non-réponse

Tableau 2

Biais relatifs de l'estimateur de la variance simpliste (v_{naive}) et de l'estimateur de la variance jackknife rajusté, à des taux de non-réponse de 5% et de 30%

Taux de non-réponse	Estimateur de la variance	Méthode d'imputation			
		Moyenne	HD	Quotient	PPV
		en pourcentage			
5%	v_{naive}	−10.7	−9.4	−2.5	−2.2
	v_J	2.7	3.6	3.4	3.7
30%	v_{naive}	−51.4	−43.4	−15.3	−12.8
	v_J	3.3	1.9	3.0	5.3

de 5% et de 30% ont été utilisés et les biais relatifs ont été calculés. Les résultats sont résumés dans le tableau 2.

Puisque l'estimateur de la variance jackknife rajusté est convergent selon le plan (p-convergent) (Rao 1992), il donne de bons résultats, comme prévu, dans le cas de l'imputation de la moyenne, de l'imputation hot-deck et de l'imputation par quotient pour une structure de réponse uniforme. (Nous avons observé d'aussi bons résultats avec d'autres ensembles de données qui ne suivent pas le modèle (15), mais il faudrait approfondir davantage l'analyse dans leur cas.) Il convient de signaler la performance relativement bonne obtenue dans le cas de l'imputation du plus proche voisin. L'estimateur proposé semble être quelque peu prudent, ce qui s'explique, pour une faible part, par le fait qu'il n'intègre pas la correction pour population finie.

4. EXTENSIONS

Bien que l'estimateur de la variance jackknife rajusté ait donné de bons résultats dans le cas de l'échantillonnage aléatoire simple, pour un mécanisme de non-réponse uniforme et une seule classe d'imputation, nous examinons ici des extensions possibles à un plan plus complexe, à plus d'une classe d'imputation et à des mécanismes de réponse non aléatoires.

4.1 Plans complexes

Dans la présente section, nous décrivons une étude de simulation qui évalue l'estimateur de la variance jackknife rajusté de Rao et Shao (1992) par rapport à l'estimateur de la variance simpliste, dans le cas d'un échantillonnage stratifié à plusieurs degrés et d'une imputation hot-deck. Plus précisément, nous utiliserons des données de l'enquête canadienne sur les finances des consommateurs (EFC), dont le plan de sondage est identique à celui de l'enquête sur la population active. La variable à l'étude, y, est le revenu total du ménage. L'EFC se fonde sur un plan complexe à plusieurs degrés avec stratification, et les unités primaires d'échantillonnage (upé) des strates utilisées dans la présente étude sont sélectionnées avec probabilité proportionnelle au nombre de logements. De façon générale, les upé sont des ensembles de logements, plus précisément des îlots urbains dans les villes et des groupes de secteurs de dénombrement (SD) du recensement dans les régions rurales. Nous avons pris comme population un échantillon de 3,870 ménages appartenant à 30 strates, et nous avons prélevé deux upé dans chaque strate. Comme dans notre étude sur l'échantillonnage aléatoire simple, des taux de non-réponse uniformes de 5% et de 30% ont été simulés, au niveau des ménages. Les valeurs manquantes ont ensuite fait l'objet d'une imputation selon la méthode du hot-deck décrite dans Rao et Shao (1992). En bref, la méthode d'imputation consiste à sélectionner les donneurs dans l'ensemble des répondants, avec remplacement, et avec probabilité proportionnelle au poids de l'enquête des donneurs.

Nous examinons d'abord le cas où il n'y a qu'une seule classe d'imputation. Soit y_{hik} la valeur observée pour la $k$ième unité de la $i$ième upé et de la $h$ième strate ($k = 1, \ldots, n_{hi}, i = 1, \ldots, n_h, h = 1, \ldots, L, n = \sum \sum n_{hi}$) et soit y^*_{hik} la valeur imputée correspondante quand l'unité (hik) est un non-répondant, c'est-à-dire quand $(hik) \in s\text{-}s_r$. L'estimateur imputé de Y est alors donné par

$$\hat{Y}_I = \sum_{(hik)\in s_r} w_{hik} y_{hik} + \sum_{(hik)\in s-s_r} w_{hik} y^*_{hik}, \qquad (16)$$

où w_{hik} est le poids de l'enquête correspondant à l'unité (hik). En vertu de la méthode d'imputation hot-deck ci-dessus, $\hat{Y}_I$ est asymptotiquement sans biais (Rao et Shao 1992).

L'espérance de $\hat{Y}_I$ en vertu de la méthode du hot-deck peut s'écrire ainsi (Rao et Shao 1992):

$$E_*(\hat{Y}_I) = \left[\sum_{(hik)\in s_r} w_{hik} y_{hik} \Big/ \sum_{(hik)\in s_r} w_{hik}\right] \times \sum_{(hik)\in s} w_{hik}$$

$$= [\hat{S}/\hat{T}] \times \hat{U}, \qquad (17)$$

notation introduisant les termes $\hat{S}$, $\hat{T}$ et $\hat{U}$. Les valeurs jackknife "à une suppression" sont alors données par

$$\hat{S}(gj) = \sum_{\substack{(hik)\in s_r \\ h\neq g}} w_{hik} y_{hik} + \frac{n_g}{n_g - 1} \sum_{\substack{(gik)\in s_r \\ i\neq j}} w_{gik} y_{gik},$$

$$\hat{T}(gj) = \sum_{\substack{(hik)\in s_r \\ h\neq g}} w_{hik} + \frac{n_g}{n_g - 1} \sum_{\substack{(gik)\in s_r \\ i\neq j}} w_{gik}, \qquad (18)$$

lorsque la $j$ième upé de la $g$ième strate est supprimée. On effectue le rajustement des valeurs imputées quand la (gj)ième upé est supprimée, $(hi) \neq (gj)$, et $(hik) \in s\text{-}s_r$, en posant

$$z^{(gj)}_{hik} = y^*_{hik} + \left[\frac{\hat{S}(gj)}{\hat{T}(gj)} - \frac{\hat{S}}{\hat{T}}\right]. \qquad (19)$$

Ensuite, comme dans les équations (12) et (13), on évalue l'estimateur de la variance jackknife en calculant d'abord, de la manière suivante, l'estimateur imputé rajusté $\hat{Y}^a_I$ quand la (gj)ième upé est supprimée:

$$\hat{Y}_I^q(gj) = \hat{S}(gj) + \sum_{(hik)\in s\text{-}s_r} w_{hik}\, z_{hik}^{(gj)} + \frac{n_g}{n_g - 1} \sum_{\substack{(hik)\in s\text{-}s_r \\ i\neq j}} w_{gik}\, z_{gik}^{(gj)}, \quad (20)$$

puis en posant

$$v_J(\hat{Y}_I) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_I^q(gj) - \hat{Y}_I)^2. \quad (21)$$

Il peut être démontré que l'estimateur v_J défini en (21) est un estimateur convergent de la variance de $\hat{Y}_I$ (Rao et Shao 1992).

Nous avons produit 10,000 échantillons de 60 upé choisies avec probabilité proportionnelle à la taille, et nous avons soumis les ménages sélectionnés à des taux de non-réponse uniformes de 5% et de 30%. Nous avons ensuite calculé l'estimateur de la variance simpliste, ainsi que l'estimateur de la variance jackknife rajusté, v_J, donné en (21). Le biais relatif (8) et la stabilité relative (9) ont été calculés pour les deux estimateurs de la variance, et les résultats sont résumés au tableau 3.

Tableau 3

Biais relatif et stabilité relative (entre parenthèses) de l'estimateur de la variance simpliste (v_{naive}) et de l'estimateur de la variance jackknife rajusté, à des taux de non-réponse de 5% et de 30%, dans le cas d'un échantillonnage stratifié à plusieurs degrés

Estimateur de la variance	Taux de non-réponse	
	5%	30%
	en pourcentage	
v_{naive}	−10.3 (88)	−43.7 (84)
v_J	−0.9 (97)	1.2 (124)

Comme en fait foi le tableau 3, l'estimateur de la variance simpliste sous-estime la variance vraie de Y en des proportions comparables à celles observées dans le cas de l'échantillonnage aléatoire simple (tableau 2), et la sous-estimation s'aggrave avec l'augmentation du taux de non-réponse. L'estimateur de la variance jackknife rajusté, par contre, donne de bons résultats aux deux niveaux de non-réponse, au prix relativement modique d'une légère diminution de sa stabilité, en comparaison de v_{naive}.

4.2 Classes d'imputation

Selon le même plan d'échantillonnage que celui décrit à la section 4.1, nous avons aussi examiné le cas où il existe plus d'une classe d'imputation, qui reflète la situation réelle. La taille du ménage, connue pour tous les ménages de l'échantillon, a été utilisée pour former deux classes d'imputation, soit les ménages à un seul membre et les ménages à plus d'un membre. L'hypothèse sous-jacente était que la propension à répondre est différente entre ces deux classes; on a supposé, en revanche, une probabilité de réponse uniforme à l'intérieur des classes d'imputation. Deux structures de non-réponse ont été évaluées. La première suppose une non-réponse uniforme de 5% dans la classe des ménages à un seul membre et une non-réponse uniforme de 10% dans la classe des ménages à plusieurs membres, tandis que la deuxième suppose des taux de non-réponse de 25% et de 30% respectivement pour ces deux classes. L'imputation hot-deck, les rajustements des valeurs imputées et les calculs des totaux rajustés donnés en (20), $\hat{Y}_{Iv}^q(gj)$, ont été effectués indépendamment dans chaque classe d'imputation, désignée par v. Les termes $\hat{Y}_{Iv}^q(gj)$ ont ensuite été additionnés pour les deux classes d'imputation, ce qui a donné $\hat{Y}_I^q(gj)$, dont on s'est servi pour obtenir l'estimation v_J, selon l'équation (21). Les résultats sont résumés au tableau 4.

Tableau 4

Biais relatif et stabilité relative (entre parenthèses) de l'estimateur de la variance simpliste (v_{naive}) et de l'estimateur de la variance jackknife rajusté, selon deux structures de non-réponse, dans le cas d'un échantillonnage stratifié à plusieurs degrés et de deux classes d'imputation

Estimateur de la variance	Taux de non-réponse	
	5% et 10%	25% et 30%
	en pourcentage	
v_{naive}	−16.7 (87)	−40.2 (84)
v_J	−1.0 (103)	1.1 (127)

Comme le montre le tableau 4, l'estimateur de la variance jackknife rajusté v_J donne de bons résultats en vertu des deux structures de non-réponse. Ces résultats, jumelés à ceux du tableau 3, démontrent la convergence et la stabilité relativement bonne de l'estimateur de la variance jackknife rajusté, même dans le cas de taux de non-réponse élevés.

4.3 Non-réponse non aléatoire

Comme nous l'avons vu ci-dessus, l'estimateur de la variance jackknife rajusté donne de bons résultats quand la non-réponse est aléatoire à l'intérieur des classes d'imputation. Pour étudier sa robustesse à l'égard de l'hypothèse d'un mécanisme de réponse uniforme, nous avons utilisé l'ensemble de données décrit à la section 2, et produit une non-réponse de la manière décrite dans Lee, Rancourt et Särndal (1991). Plus précisément, nous avons supposé que la probabilité de non-réponse était reliée à la variable x de deux façons distinctes:

$$P_L = 1 - \exp(-c_L x), \tag{22}$$

$$P_S = \exp(-c_S x), \tag{23}$$

où les constantes c_L et c_S sont choisies de façon à donner un taux de non-réponse probable de 30%. Dans le modèle P_L donné en (22), la non-réponse est en corrélation positive avec la variable x, ce qui signifie qu'une non-réponse est plus probable dans le cas des grandes unités (L). L'inverse est vrai pour le modèle P_S donné en (23), en vertu duquel ce sont les petites unités (S) qui sont le plus susceptibles de ne pas répondre. Les méthodes d'imputation qui ne tiennent pas compte de la variable x (moyenne et hot-deck) devraient normalement produire des estimateurs de $\bar{Y}$ qui sous-estiment la moyenne vraie en vertu du modèle de non-réponse (22) et qui surestiment la moyenne vraie en vertu du modèle (23). Toutefois, les méthodes d'imputation qui utilisent la variable auxiliaire (quotient et plus proche voisin) devraient produire de meilleures estimations de la moyenne. Une simulation, dont les résultats sont présentés au tableau 5 ci-dessous, a permis de confirmer ces suppositions. Comme auparavant, nous avons utilisé 100,000 répétitions.

Tableau 5

Estimations de la moyenne $\hat{Y}$ en pourcentage de la moyenne vraie lorsque la non-réponse n'est pas aléatoire et que le taux de non-réponse probable est de 30%

Modèle de non-réponse	Méthode d'imputation			
	Moyenne	HD	Quotient	PPV
	en pourcentage			
P_L	60.4	60.4	94.7	93.5
P_S	132.7	132.7	102.0	101.4

De toute évidence, l'estimation de la variance n'offre aucun intérêt lorsque les estimateurs ponctuels eux-mêmes sont fortement biaisés, comme c'est le cas pour les méthodes de la moyenne et du hot-deck. Toutefois, dans le cas des méthodes du quotient et du plus proche voisin, pour lesquelles les estimateurs ponctuels sont supérieurs, nous avons examiné la performance de l'estimateur de la variance jackknife rajusté, ainsi que d'un estimateur proposé par Särndal (1990), qui peut s'écrire (Rao 1992):

$$v_s(\bar{y}_I) = \left(\frac{\bar{x}}{\bar{x}_m}\right)^2 \frac{1}{m(m-1)} \sum_{i \in s_r} \left(y_i - \frac{\bar{y}_m}{\bar{x}_m} x_i\right)^2$$

$$+ \left(\frac{\bar{y}_m}{\bar{x}_m}\right) \frac{2m}{n^2(m-1)} \sum_{i \in s_r} \left(y_i - \frac{\bar{y}_m}{\bar{x}_m}\right) x_i \tag{24}$$

$$+ \left(\frac{\bar{y}_m}{\bar{x}_m}\right)^2 \frac{1}{n(n-1)} \sum_{i \in s} (x_i - \bar{x})^2,$$

pourvu que le facteur de correction pour population finie soit omis et que $(n-1)/n \cong 1$ et $(m-1)/m \cong 1$. Les résultats sont résumés au tableau 6.

Tableau 6

Biais relatif de l'estimateur de la variance simpliste (v_{naive}), de l'estimateur de la variance jackknife rajusté et de l'estimateur de la variance de Särndal en vertu d'une non-réponse non aléatoire de 30%

Modèle de non-réponse	Estimateur de la variance	Méthode d'imputation	
		Quotient	PPV
		en pourcentage	
P_L	v_{naive}	−22.7	−54.6
	v_J	3.9	−37.5
	v_S	−2.6	−36.8
P_S	v_{naive}	−4.0	−0.7
	v_J	3.7	7.2
	v_S	2.8	4.5

Dans le cas de l'imputation par quotient, l'estimateur de la variance simpliste donne des résultats très différents pour les deux modèles de non-réponse (−22.7% contre −4.0%). En effet, s'il est vrai que la réduction de la taille effective de l'échantillon tend à réduire la variance dans les deux cas, il y a surreprésentation des unités de grande taille parmi les unités manquantes dans le modèle P_L, ce qui tend à accentuer cet effet, alors que dans le modèle P_S, où il y a surreprésentation des petites unités parmi les unités manquantes, cet effet tend à être partiellement compensé. Deuxième observation, l'estimateur de la variance jackknife rajusté donne de bons résultats dans le cas de l'imputation par quotient, mais est plutôt décevant dans le cas de l'imputation du plus proche voisin. Cela s'explique par le fait que le présent ensemble de données suit assez bien le modèle linéaire habituel (15) et que l'estimateur de la variance jackknife rajusté s'est révélé non biaisé selon le modèle (Rao 1992) pour l'imputation par quotient. Par ailleurs, le rajustement de la méthode du quotient n'apporte rien d'intéressant à l'imputation du plus proche voisin quand la non-réponse n'est pas uniforme. L'autre rajustement applicable à l'imputation du plus proche voisin, que nous avons décrit à la section 3, donne également de piètres résultats en termes absolus (valeurs non présentées ici), bien que les estimations soient toujours prudentes. Troisième observation, la performance de l'estimateur de Särndal, v_S, équivaut grosso modo à celle de l'estimateur jackknife rajusté, en vertu aussi bien de la méthode du quotient que de celle du plus proche voisin, et d'une non-réponse non aléatoire qui dépend seulement de x.

Lorsque le mécanisme de réponse n'est pas aléatoire, et que la propension à répondre est reliée à la variable touchée par la non-réponse (y), les estimateurs ponctuels sont eux-mêmes gravement biaisés selon les quatre méthodes d'imputation. L'estimation de la variance a donc peu à

offrir, et le véritable intérêt réside dans l'estimation de l'erreur quadratique moyenne. Autrement dit, des efforts accrus doivent être axés sur l'amélioration des estimations ponctuelles et de leurs biais. Des résultats préliminaires dans ce domaine ont été présentés par Rancourt, Lee et Särndal (1992).

5. CONCLUSION

Il est bien connu que l'estimateur de la variance habituel sous-estime la variance de l'estimation de $\bar{Y}$ en présence de valeurs imputées, si ces valeurs sont traitées comme des valeurs observées. Dans la présente étude, nous avons de nouveau mis en évidence l'importante sous-estimation que produit l'estimateur de la variance simpliste lorsque des données sont imputées. Nous avons examiné plusieurs méthodes d'imputation pour évaluer dans quelle mesure le degré de sous-estimation était lié à la méthode d'imputation. Nous avons évalué un estimateur de la variance jackknife unifié de la forme proposée par Rao et Shao (1992), lequel tient compte de la variance due au processus d'imputation. L'étude a révélé certaines propriétés intéressantes de l'estimateur proposé dans le cas aussi bien de l'échantillonnage aléatoire simple que de plans d'enquête complexes. Nos observations sont résumées dans les paragraphes qui suivent.

(1) L'ampleur de la sous-estimation de la variance est fortement liée à la fois à la capacité de la méthode d'imputation de prédire les valeurs vraies, et à sa capacité de préserver la variation naturelle des données.

(2) L'estimateur de la variance jackknife rajusté qui est proposé offre une approche unifiée pour l'estimation de la variance de données imputées, qui peut être appliquée facilement à un certain nombre de méthodes d'imputation et à des plans de complexité variable.

(3) D'un point de vue pratique, aucune modification du fichier imputé initial n'est nécessaire, de sorte que l'estimation des moyennes et des totaux n'est aucunement perturbée par la nécessité d'estimer les variances.

(4) La méthode proposée peut facilement être étendue à des plans plus complexes, à plusieurs classes d'imputation et, avec prudence, au cas d'une non-réponse non aléatoire qui dépend seulement des variables auxiliaires disponibles.

(5) L'estimateur de la variance jackknife rajusté donne de bons résultats lorsque la non-réponse est uniforme ou que le modèle linéaire habituel est vérifié, ce qui découle du fait que l'estimateur est à la fois convergent selon le plan et non biaisé selon le plan et le modèle.

(6) Dans le cas du modèle P_L, dans lequel les unités ayant des valeurs y élevées sont plus susceptibles d'être non répondantes, les trois estimateurs de la variance ont un rendement extrêmement faible.

(7) Dans le cas de la non-réponse liée à la variable y, de meilleures techniques d'imputation sont nécessaires, et les estimateurs ponctuels doivent être étudiés plus à fond. Ici, il faut plutôt s'intéresser à l'estimation de l'erreur quadratique moyenne qu'à celle de la variance.

Les enquêtes actuelles étant soumises à un degré élevé d'imputation, du moins dans certaines classes d'imputation, il est clair qu'on ne peut passer sous silence l'effet de l'imputation sur l'estimation de la variance. Une surestimation de la précision peut produire des intervalles de confiance trop étroits et amener à présenter comme significatives des données qui ne le sont pas. Si la mise en oeuvre des méthodes suggérées ci-dessus est jugée trop coûteuse dans une situation particulière, il faudrait à tout le moins effectuer des études pour évaluer l'impact de l'imputation dans certains cas représentatifs. Un facteur spécial d'accroissement de la variance pourrait alors être appliqué. Toutefois, avec l'émergence de logiciels généraux d'estimation, il semble de moins en moins justifiable de ne pas recourir à des estimateurs de la variance qui font une juste évaluation de l'effet de l'imputation.

De nombreux problèmes demeurent évidemment non résolus, et sont peut-être impossibles à résoudre. Premièrement, l'imputation du plus proche voisin doit faire l'objet de travaux théoriques beaucoup plus intenses. Les rajustements de la méthode du jackknife que nous avons appliqués à cette méthode d'imputation ont donné un rendement moindre que ceux qui ont été appliqués aux autres méthodes. Il faudra peut-être trouver des fonctions plus lisses pour remplacer la méthode du plus proche voisin. Deuxièmement, la robustesse de l'estimateur proposé doit être évaluée. Il est clair qu'un rendement satisfaisant peut être obtenu si le modèle (15) est vérifié et que la non-réponse est aléatoire. Le non-respect partiel de l'une ou l'autre de ces conditions n'a pas semblé entacher le bon rendement de l'estimateur jackknife dans notre expérience limitée, mais d'autres travaux devront être entrepris dans cette voie. La dérogation aux deux conditions simultanément n'a pas encore été étudiée. Les cas de non-réponse non aléatoire dans lesquels la propension à ne pas répondre est liée à la variable y sont encore moins bien compris, bien qu'il faille mettre l'accent, dans ce cas, sur l'estimation de l'erreur quadratique moyenne plutôt que de la variance. Troisièmement, des comparaisons devraient être faites avec des résultats d'imputation multiple. Il faut reconnaître, toutefois, que des méthodes d'imputation appropriées (Rubin 1987) doivent d'abord être établies. Notons qu'aucune des méthodes d'imputation étudiées ici n'est pas appropriée en ce qui a trait à l'imputation multiple.

Il faudrait étendre l'analyse à d'autres méthodes d'imputation et à d'autres paramètres d'intérêt. La présente étude s'est limitée à quatre méthodes d'imputation simples. En pratique, des méthodes beaucoup plus complexes sont utilisées, et elles sont souvent combinées les unes aux autres. L'impact sur l'estimation de la variance du recours à plus d'une méthode d'imputation a été étudié par Rancourt, Lee et Särndal (1993); d'autres travaux sont nécessaires. Pour ce qui est d'autres méthodes d'imputation plus

complexes, l'effet de l'ajout de résidus théoriques aux données imputées pourrait, par exemple, être étudié. Toutefois, cette technique a comme unique objet la sous-estimation de V_{sam} par v_{naive}, et ne tient pas compte de l'effet de V_{imp}. Enfin, d'autres paramètres, par exemple la médiane, et l'effet de l'imputation sur leur variance restent à étudier. Des extensions multidimensionnelles pourraient aussi être envisagées: l'estimation des corrélations, des quotients et des paramètres de régression en présence d'une imputation serait sans doute un sujet intéressant.

REMERCIEMENTS

Les auteurs tiennent à remercier M. J.N.K. Rao pour son appui et ses encouragements constants, ainsi que le rédacteur associé pour ses commentaires utiles.

BIBLIOGRAPHIE

HANSEN, M., HURWITZ, W., et MADOW, W. (1953). *Sample Survey Methods and Theory*. (Volume 2), New York: J. Wiley, 139-141.

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

LEE, H., RANCOURT, E., et SÄRNDAL, C.-E. (1991). Experiments with variance estimation from survey data with imputed values. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 690-695.

RAO, J.N.K. (1990). Variance estimation under imputation for missing data. Rapport non publié, Statistique Canada.

RAO, J.N.K. (1992). Jackknife variance estimation under imputation for missing survey data. Rapport non publié, Statistique Canada.

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

RANCOURT, E., LEE, H., et SÄRNDAL, C.-E. (1992). Bias corrections for survey estimates from data with imputed values for nonignorable nonresponse. *Proceedings 1992 Annual Research Conference*, Bureau of the Census, 523-539.

RANCOURT, E., LEE, H., et SÄRNDAL, C.-E. (1993). Variance estimation under more than one imputation method. *Proceedings of the International Conference on Establishment Surveys, American Statistical Association*, 374-379.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley.

SÄRNDAL, C.-E. (1990). Méthodes pour estimer la précision des estimations d'enquête lorsqu'il y a eu imputation. Conférencier spécial invité. *Recueil: Symposium 90, Mesure et amélioration de la qualité des données*, Statistique Canada, 337-350.

Techniques d'enquête, juin 1994
Vol. 20, n° 1, pp. 57–61
Statistique Canada

Estimation pour grappes chevauchantes lorsque la taille de la population est inconnue

D.S. TRACY et S.S. OSAHAN[1]

RÉSUMÉ

Singh (1988) propose deux méthodes d'échantillonnage en vue d'estimer la moyenne d'une population en grappes chevauchantes lorsque la taille de la population est connue. Dans cet article, nous étudions des estimateurs par quotient appliqués dans ces deux méthodes en supposant que la taille réelle de la population est inconnue, ce qui est plus conforme à la réalité des enquêtes par sondage. Nous comparons l'efficacité des estimateurs appliqués dans l'une et l'autre méthodes et nous donnons un exemple numérique.

MOTS CLÉS: Grappes chevauchantes; formation de grappes avant échantillonnage; erreur quadratique moyenne; efficacité relative.

1. INTRODUCTION

Dans l'échantillonnage par grappes, les grappes sont formées soit avant l'échantillonnage (FGAVÉ), soit après l'échantillonnage (FGAPÉ). Dans les deux cas, les grappes peuvent se chevaucher ou non. Il existe déjà beaucoup d'ouvrages de recherche sur les grappes non chevauchantes. Cependant, de nombreux cas d'échantillonnage impliquent des grappes chevauchantes. Par exemple, on peut retrouver de telles grappes dans une enquête épidémiologique régionale portant sur une maladie contagieuse comme la tuberculose, qui tend à se répandre avec la propagation du SIDA (Gifford-Jones 1993). Dans ce cas-ci, les grappes peuvent être formées de sujets infectés ou de personnes qui sont en rapport étroit avec eux et qui sont plus vulnérables à cette maladie. De même, dans une étude écologique, les grappes peuvent être formées de centrales au charbon qui émettent des hydrocarbures aromatiques polycycliques (HAP), composés doués de propriétés cancérogènes. Les grappes sont constituées en fonction de la concentration de ces gaz, et des études peuvent s'imposer dans le but de surveiller la pollution atmosphérique, qui est à l'origine d'infections pulmonaires telles la bronchite. Au sujet des grappes chevauchantes, on peut consulter les travaux effectués par Goel et Singh (1977), Agarwal et Singh (1982) et Amdekar (1985) sur certains aspects de la question. Cependant, les méthodes élaborées par ces auteurs présentent toutes des lacunes.

Il y a quelques années, Singh (1988) a élaboré un estimateur très simple pour la moyenne d'une population. Il utilise cet estimateur dans deux méthodes d'échantillonnage par grappes (selon la formule FGAVÉ) en supposant que la taille de la population est connue. La première méthode prévoit un échantillonnage des grappes avec probabilités égales tandis que la seconde prévoit un échantillonnage avec probabilités proportionnelles à la taille de la grappe. En ce qui concerne l'échantillonnage des éléments des grappes, la règle des probabilités égales s'applique pour les deux méthodes. Toutefois, supposer que la taille de la population est connue n'est pas réaliste. Si c'était vraiment le cas, on connaîtrait *a priori* tous les éléments en double de la population et il suffirait de les éliminer pour accroître l'efficacité du plan d'échantillonnage. C'est pourquoi il est nécessaire d'améliorer les estimateurs de la moyenne d'une population de Singh (1988) afin de les rendre plus utiles, car ils dépendent de la taille réelle de la population. C'est exactement le but du présent article.

Nous proposons deux méthodes d'échantillonnage par grappes (formule FGAVÉ) qui utilisent des estimateurs par quotient ordinaires de la moyenne de population; ces estimateurs ne dépendent pas de la taille réelle de la population. Comme dans Singh (1988), la première méthode consiste en un échantillonnage avec remise avec probabilités égales, tandis que la seconde consiste en un échantillonnage avec probabilités inégales. Dans les deux cas, les éléments des grappes sont prélevés suivant un plan d'échantillonnage sans remise avec probabilités égales.

La population de N unités à l'étude peut être définie comme un ensemble de K grappes chevauchantes où N_i désigne le nombre d'unités contenues dans la grappe i et $\sum_{i=1}^{K} N_i = M \geq N$, la taille réelle (inconnue) de la population, (l'égalité ne vaut que dans le cas des grappes non chevauchantes). Une unité peut appartenir à plus d'une grappe à la fois. Posons y comme la caractéristique étudiée et $\bar{Y}$, comme la moyenne de la population.

Définissons

$$Z_{ij} = Y_{ij}/F_{ij}, \quad W_{ij} = 1/F_{ij}; \quad i = 1, 2, \ldots, K, \quad \text{et} \quad j = 1, 2, \ldots, N_i$$

où Y_{ij} est la valeur de y pour l'unité j dans la grappe i et F_{ij} est le nombre de fois que cette unité est observée dans les K grappes.

[1] D.S. Tracy et S.S. Osahan, Department of Mathematics and Statistics, University of Windsor, Windsor (Ontario), N9B 3P4.

Lorsqu'on dispose de données informatiques par grappe sur les unités, il est facile de connaître la valeur des F_{ij}. En reprenant l'exemple en épidémiologie cité plus haut, supposons qu'il existe des données informatiques sur des ménages ou des particuliers, accompagnées de codes d'identification pour ces unités (par ex., numéro civique, numéro d'assurance sociale ou numéro d'assurance-maladie). Alors, par une simple instruction de programme, un expérimentateur peut savoir facilement, grâce au code d'identification, combien de fois une certaine unité se répète dans les différentes grappes. En outre, si nous avons un diagramme des grappes chevauchantes et si le critère de formation des grappes ne permet pas d'éliminer les doubles, il est possible de connaître le nombre de fois que des unités sont observées dans les différentes grappes.

Nous examinons les deux méthodes d'échantillonnage dans la section 2 et nous comparons leur efficacité dans la section 3.

2. DESCRIPTION DES DEUX MÉTHODES

Les deux méthodes proposées sont décrites dans les sections 2.1 et 2.2. Nous en faisons la comparaison dans la section 3.

2.1 Méthode A

Cette méthode comporte les étapes suivantes:

a) Prélever k grappes parmi K au moyen d'un échantillonnage aléatoire simple avec remise (ÉASAR).

b) Dans la grappe échantillonnée i de taille $N_i (i = 1, \ldots, K)$, prélever n_i unités élémentaires par échantillonnage aléatoire simple sans remise (ÉASSR).

Théorème 1. L'estimateur par quotient suivant un ÉAS

$$\bar{z}_{RS} = \hat{Y}_{RS}/\hat{N}_{RS} = \frac{K}{k}\sum_{i=1}^{k} N_i \bar{z}_i \Big/ \frac{K}{k}\sum_{i=1}^{k} N_i \bar{w}_i \qquad (1)$$

a pour biais relatif (au premier degré d'approximation)

$$BR(\bar{z}_{RS}) \doteq \frac{K}{k}\left[\left(\frac{\sigma_{bw}^2}{N^2} - \frac{\sigma_{bzw}}{NY}\right)K + \sum_{i=1}^{K} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)\left(\frac{S_{iw}^2}{N^2} - \frac{S_{izw}}{NY}\right)\right] \qquad (2)$$

où

$$\sigma_{bzw} = \sum_{i=1}^{K} (N_i\bar{Z}_i - Y/K)(N_i\bar{W}_i - N/K)/K$$

$$S_{izw} = \sum_{j=1}^{N_i} (Z_{ij} - \bar{Z}_i)(W_{ij} - \bar{W}_i)/(N_i - 1),$$

$$\bar{Z}_i = \sum_{j=1}^{N_i} Z_{ij}/N_i \quad \text{et} \quad \bar{z}_i = \sum_{j=1}^{n_i} z_{ij}/n_i,$$

et σ_{bw}^2, S_{iw}^2, $\bar{W}_i$ et $\bar{w}_i$ correspondent respectivement à σ_{bzw}, S_{izw}, $\bar{Z}_i$ et $\bar{z}_i$, z étant remplacé par w et Y, par N.

Démonstration. D'après un résultat courant, le biais relatif de l'estimateur $\bar{z}_{RS}$ (au premier degré d'approximation) est

$$BR(\bar{z}_{RS}) \doteq [V(\hat{N}_{RS})/N^2] - \mathrm{Cov}(\hat{Y}_{RS},\hat{N}_{RS})/YN. \qquad (3)$$

Posons E_2 et V_2 comme l'espérance et la variance conditionnelles relatives à un échantillon donné de grappes, et E_1 et V_1 comme l'espérance et la variance relatives à tous les échantillons de grappes. Nous avons alors

$$V(\hat{N}_{RS}) = V_1E_2(\hat{N}_{RS}) + E_1V_2(\hat{N}_{RS})$$

$$= V_1\left[\frac{K}{k}\sum_{i=1}^{k} N_iE_2(\bar{w}_i)\right] + E_1\left[\frac{K^2}{k^2}\sum_{i=1}^{k} N_i^2 V_2(\bar{w}_i)\right]$$

$$= V_1\left(\frac{K}{k}\sum_{i=1}^{k} N_i\bar{W}_i\right) + E_1\left[\frac{K^2}{k^2}\sum_{i=1}^{k} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{iw}^2\right]$$

$$= \frac{K^2}{k}\sigma_{bw}^2 + \frac{K}{k}\sum_{i=1}^{K} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{iw}^2. \qquad (4)$$

De la même manière, nous avons

$$\mathrm{Cov}(\hat{Y}_{RS},\hat{N}_{RS}) = \frac{K^2}{k}\sigma_{bzw} + \frac{K}{k}\sum_{i=1}^{K} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{izw}. \qquad (5)$$

En substituant (4) et (5) dans (3), nous obtenons l'équation (2), ce qui complète la démonstration du théorème.

Théorème 2. L'erreur quadratique moyenne (EQM) de l'estimateur $\bar{z}_{RS}$ (au premier degré d'approximation) est

$$\mathrm{EQM}(\bar{z}_{RS}) \doteq$$

$$\frac{K}{kN^2}\sum_{i=1}^{K} N_i^2\left[(\bar{Z}_i - \bar{Y}\bar{W}_i)^2 + \left(\frac{1}{n_i} - \frac{1}{N_i}\right)D_i^2\right] \quad (6)$$

où $D_i^2 = S_{iz}^2 - 2\bar{Y}S_{izw} + \bar{Y}^2S_{iw}^2$, et $S_{iz}^2 = \sum_{j=1}^{N_i}(Z_{ij} - \bar{Z}_i)^2/(N_i - 1)$.

Démonstration. Au premier degré d'approximation, nous avons

$$\mathrm{EQM}(\bar{z}_{RS}) \doteq [V(\hat{Y}_{RS}) - 2\bar{Y}\,\mathrm{Cov}(\hat{Y}_{RS},\hat{N}_{RS}) + \bar{Y}^2V(\hat{N}_{RS})]/N^2. \quad (7)$$

D'après (4), on peut exprimer $V(\hat{Y}_{RS})$ par la formule

$$V(\hat{Y}_{RS}) = \frac{K^2}{k}\sigma_{bz}^2 + \frac{K}{k}\sum_{i=1}^{K} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{iz}^2 \quad (8)$$

où $\sigma_{bz}^2 = \sum_{i=1}^{K}(N_i\bar{Z}_i - Y/K)^2/K$.

En substituant (4), (5) et (8) dans l'équation (7), nous obtenons, après simplification,

$$\mathrm{EQM}(\bar{z}_{RS}) \doteq \frac{K^2}{kN^2}(\sigma_{bz}^2 - 2\bar{Y}\sigma_{bzw} + \bar{Y}^2\sigma_{bw}^2)$$

$$+ \frac{K}{kN^2}\sum_{i=1}^{K} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)(S_{iz}^2 - 2\bar{Y}S_{izw} + \bar{Y}^2S_{iw}^2). \quad (9)$$

En substituant dans (9) les expressions pour σ_{bz}^2, σ_{bzw} et σ_{bw}^2, et en simplifiant, nous obtenons l'équation (6). Dans le théorème suivant, nous définissons un estimateur de $\mathrm{EQM}(\bar{z}_{RS})$.

Théorème 3. Un estimateur convergent de $\mathrm{EQM}(\bar{z}_{RS})$ (au premier degré d'approximation) est défini

$$\widehat{\mathrm{EQM}}(\bar{z}_{RS}) = \frac{K^2}{k\hat{N}_{RS}^2}\cdot\frac{1}{k-1}\sum_{i=1}^{k} N_i^2(\bar{z}_i - \bar{z}_{RS}\bar{w}_i)^2. \quad (10)$$

Démonstration. Nous remarquons que le premier degré d'échantillonnage consiste en un ÉASAR et que les variables aléatoires $N_i\bar{z}_i$ et N_iw_i, contenues dans l'estimateur par quotient, sont distribuées de façon indépendante mais identique. On peut donc estimer l'erreur quadratique moyenne de $\bar{z}_{RS}$ en se fondant sur le résultat bien connu selon lequel l'estimateur de la variance dans un plan à plusieurs degrés peut ne prendre en compte que le premier degré (voir Särndal, Swensson et Wretman, 1992, résultats 2.9.1 et 4.5.1).

D'après (9), un estimateur sans biais de

$$\sigma_{bz}^2 + \frac{1}{K}\sum_{i=1}^{K} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{iz}^2$$

peut être défini

$$s_{bz}^2 = \frac{1}{k-1}\sum_{i=1}^{k}\left(N_i\bar{z}_i - \sum_{i=1}^{k} N_i\bar{z}_i/k\right)^2, \quad (11)$$

et un estimateur sans biais de

$$\sigma_{bzw} + \frac{1}{K}\sum_{i=1}^{K} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{izw}$$

est

$$s_{bzw} = \frac{1}{k-1}\sum_{i=1}^{k}\left(N_i\bar{z}_i - \sum_{i=1}^{k} N_i\bar{z}_i/k\right) \times \left(N_i\bar{w}_i - \sum_{i=1}^{k} N_i\bar{w}_i/k\right). \quad (12)$$

De la même manière, un estimateur indépendant de

$$\sigma_{bw}^2 + \frac{1}{K}\sum_{i=1}^{K} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{iw}^2$$

est s_{bw}^2, défini comme en (11).

À l'aide de ces résultats, il est facile de montrer qu'un estimateur convergent de $\mathrm{EQM}(\bar{z}_{RS})$ (équ. 6) est défini par l'expression

$$\widehat{\mathrm{EQM}}(\bar{z}_{RS}) = \frac{K^2}{k\hat{N}_{RS}^2}(s_{bz}^2 - 2\bar{z}_{RS}s_{bzw} + \bar{z}_{RS}^2s_{bw}^2),$$

qui peut s'écrire comme en (10).

2.2 Méthode B

Cette méthode comporte les étapes suivantes:

a) Prélever k grappes parmi K au moyen d'un échantillonnage avec remise avec probabilité proportionnelle à la taille (ÉARPPT), les probabilités de sélection étant $P_i = N_i/M$, $i = 1, \ldots, K$.

b) Même étape que celle de la méthode A.

Théorème 4. L'estimateur par quotient suivant un échantillonnage PPT,

$$\bar{z}_{RP} = \hat{Y}_{RP}/\hat{N}_{RP} = \frac{M}{k}\sum_{i=1}^{k}\bar{z}_i\Big/\frac{M}{k}\sum_{i=1}^{k}\bar{w}_i \quad (13)$$

a pour biais relatif (au premier degré d'approximation)

$$BR(\bar{z}_{RP}) \doteq \frac{M^2}{k}\left[\left(\frac{\sigma^2_{bw'}}{N^2} - \frac{\sigma_{bzw'}}{YN}\right) + \sum_{i=1}^{K} \frac{N_i}{M}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)\left(\frac{S^2_{iw}}{N^2} - \frac{S_{izw}}{YN}\right)\right] \quad (14)$$

où

$$\sigma_{bzw'} = \sum_{i=1}^{K} (\bar{Z}_i - Y/M)(\bar{W}_i - N/M)(N_i/M)$$

et $\sigma^2_{bw'}$ correspond à $\sigma_{bzw'}$, z étant remplacé par w et Y, par N.

Démonstration. D'après un résultat courant, le biais relatif approximatif (au premier degré d'approximation) est

$$BR(\bar{z}_{RP}) \doteq [V(\hat{N}_{RP})/N^2] - \mathrm{Cov}(\hat{Y}_{RP}, \hat{N}_{RP})/YN. \quad (15)$$

Nous avons

$$V(\hat{N}_{RP}) = V_1E_2(\hat{N}_{RP}) + E_1V_2(\hat{N}_{RP})$$

$$= M^2\left[V_1\frac{1}{k}\sum_{i=1}^{k} E_2(\bar{w}_i) + E_1\frac{1}{k^2}\sum_{i=1}^{k} V_2(\bar{w}_i)\right]$$

$$= M^2\left[V_1\frac{1}{k}\sum_{i=1}^{k} \bar{W}_i + E_1\frac{1}{k^2}\sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S^2_{iw}\right]$$

$$= \frac{M^2}{k}\left[\sigma^2_{bw'} + \sum_{i=1}^{K}\frac{N_i}{M}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S^2_{iw}\right]. \quad (16)$$

De la même manière, nous pouvons écrire

$$\mathrm{Cov}(\hat{Y}_{RP}, \hat{N}_{RP}) = \frac{M^2}{k}\left[\sigma_{bzw'} + \sum_{i=1}^{K}\frac{N_i}{M}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S_{izw}\right]. \quad (17)$$

En substituant (16) et (17) dans (15), nous obtenons l'équation (14).

Théorème 5. L'EQM de l'estimateur $\bar{z}_{RP}$ (au premier degré d'approximation) est définie

$$\mathrm{EQM}(\bar{z}_{RP}) \doteq \frac{M}{kN^2}\sum_{i=1}^{K} N_i \times \left[(\bar{Z}_i - \bar{Y}\bar{W}_i)^2 + \left(\frac{1}{n_i} - \frac{1}{N_i}\right)D_i^2\right]. \quad (18)$$

Démonstration. Au premier degré d'approximation, nous écrivons

$$\mathrm{EQM}(\bar{z}_{RP}) \doteq [V(\hat{Y}_{RP}) - 2\bar{Y}\,\mathrm{Cov}(\hat{Y}_{RP}, \hat{N}_{RP}) + \bar{Y}^2V(\hat{N}_{RP})]/N^2. \quad (19)$$

En outre, d'après le théorème 2.5 de Singh (1988), nous avons, par analogie,

$$V(\hat{Y}_{RP}) = \frac{M^2}{k}\sigma^2_{bz'} + \frac{M^2}{k}\sum_{i=1}^{K}\frac{N_i}{M}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S^2_{iz}, \quad (20)$$

où $\sigma^2_{bz'} = \sum_{i=1}^{K}(N_i/M)(\bar{Z}_i - Y/M)^2$. En substituant (16), (17) et (20) dans l'équation (19), puis en simplifiant, nous obtenons l'équation (18).

Théorème 6. Un estimateur convergent de $\mathrm{EQM}(\bar{z}_{RP})$ (au premier degré d'approximation) est défini

$$\widehat{\mathrm{EQM}}(\bar{z}_{RP}) = \frac{M^2}{\hat{N}^2_{RP}} \cdot \frac{1}{k(k-1)}\sum_{i=1}^{k}(\bar{z}_i - \bar{z}_{RP}\bar{w}_i)^2. \quad (21)$$

Démonstration. Comme les unités de sondage du premier degré sont prélevées suivant un ÉARPPT, l'argument présenté dans la démonstration du théorème 3 vaut pour cette démonstration-ci.

D'après l'équation (20) et les résultats 2.9.1 et 4.5.1 de Särndal, Swensson et Wretman (1992), nous pouvons définir un estimateur non biaisé de

$$\sigma^2_{bz'} + \sum_{i=1}^{K}\frac{N_i}{M}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)S^2_{iz}$$

par l'expression

$$s^2_{bz'} = \frac{1}{k-1}\sum_{i=1}^{k}\left(\bar{z}_i - \sum_{i=1}^{k}\bar{z}_i/k\right)^2. \quad (22)$$

De la même manière, si nous définissons $s_{bzw'}$ et $s^2_{bw'}$, nous pouvons montrer que

$$\widehat{\mathrm{EQM}}(\bar{z}_{RP}) = \frac{M^2}{\hat{N}^2_{RP}k}(s^2_{bz'} - 2\bar{z}_{RP}s_{bzw'} + \bar{z}^2_{RP}s^2_{bw'}),$$

qui peut s'écrire comme en (21).

3. COMPARAISON DE L'EFFICACITÉ

Dans cette section, nous comparons l'efficacité des estimateurs utilisés respectivement dans les méthodes A et B.

Remarque. L'estimateur $\bar{z}_{RS}$, de la méthode B, devrait normalement être plus efficace que l'estimateur $\bar{z}_{RP}$, de la méthode A. Voici pourquoi. D'après (6) et (18), nous avons

Voici pourquoi. D'après (6) et (18), nous avons

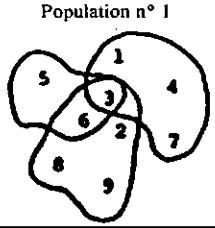
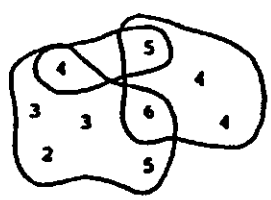
$$\mathrm{EQM}(\bar{z}_{RS}) - \mathrm{EQM}(\bar{z}_{RP}) \doteq \frac{M}{kN^2} \sum_{i=1}^{K} N_i \times \left[(\bar{Z}_i - \bar{Y}\bar{W}_i)^2 + \left(\frac{1}{n_i} - \frac{1}{N_i} \right) D_i^2 \right] \left(\frac{KN_i}{M} - 1 \right).$$

Si la taille de la grappe, N_i, augmente, le facteur $(KN_i/M - 1)$ augmentera aussi. L'autre facteur du terme en sommation est $N_i[(\bar{Z}_i - \bar{Y}\bar{W}_i)^2 + (1/n_i - 1/N_i)D_i^2]$, il représente la part de l'EQM de $\bar{z}_{RP}$ (équ. 18) attribuable à la variabilité de z et de w dans la grappe i (sans la constante M/kN^2). Si la taille de la grappe, N_i, augmente, la part de $\mathrm{EQM}(\bar{z}_{RP})$ attribuable à la grappe i devrait aussi augmenter, ce qui fait que la covariance des deux facteurs est positive. Donc, l'estimateur $\bar{z}_{RP}$ devrait avoir une EQM moins élevée que celle de $\bar{z}_{RS}$.

Exemple numérique. Dans cet exemple, nous appliquons les deux méthodes d'échantillonnage proposées à deux petites populations afin de mettre en lumière le calcul des valeurs F_{ij}, Z_{ij} et W_{ij} et la comparaison de ces valeurs. Pour les deux populations, $K = 3$, $k = 2$, $M = 12$ et $N = 9$. Il y a chevauchement dès qu'une unité est présente dans plus d'une grappe à la fois. Les populations sont décrites dans le tableau 1.

Tableau 1

Comparaison des deux méthodes pour deux petites populations

	Population n° 1			Population n° 2		
						
N_i	3	4	5	2	4	6
n_i	1	2	2	1	2	2
Y_{ij}	3,5,6	1,3,4,7	2,3,6,8,9	4,5	4,4,5,6	2,3,3,4,5,6
F_{ij}	3,1,2	1,3,1,1	1,3,2,1,1	2,2	1,1,2,2	1,1,1,2,1,2
Z_{ij}	1,5,3	1,1,4,7	2,1,3,8,9	2,2.5	4,4,2.5,3	2,3,3,2,5,3
W_{ij}	⅓,1,½	1,⅓,1,1	1,⅓,½,1,1	½,½	1,1,½,½	1,1,1,½,1,½
F	1.38	10.16	18.12	.24	.77	2.94
$\mathrm{EQM}(\bar{z}_{RS})$	2.09			0.45		
$\mathrm{EQM}(\bar{z}_{RP})$	1.83			0.33		
E.R.	114.21			136.36		
B.R.$(\bar{z}_{RS})$	−.0105			.0348		
B.R.$(\bar{z}_{RP})$	−.0047			−.0037		

L'analyse des résultats du tableau 1 confirme la théorie exposée dans cet article. Dans le cas des deux populations, le facteur $F = N_i[(\bar{Z}_i - \bar{Y}\bar{W}_i)^2 + (1/n_i - 1/N_i)D_i^2]$ augmente avec N_i, ce qui fait que $\mathrm{EQM}(\bar{y}_{RP}) < \mathrm{EQM}(\bar{y}_{RS})$, comme nous le disions plus haut.

4. CONCLUSION

Dans cet article, nous avons pu examiner la question du chevauchement des grappes sans nous embarrasser de l'hypothèse de la taille de population connue posée par Singh (1988). En outre, nous avons pu comparer les deux méthodes de façon plus directe, alors que Singh devait s'appuyer sur les résultats de Hansen et Hurwitz (1943).

REMERCIEMENTS

Cette étude a été rendue possible en partie grâce à une subvention (n° A-3111) du Conseil de recherches en sciences naturelles et en génie du Canada. Nous tenons aussi à exprimer toute notre reconnaissance aux arbitres et au rédacteur en chef, qui, par les précieux commentaires qu'ils ont faits sur une version antérieure de l'article, ont contribué à améliorer le produit.

BIBLIOGRAPHIE

AGARWAL, D.K., et SINGH, P. (1982). On cluster sampling strategies using ancillary information. *Sankhyā*, B, 44, 184-192.

AMDEKAR, S.J.(1985). An unbiased estimator in overlapping clusters. *Bulletin of the Calcutta Statistical Association*, 15, 231-232.

GIFFARD-JONES, W. (1993). The doctor game. *The Windsor Star*, April 15, 1993.

GOEL, B.B.P.S., et SINGH, D. (1977). On the formation of clusters. *Journal of the Indian Society of Agricultural Statistics*, 29, 53-68.

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

SÄRNDAL, C-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SINGH, S. (1988). Estimation in overlapping clusters. *Communications in Statistics, Theory and Methods*, 17, 613-621.

Techniques d'enquête, juin 1994
Vol. 20, n° 1, pp. 63–68
Statistique Canada

Échantillonnage avec PPT en deux occasions

N.G.N. PRASAD et J.E. GRAHAM[1]

RÉSUMÉ

La "méthode des groupes aléatoires" pour un échantillonnage avec probabilité proportionnelle à la taille (PPT) est étendue à un échantillonnage effectué en deux occasions. Nous utilisons de l'information sur une variable d'étude observée à la première occasion pour prélever la portion appariée de l'échantillon à la deuxième occasion. Nous examinons deux ensembles de données réels en vue d'une illustration numérique et d'une comparaison avec d'autres méthodes existantes.

MOTS CLÉS: Estimateur composite; comparaisons d'efficacité; méthode des groupes aléatoires; probabilité proportionnelle à la taille.

1. INTRODUCTION

Il est très fréquent, dans les enquêtes à passages répétés, qu'un plan de sondage avec remplacement partiel soit utilisé, en raison notamment de l'efficacité accrue de l'estimation qu'on espère ainsi obtenir, et aussi parce que cela permet de réduire le fardeau de réponse. Essentiellement, après chaque passage, une fraction des unités observées est retranchée de l'échantillon et remplacée par un nouveau sous-échantillon fraîchement prélevé dans la population. Cet ensemble d'unités non appariées est observé au passage suivant, en même temps que l'ensemble restant d'unités appariées. La littérature regorge d'analyses sur l'échantillonnage avec probabilités de sélection égales en deux occasions et sur les méthodes d'estimation connexes. Un cas particulièrement important survient lorsque les unités, à une occasion donnée, sont prélevées avec des probabilités de sélection inégales. Dans les travaux publiés jusqu'ici, on se sert de l'information recueillie à l'occasion précédente en vue d'améliorer l'estimateur habituel du total ou de la moyenne pour l'occasion courante, en utilisant une méthode d'estimation fondée sur les différences. Dans le présent article, nous présentons un plan visant un échantillonnage en deux occasions, et une méthode d'estimation connexe, qui se servent d'une information recueillie à l'occasion 1 (précédente) pour la sélection du sous-échantillon à observer à l'occasion 2 (courante). Par souci d'exhaustivité en même temps que de concision, nous n'examinons dans la présente section que les méthodes de sélection avec probabilités inégales aux deux occasions.

Considérons une population finie de N unités, désignées $1, 2, \ldots, N$, et un échantillonnage en deux occasions: 1 (occasion précédente) et 2 (occasion courante). Soient y_{1i} et y_{2i} les valeurs d'une caractéristique y pour la $i$ième unité observée à la première et à la deuxième occasion respectivement. Soient Y_1 et Y_2 les totaux respectifs pour la population. Supposons qu'une mesure de taille x soit connue pour chacune des unités de la population.

1.1 Le plan de Des Raj

Raj (1965) a examiné le plan d'échantillonnage avec PPT (probabilité proportionnelle à la taille) suivant: à la première occasion, un échantillon s de taille n est prélevé avec probabilités p_i proportionnelles aux valeurs x_i, $i = 1, 2, \ldots, N$, et avec remplacement. À la deuxième occasion, un échantillon aléatoire simple s_1 de m unités est prélevé dans s sans remplacement et un échantillon indépendant s_2 de $u = n - m$ unités est prélevé avec PPT et avec remplacement dans l'ensemble de la population. Les totaux Y_1 et Y_2 sont alors respectivement estimés sans biais par:

$$\hat{Y}_1 = \sum_{i \in s} y_{1i}/(np_i) \tag{1.1}$$

et

$$\hat{Y}_2 = Q\hat{Y}_{2u} + (1 - Q)\hat{Y}_{2m}, \tag{1.2}$$

où

$$\hat{Y}_{2u} = \sum_{i \in s_2} y_{2i}/(up_i), \tag{1.3}$$

$$\hat{Y}_{2m} = \sum_{i \in s} y_{1i}/(np_i) + \sum_{i \in s_1} (y_{2i} - y_{1i})/(mp_i), \tag{1.4}$$

et Q est un coefficient de pondération, $0 \leq Q \leq 1$. Si l'on suppose que

$$V_1 = \sum_{i=1}^{N} (y_{1i}/p_i - Y_1)^2 p_i = V_2$$

$$= \sum_{i=1}^{N} (y_{2i}/p_i - Y_2)^2 p_i = V, \tag{1.5}$$

[1] N.G.N. Prasad, Associate Professor, Department of Statistics and Applied Probability, University of Alberta, Edmonton, Alberta, Canada T6G 2G1; J.E. Graham, Professor, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6.

on trouve que la variance minimum de $\hat{Y}_2$ est

$$V_{min}(\hat{Y}_2) = V[1 + \sqrt{2(1-\delta)}/(2n)], \quad (1.6)$$

où δ est donné par

$$V\delta = \sum_{i=1}^{N} (y_{1i}/p_i - Y_1)(y_{2i}/p_i - Y_2)p_i. \quad (1.7)$$

1.2 Le plan de Ghangurde-Rao (G-R)

En utilisant un plan d'échantillonnage avec PPT sans remplacement, Ghangurde et Rao (1969) ont étendu la "méthode de Rao-Hartley-Cochran (RHC)", aussi appelée "méthode des groupes aléatoires" (voir Rao, Hartley et Cochran 1962) à l'échantillonnage en deux occasions. Dans la méthode RHC, la population de N unités est divisée au hasard en n groupes de tailles $N_1, N_2, \ldots, N_n$ telles que $\sum_{h=1}^{n} N_h = N$, et un échantillon d'une unité est tiré indépendamment de chacun des n groupes avec probabilités proportionnelles aux probabilités de sélection initiales, p_i. Dans la méthode G-R, la population est d'abord divisée au hasard en n groupes de taille N/n (qu'on suppose être un entier). À la première occasion, une unité est tirée de chaque groupe aléatoire (comme ci-dessus), ce qui produit un échantillon s de n unités. À la deuxième occasion, un échantillon aléatoire simple s_1 de $m = \lambda n (o < \lambda < 1)$ unités appariées est tiré de s sans remplacement, et un échantillon indépendant s_2 de $u = n - m$ unités est tiré de l'ensemble de la population de N unités par la même méthode que celle utilisée pour produire s. On obtient alors un estimateur composite de Y_2 sous la forme suivante

$$\hat{Y}_2' = Q'\hat{Y}_{2u}' + (1 - Q')\hat{Y}_{2m}', \quad (1.8)$$

où $0 \leq Q' \leq 1$,

$$\hat{Y}_{2u}' = \sum_{i \in s_2} \frac{y_{2i}P_i^*}{p_i}, \quad (1.9)$$

et

$$\hat{Y}_{2m}' = \sum_{i \in s} \frac{y_{1i}P_i}{p_i} + nm^{-1} \sum_{i \in s_1} \frac{(y_{2i} - y_{1i})P_i}{p_i}, \quad (1.10)$$

P_i et P_i^* désignant les totaux des valeurs p_i pour les groupes contenant la $i$ième unité ($i = 1, 2, \ldots, N$) dans la sélection de s et s_2 respectivement. En vertu de l'hypothèse (1.5), la variance de $\hat{Y}_2'$ (avec valeurs optimales de Q' et λ) est donnée par

$$V_{min}(\hat{Y}_2') = \frac{NV}{2n(N-1)} \times [1 - n/N + \sqrt{2(1-\delta)}(1+\gamma)n/N], \quad (1.11)$$

où

$$\gamma = \frac{(1-\rho')V'}{(1-\delta)V} - 1,$$

$$V' = N^{-1} \sum_{i=1}^{N} (y_{1i} - \bar{Y}_1)^2 = N^{-1} \sum_{i=1}^{N} (y_{2i} - \bar{Y}_2)^2$$

et

$$\rho' = N^{-1} \sum_{i=1}^{N} (y_{1i} - \bar{Y}_1)(y_{2i} - \bar{Y}_2)/V'.$$

1.3 Le plan de Chotai

Chotai (1974), en posant comme hypothèse additionnelle que n/m est un entier, a modifié le plan d'échantillonnage G-R pour la deuxième occasion. Un échantillon s est prélevé comme dans le plan G-R à la première occasion. À la deuxième occasion, les n unités de l'échantillon s sont réparties au hasard en $m (= \lambda n)$ groupes de taille n/m. Une unité est tirée indépendamment de chacun des m groupes avec probabilités proportionnelles aux P_i, tels que définis dans le plan G-R. Cette sélection donne l'échantillon s_1. La sélection de s_2 se fait de la même façon que dans le plan G-R. On dispose alors d'un estimateur composite de Y_2 sous la forme

$$\hat{Y}_2^C = Q^C\hat{Y}_{2u}^C + (1 - Q^C)\hat{Y}_{2m}^C, \quad (1.12)$$

où $0 \leq Q^C \leq 1$,

$$\hat{Y}_{2u}^C = \sum_{i \in s_2} \frac{y_{2i}P_i^*}{p_i}, \quad (1.13)$$

et

$$\hat{Y}_{2m}^C = \sum_{i \in s_1} \frac{(y_{2i} - y_{1i})P_i^+}{p_i} + \sum_{i \in s} \frac{y_{1i}P_i}{p_i}. \quad (1.14)$$

Ici, P_i et P_i^* sont définis comme dans le plan G-R, et P_i^+ désigne le total des valeurs P_i pour les groupes aléatoires de s contenant la $i$ième unité ($i = 1, 2, \ldots, N$) dans la sélection de s_1. La variance minimum de $\hat{Y}_2^C$ en vertu de l'hypothèse (1.5), obtenue à l'aide des valeurs optimales de Q^C et λ, est donnée par

$$V_{min}(\hat{Y}_2^C) = \frac{NV}{2n(N-1)}[1 - n/N + \sqrt{2(1-\delta)}]. \quad (1.15)$$

En vertu de ce plan, mais sans l'hypothèse (1.5), Chotai a aussi examiné un estimateur de Y_2 (semblable à l'estimateur de Kulldorff pour l'échantillonnage aléatoire simple; voir Kulldorff 1963) donné par

$$\hat{Y}_2^{CM} = Q^{CM}\hat{Y}_{2u}^{C} + (1 - Q^{CM})\hat{Y}_{2m}^{CM}, \qquad (1.16)$$

où $\hat{Y}_{2u}^{C}$ est défini comme en (1.13), Q^{CM} $(0 \leq Q^{CM} \leq 1)$ est un coefficient de pondération donné qui doit être déterminé et

$$\hat{Y}_{2m}^{CM} = \sum_{i \in s_1} \frac{(y_{2i} - \beta y_{1i})P_i^{+}}{p_i} + \beta \sum_{i \in s} \frac{y_{1i}P_i}{p_i}, \qquad (1.17)$$

avec

$$\beta = \delta \left[\frac{\sum_{i=1}^{N} p_i(y_{2i}/p_i - Y_2)^2}{\sum_{i=1}^{N} p_i(y_{1i}/p_i - Y_1)^2} \right] = \delta \frac{V_2}{V_1}, \qquad (1.18)$$

et δ tel que défini en (1.7). La variance minimum de $\hat{Y}_2^{CM}$, selon les valeurs optimales de Q^{CM} et λ, est donnée par

$$V_{min}(\hat{Y}_2^{CM}) = \frac{N}{2n(N-1)}(1 + \sqrt{1 - \delta^2} - n/N)V_2. \qquad (1.19)$$

Pour utiliser effectivement $\hat{Y}_2^{CM}$, il est évidemment nécessaire de trouver d'abord la valeur de β, ce qui, en général, n'est pas possible en pratique. On peut utiliser une estimation de β fondée sur l'échantillon disponible, mais cela introduira un biais dans l'estimateur $\hat{Y}_2^{CM}$.

2. AUTRES PLANS POSSIBLES POUR UN ÉCHANTILLONNAGE AVEC PPT EN DEUX OCCASIONS

Nous présentons maintenant une autre méthode possible d'échantillonnage et d'estimation, qui n'exige pas de connaître la valeur β telle que définie en (1.18). Selon ce plan, on se sert d'une information recueillie à la première occasion pour sélectionner l'échantillon à la deuxième occasion. La méthode se fonde sur une technique que Prasad et Srivenkataramana (1980) ont élaborée dans le contexte d'un échantillonnage double dans lequel un sous-échantillon de deuxième phase est prélevé selon l'information obtenue d'un échantillon initial. À des fins de simplicité, nous examinons d'abord son application au plan de Raj (1965) décrit plus haut.

2.1 Modification du plan de Des Raj

À la première occasion, un échantillon s de taille n est sélectionné avec probabilités p_i proportionnelles aux valeurs x_i et avec remplacement. À la deuxième occasion, plutôt que de prélever un sous-échantillon aléatoire simple sans remplacement, on prélève un sous-échantillon s_1 de m unités dans s, selon un plan avec PPT et avec remplacement, en prenant comme mesure de la taille $z_i = y_{1i}/x_i$, où y_{1i} est la valeur observée de la caractéristique y pour l'unité i à la première occasion. Un échantillon s_2 de taille $u = n - m$ est tiré indépendamment de s, comme dans Raj (1965). Un estimateur composite de Y_2 est donné par

$$\tilde{Y}_2 = Q\,\hat{Y}_{2u} + (1 - Q)\tilde{Y}_{2m},$$

où $\hat{Y}_{2u}$ est tel que défini en (1.3) et

$$\tilde{Y}_{2m} = \frac{1}{nm} \sum_{i \in s_1} \frac{(y_{2i}/p_i)}{(y_{1i}/p_i)} \sum_{i \in s} (y_{1i}/p_i),$$

Q étant un coefficient de pondération, $0 \leq Q \leq 1$. La variance minimum de $\tilde{Y}_2$, obtenue en minimisant la variance de $\tilde{Y}_2$ par rapport à Q, est donnée par

$$V_{min}(\tilde{Y}_2) = V_1C_1(n + C_1m)^{-1},$$

où $C_1 = \sum_{i=1}^{N} (y_{2i}/p_{1i} - Y_2)^2 p_{1i} V_1^{-1}$, avec $p_{1i} = y_{1i}/Y_1$ et V_1 tel que défini en (1.5).

2.2 Modification du plan de Chotai

Supposons, comme dans Chotai (1974), que N, n et $m(< n)$ soient tous des entiers positifs tels que N/n, N/u et n/m sont aussi des entiers. Alors:

1. À la première occasion, prélevons un échantillon s de taille n selon une méthode identique à celle du plan G-R. Pour cet ensemble d'unités, des observations y_{1i}, $i = 1, \ldots, n$, sont faites au sujet d'une caractéristique y.
2. À la deuxième occasion, a) répartissons les n unités de s au hasard en m groupes de taille n/m et tirons indépendamment une unité avec PPT, $p_i^{*} = (y_{1i}P_i)/p_i$, de chacun des m groupes afin de constituer un sous-échantillon s_1, où P_i est tel que défini dans le plan G-R; b) sélectionnons s_2, un échantillon entièrement nouveau de $u = n - m$ unités dans la population entière, et observons les valeurs y de la deuxième occasion, y_{2i}, pour ces u unités de la même manière que dans le plan G-R.

Notons que la différence entre la méthode proposée et celle de Chotai (1974) réside dans la sélection de s_1: dans le premier cas, on se sert de l'information recueillie à la première occasion pour prélever s_1 à la deuxième occasion.

Examinons maintenant un estimateur du total pour la deuxième occasion, Y_2, qui exploite la méthode proposée. Posons

$$y_{2i}^* = \frac{y_{2i}P_i}{p_i}.$$

Un estimateur composite de Y_2 est donné par

$$\hat{Y}_2^* = Q^{**}\hat{Y}_{2u}^C + (1 - Q^{**})\hat{Y}_{2m}^*, \qquad (2.1)$$

où $\hat{Y}_{2u}^C$ est défini comme en (1.13), $0 \leq Q^{**} \leq 1$ et

$$\hat{Y}_{2m}^* = \sum_{i \in s_1} \frac{y_{2i}^* \tilde{P}_i}{p_i^*}.$$

Ici, $\tilde{P}_i$ désigne le total des valeurs p_i^* associées aux unités appartenant au groupe aléatoire d'où provient la $i$ième unité de s_1. Supposons que E_1 et E_2 désignent l'espérance et que V_1 et V_2 désignent la variance sur tous les s et pour un s donné, respectivement. On peut alors constater que $\hat{Y}_{2m}^*$, et donc $\hat{Y}_2^*$ pour Y_2, sont sans biais, car l'espérance de $\hat{Y}_{2m}^*$ est

$$E(\hat{Y}_{2m}^*) = E_1E_2(\hat{Y}_{2m}^*) = E_1\left(\sum_{i \in s} \frac{y_{2i}P_i}{p_i}\right) = Y_2. \qquad (2.2)$$

Pour obtenir la variance de $\hat{Y}_{2m}^*$, considérons

$$V_2(\hat{Y}_{2m}^*) = \frac{n - m}{m(n-1)} \sum_{i \in s} \left(\frac{y_{2i}^*}{p_i^*} - \sum_{i \in s} y_{2i}^*\right)^2 p_i^*$$

$$= \frac{n - m}{m(n-1)}\left[\sum_{i \in s} \frac{(y_{2i}^2/y_{1i})}{p_i} P_i \sum_{i \in s} \frac{y_{1i}P_i}{p_i} - \left(\sum_{i \in s} \frac{y_{2i}P_i}{p_i}\right)^2\right],$$

ce qui donne, après une importante simplification algébrique

$$E_1V_2(\hat{Y}_{2m}^*) = \frac{N(n-m)}{mn(N-1)}\sigma_3^2,$$

où

$$\sigma_3^2 = \sum_{i=1}^{N} \left(\frac{y_{2i}}{y_{1i}} Y_1 - Y_2\right)^2 \frac{y_{1i}}{Y_1}.$$

Puisque

$$V_1E_2(\hat{Y}_{2m}^*) = \frac{N-n}{n(N-1)}\sigma_2^2,$$

on a

$$V(\hat{Y}_{2m}^*) = \frac{N}{n(N-1)}\left[(1 - n/N) + \frac{1-\lambda}{\lambda} h\right]\sigma_2^2, \qquad (2.3)$$

où

$$h = \frac{\sigma_3^2}{\sigma_2^2},\ \sigma_2^2 = V_2 = \sum_{i=1}^{N} (y_{2i}/p_i - Y_2)^2 p_i \quad \text{et} \quad \lambda = \frac{m}{n}.$$

Parce que $\hat{Y}_{2u}^C$ et $\hat{Y}_{2m}^*$ sont indépendants, la variance de $\hat{Y}_2^*$ est donnée par

$$V(\hat{Y}_2^*) = Q^{**2}V(\hat{Y}_{2u}^C) + (1 - Q^{**})^2V(\hat{Y}_{2m}^*),$$

où

$$V(\hat{Y}_{2u}^C) = \frac{N-u}{u(N-1)}\sigma_2^2,$$

et $V(\hat{Y}_{2m}^*)$ est donnée par (2.3).

On obtient la variance minimum de $V(\hat{Y}_2^*)$ en utilisant les valeurs optimales de Q^{**} et λ, données respectivement par

$$Q^{**} = \frac{(1 - n/N) + \frac{(1-\lambda)}{\lambda} h}{(1 - n/N) + \frac{(1-\lambda)}{\lambda} h + \frac{(1 - (1-\lambda)n/N)}{(1-\lambda)}},$$

et

$$\lambda = \frac{\sqrt{h}}{1 + \sqrt{h}}.$$

Donc, la variance minimum de $V(\hat{Y}_2^*)$ est donnée par

$$V_{min}(\hat{Y}_2^*) = \frac{N\sigma_2^2}{n(N-1)}[1 - n/N + \sqrt{h}]. \qquad (2.4)$$

Notons que la quantité h reflète l'efficacité de l'estimateur utilisant les p_i comme probabilités de sélection initiales par rapport à celle de l'estimateur fondé sur les probabilités de sélection initiales y_{1i}/Y_1. Une valeur "peu élevée" de h entraîne un accroissement de l'efficacité de la méthode proposée par rapport à celle de Chotai.

3. COMPARAISONS NUMÉRIQUES DE L'EFFICACITÉ

Les estimateurs composites $\hat{Y}_2^C$ défini en (1.12), $\hat{Y}_2^{CM}$ défini en (1.16) et $\hat{Y}_2^*$ défini en (2.1) sont maintenant comparés à leurs valeurs optimales Q et λ respectives. L'efficacité du plan proposé en 2.2 par rapport à la méthode de Chotai (1974) est examinée au moyen d'une comparaison des deux efficacités relatives suivantes:

$$\text{RE1} = \frac{V_{min}(\hat{Y}_2^C)}{V_{min}(\hat{Y}_2^*)} = \frac{(1 - n/N) + \sqrt{2(1 - \delta)}}{(1 - n/N) + \sqrt{h}}$$

et

$$\text{RE2} = \frac{V_{min}(\hat{Y}_2^{CM})}{V_{min}(\hat{Y}_2^*)} = \frac{(1 - n/N) + \sqrt{1 - \delta^2}}{(1 - n/N) + \sqrt{h}},$$

obtenues respectivement à l'aide de (1.15) et (2.4), et de (1.19) et (2.4). Il s'ensuit que le plan proposé est supérieur au plan de Chotai utilisant un estimateur de Kulldorff (qui dépend de la constante inconnue β) dans le cas des populations pour lesquelles $h < (1 - \delta^2)$. Afin de permettre des comparaisons numériques utiles, nous utilisons deux ensembles de données déjà examinés dans la littérature.

Ensemble de données A: Cet ensemble de données a trait à la superficie de culture du blé en 1964 (y_2) et en 1963 (y_1), ainsi qu'à la superficie cultivée en 1961 (x), pour 34 villages de l'Inde (voir Murthy 1967). Les valeurs des paramètres pour cet ensemble de données sont $\delta = 0.6404$ et $h = 0.1868$.

Ensemble de données B: Cet ensemble de données a trait à la superficie de culture du blé en 1937 (y_2) et en 1936 (y_1), ainsi qu'à la superficie cultivée en 1930 (x) pour un échantillon de 34 villages de l'Inde (voir Sukhatme, P.V. et Sukhatme, B.V. 1970). Les valeurs correspondantes des paramètres pour cet ensemble de données sont $\delta = 0.7635$ et $h = 0.3811$.

D'après ces valeurs de δ et h, les deux efficacités relatives RE1 et RE2 (exprimées en pourcentage) ont été calculées pour certaines valeurs de n/N, et les résultats sont présentés aux tableaux 1 et 2.

Tableau 1

RE1 en % pour les ensembles de données A et B

n/N	Ensemble A	Ensemble B
0.05	130.09	124.30
0.10	131.22	125.21
0.15	132.43	126.19
0.20	133.75	127.25
0.25	135.18	128.41
0.30	136.73	129.66

Tableau 2

RE2 en % pour les ensembles de données A et B

n/N	Ensemble A	Ensemble B
0.05	104.49	101.82
0.10	104.64	101.88
0.15	104.80	101.94
0.20	104.97	102.01
0.25	105.15	102.08
0.30	105.34	102.16

Un examen du tableau 1 permet de conclure que le plan proposé est supérieur à celui de Chotai (1974). Le gain d'efficacité passe de 30% à 37% pour l'ensemble de données A et de 24% à 30% pour l'ensemble de données B avec l'augmentation de 0.05 à 0.30 de la fraction de sondage. Notons que le gain d'efficacité est plus élevé pour l'ensemble de donnés A que pour l'ensemble de données B en raison de la différence entre les valeurs des paramètres h (0.1868 contre 0.3811) et δ (0.6404 contre 0.7635). Rappelons que h évalue l'efficacité de p_i en tant que mesure de la taille pour l'unité i par rapport à celle de y_{1i} en tant que mesure de la taille dans l'estimation du total Y_2 pour l'occasion courante, tandis que δ est la corrélation entre y_{1i}/p_i et y_{2i}/p_i telle que définie en (1.7). Quand h est relativement petit, on obtient avec la méthode proposée des gains d'efficacité supérieurs à ceux obtenus quand h est élevé. Dans les deux cas, toutefois, les gains d'efficacité que procure la méthode proposée sont dignes d'intérêt.

Les gains d'efficacité qu'engendre la méthode proposée par rapport au plan de Chotai utilisant un estimateur de Kulldorff (résultats présentés au tableau 2) sont minimes, variant de 4.5% à 5.3% pour l'ensemble de données A et de 1.8% à 2.2% pour l'ensemble de données B. Toutefois, il est nécessaire de connaître la valeur de β pour pouvoir utiliser l'estimateur de Kulldorff, ce qui n'est pas le cas en pratique. La stratégie proposée, donc, donne de bons résultats tant du point de vue de son application en situation réelle que de celui du gain d'efficacité.

Il existe des situations dans lesquelles l'information auxiliaire nécessaire au calcul des probabilités de sélection initiales n'est pas disponible. Un échantillonnage aléatoire simple peut alors être utilisé en remplacement de la méthode RHC pour la sélection de l'échantillon de la première occasion; on peut ensuite appliquer la méthode RHC au prélèvement de s_1, en utilisant l'information sur la variable à l'étude fournie par l'échantillonnage aléatoire simple effectué à la première occasion. La théorie relative à une telle méthode s'obtient directement à titre de cas spécial de celui présenté, en prenant $p_i = 1/N$, $i = 1, \ldots, N$. D'importants gains d'efficacité devraient être obtenus dans ce cas.

REMERCIEMENTS

Les auteurs remercient le professeur J.N.K. Rao de leur avoir suggéré l'étude de ce problème, et sont reconnaissants envers l'arbitre pour les suggestions constructives formulées. Cette recherche a bénéficié de subventions du Conseil de recherches en sciences naturelles et en génie du Canada.

BIBLIOGRAPHIE

CHOTAI, J. (1974). A note on the Rao-Hartley-Cochran method for PPS sampling over two occasions. *Sankhyā*, Séries C, 36, 173-180.

GHANGURDE, P., et RAO, J.N.K. (1969). Some results on sampling over two occasions. *Sankhyā*, Séries A, 31, 463-472.

KULLDORFF, G. (1963). Some problems of optimum allocation for sampling on two occasions. *Revue de l'Institut International de Statistique*, 31, 24-57.

MURTHY, N.N. (1967). *Sampling Theory and Methods*. Calcutta, India: Statistical Publishing Society.

PRASAD, N.G.N., et SRIVENKATARAMANA, T. (1980). Double sampling with PPS selection. *Vignana Bharathi*, 6, 52-58.

RAJ, D. (1965). On sampling over two occasions with probabilities proportional to size. *Annals of Mathematical Statistics*, 36, 327-330.

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

SUKHATME, P.V., et SUKHATME, B.V. (1970). *Sampling Theory of Surveys With Applications*. Ames, Iowa: Iowa State University Press.

Techniques d'enquête, juin 1994
Vol. 20, n° 1, pp. 69–78
Statistique Canada

Stratification multidimensionnelle par programmation linéaire

R.R. SITTER et C.J. SKINNER[1]

RÉSUMÉ

Rao et Nigam (1990, 1992) ont montré comment une classe de plans à échantillonnage contrôlé peut être produite par programmation linéaire. Dans cet article, leur méthode est appliquée à la stratification multidimensionnelle. L'examen des plans de sondage produits dans des applications particulières, ainsi que l'évaluation des erreurs quadratiques moyennes, permettent d'établir une comparaison avec les méthodes existantes. La méthode proposée est d'utilisation relativement simple et semble avoir une performance raisonnable sur le plan des erreurs quadratiques moyennes. Les calculs nécessaires peuvent toutefois augmenter rapidement, à mesure que s'accroît le nombre de cellules dans la classification multidimensionnelle. L'estimation de la variance est également examinée.

MOTS CLÉS: Choix contrôlé; programmation linéaire, échantillonnage à plusieurs degrés; échantillonnage stratifié.

1. INTRODUCTION

Souvent, le concepteur d'une enquête dispose de plusieurs variables pour faire une stratification et choisit naturellement de définir les strates comme étant les cellules formées par la classification multidimensionnelle de ces variables. Un problème que pose cette méthode, notamment lorsqu'on prélève les unités primaires d'échantillonnage (upé) pour des enquêtes sur les ménages, vient du fait que la taille désirée de l'échantillon peut être inférieure au nombre total de cellules, auquel cas les méthodes de stratification classiques peuvent ne pas s'appliquer.

Une illustration, basée sur un exemple hypothétique de Bryant et coll. (1960), est donnée au tableau 1. Des localités (upé) sont classées selon deux variables de stratification: le type de localité (trois catégories) et la région (cinq catégories). La taille désirée de l'échantillon, $n = 10$, est inférieure au nombre total de cellules, 15. Cet exemple fait apparaître un problème connexe. Les valeurs du tableau 1 sont les nombres attendus selon une stratification proportionnelle, c'est-à-dire les proportions de la population multipliées par la taille de l'échantillon. Même si la taille de l'échantillon était doublée, afin qu'elle dépasse le nombre de cellules, les nombres attendus ne seraient toujours pas des entiers. Le fait d'arrondir ces valeurs à des nombres entiers aurait sans doute un effet pratiquement négligeable pour des nombres attendus élevés, mais le choix de la méthode d'arrondissement est plus préoccupant lorsque ces nombres sont très faibles.

Une solution au problème que pose le trop grand nombre de cellules consiste à laisser de côté une ou plusieurs des variables de stratification, ou encore à fusionner certaines catégories. Ont également été proposées diverses autres méthodes en vertu desquelles on tente de maintenir un certain "contrôle" pour toutes les catégories de l'ensemble des variables de stratification, en permettant des formes différentes de sélection aléatoire des cellules.

Tableau 1

Nombres probables d'unités de l'échantillon dans les cellules en vertu d'une stratification proportionnelle avec $n = 10$

Région	Type de localité			
	Urbaine	Rurale	Métropolitaine	Total
1	1.0	0.5	0.5	2.0
2	0.2	0.3	0.5	1.0
3	0.2	0.6	1.2	2.0
4	0.6	1.8	0.6	3.0
5	1.0	0.8	0.2	2.0
Total	3.0	4.0	3.0	10.0

Goodman et Kish (1950) ont proposé une méthode qu'ils ont baptisée "choix contrôlé". Jessen (1970) se dit d'avis que "cette méthode est un peu complexe et son utilisation en échantillonnage appliqué apparaît limitée" (p. 778). Waterton (1983) illustre cette complexité, tandis que Bryant et coll. (1960) proposent une méthode beaucoup plus simple pour une stratification bidimensionnelle. Cette méthode se caractérise par une indépendance des nombres attendus d'unités échantillonnées entre les lignes et les colonnes du tableau bidimensionnel. Si les lignes et les colonnes sont également indépendantes dans la population, aucun problème ne se pose, mais si l'on note un degré important de dépendance (comme c'est souvent le cas), il faut habituellement revoir la pondération, ce qui peut rendre cette solution inintéressante en pratique et accroître la variance, comme il est démontré à la section 5. Jessen (1970) signale qu'une autre limite de la méthode de Bryant et coll. (1960) vient de l'impossibilité de contraindre des cellules déterminées à avoir une taille nulle. Il propose deux méthodes, applicables à des stratifications bidimensionnelles et tridimensionnelles, mais ces méthodes demeurent d'application assez difficile et, comme le notent Causey et coll. (1985), elles ne produisent pas toujours une solution.

[1] R.R. Sitter, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6; C.J. Skinner, Department of Social Statistics, University of Southampton S09 5NH, U.K.

Toutes les méthodes ci-dessus peuvent être appliquées manuellement avec un degré d'effort variable, mais aucune ne tire parti du potentiel de l'informatique moderne. Dans cet article, nous allons montrer comment les techniques informatiques de la programmation linéaire peuvent être appliquées au problème de la stratification multidimensionnelle, comme l'ont fait auparavant Rao et Nigam (1990, 1992). La méthode que nous proposons peut être considérée comme un complément de la méthode de programmation linéaire proposée par Causey et coll. (1985). Le choix entre les deux dépendra de la nature du problème de stratification et du logiciel disponible. Notre méthode a comme inconvénient possible d'être beaucoup plus exigeante en calculs, puisque le nombre d'inconnues dans notre problème de programmation linéaire peut être aussi élevé que $\binom{k}{n}$, où k est le nombre de cellules du tableau et n est la taille de l'échantillon, tandis qu'il n'y a que k inconnues dans la méthode de Causey et coll. (1985). Nous ferons toutefois quelques suggestions destinées à réduire les calculs exigés par notre méthode. Il s'agit d'une méthode qui est susceptible d'offrir plusieurs avantages. Premièrement, la correspondance est directe entre le problème de stratification et le problème de programmation linéaire, de sorte que la programmation informatique est immédiate, tandis que la méthode de Causey et coll. est moins directe, ce qui oblige à simuler le comportement de fonctions non linéaires à l'aide de fonctions linéaires (p. 904) et à imbriquer des problèmes de programmation linéaire répétés dans un algorithme comportant une récursivité additionnelle. Deuxièmement, notre méthode donne toujours une solution, tandis que celle de Causey et coll. n'en produit pas toujours une, par exemple dans des cas de stratification tridimensionnelle. Troisièmement, la fonction objective de notre problème de programmation linéaire peut être naturellement modifiée en fonction des différents objectifs du problème de stratification, par exemple pour une stratification tridimensionnelle dans laquelle il est plus important d'"équilibrer" l'échantillon à l'égard des deux premières variables de stratification qu'à l'égard de la troisième. Quatrièmement, notre méthode peut être naturellement modifiée pour forcer les probabilités d'inclusion conjointe de cellules à être positives, de façon à permettre une estimation non biaisée de la variance.

2. MÉTHODE PROPOSÉE

2.1 Concepts de base

Examinons, pour commencer, la forme la plus simple de stratification bidimensionnelle. Soit une population de N unités répartie entre les RC cellules d'un tableau bidimensionnel résultant de la classification croisée d'un facteur de stratification ligne comprenant R catégories et d'un facteur colonne comprenant C catégories. Soit N_{ij} le nombre d'unités de la cellule ij, c'est-à-dire de l'ensemble d'unités appartenant à la fois à la ligne i et à la colonne j, et soit $P_{ij} = N_{ij}/N$ la proportion correspondante. Supposons que le paramètre d'intérêt soit la moyenne de la population, $\bar{Y}$, d'une variable Y.

Examinons la méthode d'échantillonnage à deux degrés suivantes. Premièrement, des tailles d'échantillon n_{ij} sont déterminées pour chaque cellule d'après une procédure aléatoire donnée. Si s désigne la matrice $R \times C$ des $(n_{ij},\ i = 1, \ldots, R,\ j = 1, \ldots, C)$, cette procédure attribue une probabilité $p(s)$ à chaque s d'un ensemble de S matrices possibles. Pour bien montrer que les n_{ij} dépendent de s, nous écrivons $n_{ij}(s)$. Deuxièmement, un échantillon aléatoire simple de $n_{ij}(s)$ unités est prélevé dans la cellule ij et les valeurs de Y propres aux unités de l'échantillon sont enregistrées.

Nous limitons notre examen aux plans comportant une taille d'échantillon fixe $n > 0$, c'est-à-dire que nous restreignons S à l'ensemble S_n de toutes les matrices telles que

$$\sum_{i=1}^{R} \sum_{j=1}^{C} n_{ij}(s) = n.$$

Nous limitons aussi notre attention à une stratification proportionnelle telle que

$$\sum_{s \in S_n} n_{ij}(s)p(s) = nP_{ij} \quad \text{pour} \quad i = 1, \ldots, R, \quad j = 1, \ldots, C. \tag{2.1}$$

Il résulte de (2.1) que la moyenne non pondérée simple de l'échantillon, $\bar{y}(s)$, est un estimateur sans biais de $\bar{Y}$. Nous proposons de choisir un (ou le) plan d'échantillonnage $p(s)$ qui minimise le degré de "non-désirabilité" attendu de l'échantillon s, en résolvant le problème:

$$\underset{p \in P}{\text{minimiser}} \sum_{s \in S_n} w(s)p(s), \tag{2.2}$$

sous la contrainte (2.1), où $w(s)$ est une fonction de perte (à préciser) qui s'applique à l'échantillon s, et P est la classe des plans d'échantillonnage possibles sur S_n qui remplissent la condition

$$0 \leq p(s) \leq 1 \quad \text{pour tout} \quad s \in S_n. \tag{2.3}$$

Notons qu'il découle de (2.1) que $\sum_{s \in S_n} p(s) = 1$. L'observation clé de Rao et Nigam (1990, 1992) est que la fonction objective énoncée en (2.2), ainsi que les contraintes d'égalité et d'inégalité énoncées en (2.1) et (2.3), sont toutes des fonctions linéaires en $p(s)$, et donc qu'on peut résoudre ce problème directement par programmation linéaire, les inconnues étant $p(s)$, $s \in S_n$. Le principal obstacle à l'application de cette méthode vient du fait que le nombre d'éléments de S_n est souvent très élevé et qu'il est difficile, même avec les ressources informatiques modernes, de résoudre par programmation linéaire un problème qui comporte un grand nombre d'inconnues.

Il vaut mieux, par conséquent, restreindre l'examen à un sous-ensemble de S_n. Une restriction naturelle consiste à ne tenir compte que des matrices s pour lesquelles $n_{ij}(s)$ est égal soit à $I_{ij} = [nP_{ij}]$, le plus grand entier inférieur à nP_{ij}, soit à $I_{ij} + 1$. Si l'on pose $\tilde{n}_{ij}(s) = n_{ij}(s) - I_{ij}$ et $r_{ij} = nP_{ij} - I_{ij}$, le problème devient

$$\underset{p \in P}{\text{minimiser}} \sum_{s \in \tilde{S}_{\tilde{n}}} w(s)p(s), \tag{2.4}$$

sous les contraintes

$$\sum_{s \in \tilde{S}_{\tilde{n}}} \tilde{n}_{ij}(s)p(s) = r_{ij}, \tag{2.5}$$

$$\sum_{s \in \tilde{S}_{\tilde{n}}} p(s) = 1, \quad 0 \leq p(s) \leq 1 \quad \text{pour tout} \quad s \in \tilde{S}_{\tilde{n}}, \tag{2.6}$$

où $\tilde{S}_{\tilde{n}}$ est l'ensemble des matrices $R \times C$ où tous les éléments sont 0 ou 1 et où la somme des éléments est $\tilde{n} = n - \sum_{ij} I_{ij}$. On peut constater, évidemment, que si tous les I_{ij} sont égaux à zéro, le problème est exactement le même qu'avant. Le nombre d'éléments dans $\tilde{S}_{\tilde{n}}$, qui détermine l'ampleur du travail de calcul de la programmation linéaire, est maintenant $\binom{RC}{\tilde{n}}$. Ce nombre peut encore être très élevé, toutefois, et peut être réduit encore davantage par un choix approprié de la fonction de perte $w(s)$, comme nous le verrons dans la section suivante.

Dans l'exemple du tableau 1, on ramènerait la situation à celle représentée au tableau 2, tout en ne permettant que des tailles de cellule égales à 0 ou 1, et en ajoutant ensuite 1 aux cellules (1,1), (3,3), (4,2) et (5,1) dans la solution finale. Ainsi, $n = 10$, mais $\tilde{n} = 6$.

Tableau 2

Tableau des valeurs r_{ij}, d'après le tableau 1, avec $\tilde{n} = 6$

Région	Type de localité			
	Urbaine	Rurale	Métropolitaine	Total
1	0.0	0.5	0.5	1.0
2	0.2	0.3	0.5	1.0
3	0.2	0.6	0.2	1.0
4	0.6	0.8	0.6	2.0
5	0.0	0.8	0.2	1.0
Total	1.0	3.0	2.0	6.0

2.2 Choix de la fonction de perte *w(s)*

La méthode proposée est particulièrement souple du fait qu'elle laisse à l'utilisateur le choix de la fonction $w(s)$ incluse dans l'énoncé de la fonction objective en (2.2). La façon de procéder habituelle, dans le cas d'une stratification bidimensionnelle (p. ex. Jessen 1970; Causey et coll. 1985), consiste à exiger que l'échantillon prélevé s respecte les contraintes marginales:

$$| n_{i\cdot}(s) - nP_{i\cdot} | < 1 \quad i = 1, \ldots, R, \tag{2.7}$$

$$| n_{\cdot j}(s) - nP_{\cdot j} | < 1 \quad j = 1, \ldots, C, \tag{2.8}$$

où

$$n_{i\cdot}(s) = \sum_j n_{ij}(s), \quad n_{\cdot j}(s) = \sum_i n_{ij}(s)$$

$$P_{i\cdot} = \sum_j P_{ij}, \quad P_{\cdot j} = \sum_i P_{ij}.$$

Cette exigence peut être remplie, dans notre méthode, si l'on définit $w(s)$ comme étant (effectivement) infini pour les échantillons s ne satisfaisant pas à (2.7) ou à (2.8), ou plus simplement si l'on exclut ces échantillons de l'ensemble S_n. La difficulté que pose cette approche classique vient du fait qu'il n'existe peut-être pas de solution au problème d'optimisation avec contraintes.

Si, toutefois, nous utilisons dans notre méthode une fonction de perte comme

$$w(s) = \sum_{i=1}^{R} (n_{i\cdot}(s) - nP_{i\cdot})^2 + \sum_{j=1}^{C} (n_{\cdot j}(s) - nP_{\cdot j})^2, \tag{2.9}$$

il existera toujours une solution optimale dans un ensemble S_n suffisamment grand. Il peut être avantageux, du point de vue des besoins de calcul, de restreindre au départ l'ensemble S_n aux seuls échantillons respectant les contraintes (2.7) et (2.8), ou même à un sous-ensemble de ces derniers, puis d'élargir l'ensemble, si nécessaire, en remplaçant par exemple 1 par 2 dans (2.7) et (2.8), jusqu'à ce qu'une solution soit trouvée.

Examinons maintenant un aspect plus fondamental et voyons en quoi des contraintes comme (2.7) et (2.8) sont pertinentes. D'un point de vue non statistique, l'équilibrage d'un échantillon par rapport à des facteurs ayant une distribution connue pour l'ensemble de la population peut rassurer les utilisateurs quant à la "représentativité" de l'échantillon. D'un point de vue statistique, compte tenu de notre contrainte d'une estimation sans biais (2.1), il est naturel de se demander comment choisir la fonction de perte de façon à améliorer l'efficacité. On peut examiner cette question en posant $w(s)$ comme étant l'erreur quadratique moyenne $E_m(\bar{y}(s) - \bar{Y})^2$ en vertu d'un modèle m. Dans ce cas, la solution du problème d'optimisation (2.2) minimise l'espérance, selon le plan, de l'erreur quadratique moyenne liée au modèle ou, de façon équivalente puisque nous exigeons une estimation non biaisée selon le plan, l'espérance, selon le modèle, de la variance liée au plan.

Considérons, par exemple, le modèle d'analyse de variance des effets principaux

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},$$

où y_{ijk} est la $k$ième valeur de Y dans la cellule ij, μ est une moyenne fixe et α_i, β_j et ϵ_{ijk} sont des effets aléatoires indépendants de moyenne zéro et de variances σ_α^2, σ_β^2 et σ_ϵ^2 respectivement. On a alors, sans tenir compte des termes de correction pour population finie,

$$E_m(\bar{y}(s) - \bar{Y})^2 = \sigma_\alpha^2 \sum_i (n_{i\cdot}(s)/n - P_{i\cdot})^2 + \sigma_\beta^2 \sum_j (n_{\cdot j}(s)/n - P_{\cdot j})^2 + \sigma_\epsilon^2/n. \tag{2.10}$$

Par conséquent, si $\sigma_\alpha^2 = \sigma_\beta^2$, la variance probable de $\bar{y}(s)$ liée au plan selon ce modèle est minimisée par le choix de la fonction de perte définie en (2.9). Autre possibilité: si l'on disposait d'information préalable sur le ratio probable entre la variance entre lignes et la variance entre colonnes, il pourrait être judicieux, pour des besoins d'efficacité, de modifier la fonction de perte décrite en (2.9) en multipliant le premier terme du côté droit de (2.9) par ce ratio estimatif.

Par contre, si l'on pressent que la valeur de Y sera assujettie à une forte interaction entre les facteurs lignes et les facteurs colonnes, il n'est peut-être pas approprié de tenter simplement d'équilibrer l'échantillon selon les marges. Par exemple, si l'un des facteur de stratification est "urbain/rural" et que l'autre est un indicateur économique X, et que l'on sait que Y a un lien positif avec X dans les régions urbaines et un lien négatif dans les régions rurales, il sera sans doute plus efficace de stratifier partiellement selon X *séparément* pour les régions rurales et les régions urbaines, que d'équilibrer complètement selon les deux marges. Voir Bryant et coll. (1960, section 9) pour des commentaires connexes sur l'efficacité dans le cas d'une stratification bidimensionnelle.

2.3 Stratifications de dimensions supérieures

On peut, de façon naturelle, étendre la méthode proposée à trois facteurs de stratification ou plus, en définissant s comme étant la matrice à r dimensions correspondante. La fonction de perte comprendra généralement des termes additionnels; par exemple, pour une stratification à trois dimensions, nous pourrions poser

$$w(s) = \lambda_1 \sum_{i=1}^{R_1} (n_{i..}(s) - nP_{i..})^2 + \lambda_2 \sum_{j=1}^{R_2} (n_{.j.}(s) - nP_{.j.})^2 + \lambda_3 \sum_{k=1}^{R_3} (n_{..k}(s) - nP_{..k})^2$$

selon une notation évidente, où λ_1, λ_2 et λ_3 sont inclus pour représenter l'importance relative d'un équilibrage selon chacun des trois facteurs, et pourraient être des estimations préalables des variances des moyennes de Y entre les catégories des trois facteurs de stratification, comme en (2.10).

2.4 Échantillonnage à plusieurs degrés

Une importante application pratique de la stratification multidimensionnelle consiste à sélectionner les unités primaires d'échantillonnage (upé) d'un échantillon à plusieurs degrés, car il est fréquent qu'on dispose d'information sur plusieurs variables de stratification.

Dans la méthode exposée à la section 2.1, les probabilités de sélection de chaque unité de la population sont $E(n_{ij}(s)/N_{ij}) = n/N$. Si l'on désire choisir des upé avec probabilité égale, on peut appliquer immédiatement cette méthode en considérant les upé comme les unités et en remplaçant les valeurs observées de Y par des estimateurs sans biais des totaux pour les upé. Supposons toutefois que nous souhaitions choisir des upé avec probabilités inégales, par exemple nz_{ijk} pour l'upé k dans la cellule ij, où généralement z_{ijk} sera égal à $M_{ijk}/\sum_{ijk} M_{ijk}$, M_{ijk} étant une quelconque mesure de la taille de l'upé k dans la cellule ij. On peut alors modifier simplement le processus, en posant P_{ij} égal à la somme des z_{ijk} pour les upé k de la cellule ij. Alors, si $n_{ij}(s) > 0$, un échantillon de $n_{ij}(s)$ upé est prélevé dans la cellule ij selon une méthode dans laquelle la probabilité est proportionnelle à z_{ijk}.

3. EXEMPLES

Exemple 1: Bryant, Hartley et Jessen (1960)

Nous allons d'abord appliquer la méthode à l'exemple hypothétique de Bryant et coll. (1960) donné au tableau 1. Nous commençons par ramener le problème à la forme décrite en (2.4), (2.5) et (2.6), avec les r_{ij} donnés au tableau 2. La fonction de poids présentée en (2.9) devient, dans ce problème réduit de programmation linéaire,

$$w(s) = \sum_{i=1}^{5} (\tilde{n}_{i.}(\tilde{s}) - r_{i.})^2 + \sum_{j=1}^{3} (\tilde{n}_{.j}(\tilde{s}) - r_{.j})^2.$$

À l'aide d'un logiciel courant de programmation linéaire de la bibliothèque NAG FORTRAN, nous obtenons la solution donnée au tableau 3. Les valeurs I_{ij} ont été ajoutées à la solution de telle façon que $n_{ij} = I_{ij} + \tilde{n}_{ij}(\tilde{s})$. Il se révèle, pour cette solution, que chaque s pour lequel $p(s) > 0$, a des marges $n_{i.}(s)$ et $n_{.j}(s)$ qui correspondent exactement aux marges désirées, c'est-à-dire que la solution rend (2.4) égal à zéro.

Tableau 3

Solution de l'exemple 1

s	$p(s)$	s	$p(s)$
1 1 0		1 1 0	
1 0 0		0 0 1	
0 1 1	0.2	0 1 1	0.1
0 2 1		1 1 1	
1 0 1		1 1 0	
1 1 0		1 0 1	
0 0 1		0 1 0	
0 0 2	0.2	1 0 1	0.2
1 2 0		0 2 1	
1 1 0		1 1 0	
1 0 1		1 0 1	
0 1 0		0 0 1	
0 1 1	0.1	0 1 1	0.2
1 1 1		1 2 0	
1 1 0		1 1 0	

Exemple 2: Jessen (1970)

Jessen (1970) a proposé deux méthodes, pour la stratification bidimensionnelle et la stratification tridimensionnelle. Les deux méthodes, très complexes, obligent à trouver l'ensemble d'échantillons qui correspondent exactement aux marges. Aucune d'elle ne donne nécessairement une solution. Jessen (1970) applique les deux méthodes à un exemple hypothétique simple pour lequel chacune produit une solution. Cet exemple est présenté au tableau 4. Dans ce cas, puisque tous les $nP_{ij} < 1$, les problèmes de programmation linéaire définis respectivement par (2.1), (2.2) et (2.3) et par (2.4), (2.5) et (2.6) sont identiques. Nous avons appliqué notre méthode à ce problème, en utilisant encore la fonction $w(s)$ définie en (2.9). En essayant plusieurs valeurs de départ dans le programme d'optimisation, nous avons pu obtenir trois solutions différentes, chacune rendant (2.2) égal à zéro et satisfaisant aux contraintes. Ces solutions sont présentées au tableau 5. Les deux premières sont les mêmes que celles obtenues respectivement par la méthode 2 et par la méthode 3 de Jessen.

Tableau 4

Exemple 2: Jessen (1970)

Nombres probables d'unités de l'échantillon dans les cellules en vertu d'une stratification proportionnelle avec $n = 6$

Lignes	Colonnes			$nP_{i\cdot}$
	1	2	3	
1	0.8	0.5	0.7	2.0
2	0.7	0.8	0.5	2.0
3	0.5	0.7	0.8	2.0
$nP_{\cdot j}$	2.0	2.0	2.0	6.0

Tableau 5

Solution de l'exemple 2

s	$p_1(s)$	$p_2(s)$	$p_3(s)$
1 0 1 1 1 0 0 1 1	0.5	0.4	0.3
1 1 0 0 1 1 1 0 1	0.3	0.2	0.1
0 1 1 1 0 1 1 1 0	0.2	0.1	0.0
1 1 0 1 0 1 0 1 1	0.0	0.1	0.2
1 0 1 0 1 1 1 1 0	0.0	0.1	0.2
0 1 1 1 1 0 1 0 1	0.0	0.1	0.2

Exemple 3: Causey, Cox et Ernst (1985)

Causey et coll. (1985) donnent un exemple d'une stratification tridimensionnelle pour laquelle leur méthode est incapable de donner une solution. Ils examinent une population soumise à une stratification $2 \times 2 \times 2$ de laquelle un échantillon de taille $n = 2$ doit être tiré; les valeurs suivantes sont données pour la taille attendue de l'échantillon dans la ijkième cellule, n_{ijk}:

$$n_{111} = n_{221} = n_{122} = n_{212} = .5$$

$$n_{121} = n_{211} = n_{112} = n_{222} = 0.$$

Si nous appliquons notre méthode d'une manière semblable à celle employée dans les exemples 1 et 2, nous obtenons la solution donnée au tableau 6. Dans ce cas, la fonction objective n'a pas atteint zéro, de sorte qu'il n'y a pas une correspondance exacte avec les marges dans chaque échantillon.

Tableau 6

Solution de l'exemple 3

s				$p(s)$
$i = 1$		$i = 2$		
1 0 0 0		0 1 0 0		0.5
0 0 0 1		0 0 1 0		0.5

4. COMPARAISON DES ERREURS QUADRATIQUES MOYENNES

Dans la présente section, l'erreur quadratique moyenne (EQM) du plan proposé avec estimateur $\bar{y}$ sera comparée à celle du plan de Bryant et coll. (1960) pour chacun des deux estimateurs qu'ils proposent, soit $\bar{y}_U$ et $\bar{y}_B$, où les indices U et B indiquent que le premier estimateur est non biaisé et que le second est biaisé. Désignons les cellules par c (ij dans le cas bidimensionnel), désignons une unité d'une cellule par k (et si nécessaire l), et supprimons le s dans $n_c(s)$, pour simplifier la notation. La probabilité de sélection de n'importe quelle unité k de la cellule c est

$$\pi_{ck} = E[n_c]/N_c = E[n_c]/(NP_c) \qquad (4.1)$$

et la probabilité de sélection conjointe de l'unité k de la cellule c et de l'unité k' de la cellule c' est

$$\pi_{ckc'k'} = \begin{cases} \dfrac{E[n_c(n_c-1)]}{N_c(N_c-1)} & \text{si } c = c' \\ \dfrac{E[n_c n_{c'}]}{N_c N_{c'}} & \text{si } c \neq c'. \end{cases} \qquad (4.2)$$

Si N est grand, cela équivaut approximativement à

$$\pi_{ckc'k'} \doteq \frac{E(n_c n_{c'})}{N^2 P_c P_{c'}} - \frac{E(n_c)}{N^2 P_c^2} I_{[c=c']}, \qquad (4.3)$$

où

$$I_{[c=c']} = \begin{cases} 1 & \text{si } c = c' \\ 0 & \text{si } c \neq c'. \end{cases}$$

Les espérances de notre plan ne seront pas les mêmes que celles du plan de Bryant et coll., de sorte que les π_{ck} et les $\pi_{ckc'k'}$ seront différents. En gardant ce fait à l'esprit, nous pouvons obtenir la variance de $\bar{y}$, $\bar{y}_U$ et $\bar{y}_B$ sous une forme généralisée en termes des valeurs π_{ck} et $\pi_{ckc'k'}$, et donc disposer d'une base de comparaison. Pour ce faire, prenons un estimateur de la forme $\bar{z} = \sum_c \sum_k w_c y_{ck}/n$, où les valeurs w_c sont des constantes connues fixes, indépendantes de k. Si nous nous limitons au cas où $n_{i\cdot} = nP_{i\cdot}$ et $n_{\cdot j} = nP_{\cdot j}$, c'est-à-dire que les valeurs marginales sont entières, les estimateurs donnés par Bryant et coll. ainsi que notre estimateur sont tous de cette forme. Nous supposerons dorénavant que c'est le cas. Après remplacement de l'indice c par ij pour la stratification bidimensionnelle, $\bar{y}_U$ et $\bar{y}_B$ sont de la même forme que $\bar{z}$ avec $w_c = w_{ij} = G_{ij} = P_{ij}/(P_{i\cdot}P_{\cdot j})$ et $w_c = w_{ij} = 1$ respectivement. L'estimateur $\bar{y}$ est également de la même forme que $\bar{z}$, avec $w_c = w_{ij} = 1$.

Nous pouvons maintenant obtenir une forme générale pour la variance de $\bar{z}$, en nous rappelant que les π_{ck} et les $\pi_{ckc'k'}$ de notre plan ne seront pas les mêmes que celles du plan de Bryant et coll.:

$$V(\bar{z}) = \frac{1}{2n^2} \sum_c \sum_{c'} \sum_k \sum_{k'} (\pi_{ck}\pi_{c'k'} - \pi_{ckc'k'}) (w_c y_{ck} - w_{c'} y_{c'k'})^2. \qquad (4.4)$$

En utilisant (4.1) et (4.3), on obtient

$$V(\bar{z}) = \frac{1}{2n^2} \sum_c \frac{w_c^2 E(n_c)}{N^2 P_c^2} \sum_k \sum_{k'} (y_{ck} - y_{ck'})^2 - \frac{1}{2n^2} \sum_c \sum_{c'} \frac{\mathrm{Cov}(n_c, n_{c'})}{N^2 P_c P_{c'}} \sum_k \sum_{k'} (w_c y_{ck} - w_{c'} y_{c'k'})^2. \qquad (4.5)$$

Notons que

$$\sum_k \sum_l (y_{ck} - y_{cl})^2 = 2N^2 P_c^2 S_c^2$$

et que

$$\sum_k \sum_{k'} (w_c y_{ck} - w_{c'} y_{c'k'})^2 = N^2 P_c P_{c'} [w_c^2 S_c^2 + w_{c'}^2 S_{c'}^2 + (w_c \bar{Y}_c - w_{c'} \bar{Y}_{c'})^2],$$

où S_c^2 désigne la variance de la population de la cellule c. L'équation (4.5) se réduit alors à

$$V(\bar{z}) = \frac{1}{n^2} \sum_c w_c^2 E(n_c) S_c^2 - \frac{1}{2n^2} \sum_c \sum_{c'} \mathrm{Cov}(n_c, n_{c'}) [w_c^2 S_c^2 + w_{c'}^2 S_{c'}^2 + (w_c \bar{Y}_c - w_{c'} \bar{Y}_{c'})^2]$$

$$= v_1 + v_2, \quad \text{par exemple.} \qquad (4.6)$$

Le premier terme, v_1, peut être interprété comme la variance stratifiée habituelle pour les tailles d'échantillon fixes $E(n_c)$ à l'intérieur des "strates" bidimensionnelles (évidemment, dans notre cas, les $E(n_c)$ ne seront pas généralement des entiers). Le second terme, v_2, peut être interprété comme l'accroissement de variance attribuable à la variabilité des n_c et à la corrélation qui existe entre eux. Nous y reviendrons à la fin de la présente section. Nous revenons maintenant à la notation $c = ij$ et comparons les variances pour la stratification bidimensionnelle.

Examinons d'abord v_1 dans (4.6). Pour la méthode de Bryant et coll. $E(n_{ij}) = nP_{i\cdot}P_{\cdot j}$, $\bar{y}_U = \sum_i \sum_j \sum_k G_{ij} y_{ijk}/n$, $G_{ij} = P_{ij}/(P_{i\cdot}P_{\cdot j})$ et $\bar{y}_B = \sum_i \sum_j \sum_k y_{ijk}/n$.

Ainsi,

$$v_1(\bar{y}_U) = \sum_i \sum_j P_{ij} G_{ij} S_{ij}^2/n,$$

(ce qui est identique au premier terme de l'équation (12) de Bryant et coll.) et

$$v_1(\bar{y}_B) = \sum_i \sum_j P_{i\cdot}P_{\cdot j} S_{ij}^2/n.$$

Dans le cas de notre méthode, $E(n_{ij}) = nP_{ij}$ et $\bar{y} = \sum_i \sum_j \sum_k y_{ijk}/n$, de sorte que

$$v_1(\bar{y}) = \sum_i \sum_j P_{ij} S_{ij}^2/n.$$

Examinons maintenant v_2. Il n'est pas difficile de montrer qu'aussi bien pour la méthode de Bryant et coll. que pour la nôtre (voir l'annexe)

$$\sum_i \mathrm{Cov}(n_{ij}, n_{i'j'}) = \sum_j \mathrm{Cov}(n_{ij}, n_{i'j'}) = 0. \qquad (4.7)$$

En tenant compte de ce fait et en remplaçant c et c' par ij et $i'j'$ respectivement dans l'expression v_2 donnée en (4.6), nous réduisons v_2 à

$$v_2 = \frac{1}{n^2} \sum_i \sum_j \sum_{i'} \sum_{j'} \mathrm{Cov}(n_{ij}, n_{i'j'}) w_{ij} w_{i'j'} \bar{Y}_{ij} \bar{Y}_{i'j'}.$$

En remplaçant w_{ij} par G_{ij}, on obtient $v_2(\bar{y}_U)$, et par des manipulations algébriques simples, on peut montrer qu'il s'agit de l'équivalent du terme 2 de l'équation (12) de Bryant et coll. Si l'on remplace w_{ij} par 1, on obtient la forme de $V(\bar{y}_B)$ et de $V(\bar{y})$, en notant que $\text{Cov}(n_{ij}, n_{i'j'})$ ne sera pas la même pour les deux. On peut donc constater que v_2 dépend seulement des moyennes des cellules, tandis que v_1 dépend seulement des variances intra-cellules.

Enfin, il est utile de noter que

$$\text{biais}(\bar{y}_B) = -\sum_i \sum_j (P_{ij} - P_{i\cdot}P_{\cdot j})\bar{Y}_{ij}, \qquad (4.8)$$

car pour comparer les trois estimateurs, l'erreur quadratique moyenne (EQM) sera la mesure pertinente, et ce biais contribuera à $\text{EQM}(\bar{y}_B)$.

La combinaison des expressions de v_1, v_2 et $\text{biais}(\bar{y}_B)$ ci-dessus permet de faire une comparaison analytique de l'erreur quadratique moyenne de la méthode proposée et de celle de la méthode de Bryant et coll. (1960), en utilisant soit $\bar{y}_U$, soit $\bar{y}_B$. Il est difficile, toutefois, de faire des énoncés généraux sur la performance relative des différentes stratégies. C'est pourquoi nous introduisons maintenant certaines hypothèses de modèles, dont nous tirerons des approximations des différentes composantes des expressions de l'erreur quadratique moyenne dans certaines conditions précises. Nous examinons d'abord le modèle additif:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},$$

où y_{ijk} est la $k$ième observation dans la ijième cellule, α_i et β_j sont des effets fixes et les ϵ_{ijk} sont des erreurs indépendantes de moyenne zéro et variance commune σ^2. On a alors $\text{E}_m(S_{ij}^2) = \sigma^2$ et $\text{E}_m(\bar{Y}_{ij}\bar{Y}_{i'j'}) \doteq (\mu + \alpha_i + \beta_j)(\mu + \alpha_{i'} + \beta_{j'})$. Donc, l'espérance, selon le modèle, de la variance liée au plan est obtenue en remplaçant S_{ij}^2 par σ^2 et $\bar{Y}_{ij}$ par $\mu + \alpha_i + \beta_j$ dans les formules de v_1 et v_2, pour les divers estimateurs. Dans ce cas, $v_2(\bar{y}_B) = 0$. Ce fait a été constaté par Bryant et coll., dans la comparaison de $\bar{y}_U$ et de $\bar{y}_B$. Le terme de biais, dans ce cas, sera nul, à moins qu'on ait arrondi les valeurs des marges, c'est-à-dire que $\text{biais}(\bar{y}_B) = 0$ pourvu que $n_{i\cdot} = nP_{i\cdot}$ et $n_{\cdot j} = nP_{\cdot j}$, comme c'est le cas dans leur exemple. Cela découle directement de (4.8) et

$$\sum_i (P_{ij} - P_{i\cdot}P_{\cdot j}) = \sum_j (P_{ij} - P_{i\cdot}P_{\cdot j}) = 0.$$

Ce résultat a également été démontré par Bryant et coll., à la page 119, équation (47). En utilisant (4.7), on peut montrer facilement que $v_2(\bar{y}) = 0$ également. Cette constatation, ainsi que le fait que $\bar{y}$ soit non biaisé et que $v_1(\bar{y}_B) = v_1(\bar{y}) = \sigma^2/n$ dans ce cas, a pour conséquence que $\text{EQM}(\bar{y}_B) = \text{EQM}(\bar{y})$ dans cette situation, c'est-à-dire que la méthode proposée a la même erreur quadratique moyenne que celle de Bryant et coll. utilisant l'estimateur biaisé. Nous démontrons dans ce qui suit que même quand ce modèle additif s'applique ($\gamma = 0$ ci-dessous), $v_2(\bar{y}_U)$ peut être élevé, tandis que $v_1(\bar{y}_U) > v_1(\bar{y})$.

Pour comparer plus à fond les estimateurs, examinons la situation de l'exemple 1. Le développement ci-dessus nous permet d'obtenir les erreurs quadratiques moyennes des trois estimateurs de cet exemple, pourvu que nous connaissions les S_{ij} et les $\bar{Y}_{ij}$ et que nous puissions calculer les $\text{Cov}(n_{ij}, n_{i'j'})$ pour la méthode de Bryant et coll. ainsi que pour notre méthode. Les covariances, pour la méthode de Bryant et coll., sont données dans leur article en fonction des P_{ij}, tandis que celles de notre méthode peuvent être tirées de la solution du tableau 3. Nous allons considérer des variantes non additives du modèle ci-dessus, c'est-à-dire

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma\alpha_i\beta_j + \epsilon_{ijk},$$

pour diverses valeurs de γ. Pour simplifier la présentation, posons $\mu = 1$, $\alpha_i = i - 3$, $\beta_j = j - 2$ (notons qu'en fait, l'erreur quadratique moyenne de chaque stratégie ne varie pas en fonction du choix de μ). Donc, on obtient l'espérance, selon le modèle, de la variance liée au plan en remplaçant S_{ij}^2 par 1 et $\bar{Y}_{ij}$ par $1 + (i - 3) + (j - 2) + \gamma(i - 3)(j - 2)$ dans les formules donnant v_1 et v_2 pour les divers estimateurs. Le tableau 7 donne les valeurs résultantes de v_1, de v_2 et de l'erreur quadratique moyenne (EQM) pour les trois estimateurs (ainsi que le terme quadratique de biais pour $\bar{y}_B$), pour diverses valeurs de γ. On peut conclure de l'examen du tableau 7 que dans le cas d'un modèle additif, $\gamma = 0$, $\bar{y}_B$ et $\bar{y}$ ont une performance identique, tandis que $\bar{y}_U$ est moins performant. À mesure que le caractère non additif du modèle s'accroît et que $|\gamma|$ augmente, les deux estimateurs de la stratégie de Bryant

Tableau 7

Comparaison des erreurs quadratiques moyennes (EQM) pour trois estimateurs

	Plan de Bryant, Hartley et Jessen							Plan proposé		
	$\bar{y}_U$			$\bar{y}_B$				$\bar{y}$		
γ	v_1	v_2	EQM	v_1	v_2	Biais^2	EQM	v_1	v_2	EQM
0	.125	.105	.230	.100	.000	.000	.100	.100	.000	.100
±.5	.125	.063	.188	.100	.033	.002	.135	.100	.018	.118
±1	.125	.105	.230	.100	.131	.008	.239	.100	.071	.171
±2	.125	.440	.565	.100	.523	.032	.655	.100	.284	.384
±3	.125	1.111	1.236	.100	1.176	.073	1.349	.100	.638	.738

et coll. tendent à avoir une performance identique et, pour chacun d'eux, l'erreur quadratique moyenne s'accroît progressivement par rapport à celle de la stratégie proposée. Cette tendance est principalement attribuable à la composante v_2 de l'erreur quadratique moyenne des trois estimateurs. Le terme de biais de $\bar{y}_B$ a une importance moindre, bien que son influence puisse s'accroître avec l'augmentation de la taille n.

Le fait que v_2, dans le plan de Bryant et coll., s'accroisse plus rapidement avec l'augmentation de $|\gamma|$ semble attribuable à la plus grande variabilité de chaque n_{ij} dans ce plan. Notons qu'il aurait été possible de réduire quelque peu cette variabilité en appliquant plutôt au tableau 2 une variante de la méthode de Bryant et coll., comme nous l'avons fait pour la méthode proposée, mais il aurait fallu calculer des poids G_{ij} rajustés pour $\bar{y}_U$, et il serait difficile de traiter les valeurs 0.0 des cellules du tableau 2. Toutefois, même après avoir fait cela, les $\tilde{n}_{ij}$ pourraient toujours, dans ce plan, prendre des valeurs autres que 0 et 1; par exemple, n_{42} pourrait prendre les valeurs 0, 1 ou 2. Cette variabilité supérieure des n_c est inhérente à la méthode de Bryant et coll. Par exemple, supposons que $n_{1\cdot} = n_{\cdot 1} = 5$. Selon la méthode de Bryant et coll., n_{11} peut prendre les valeurs 0, 1, 2, 3, 4 ou 5, tandis que selon la méthode proposée, il ne peut prendre que les valeurs $[nP_{11}]$ ou $[nP_{11}] + 1$. Si $nP_{11} < 1$, la technique utilisée pour passer du tableau 1 au tableau 2 n'améliorera pas les choses.

5. ESTIMATION DE LA VARIANCE

Dans la présente section, nous allons nous pencher sur l'estimation de la variance pour la méthode que nous proposons. En utilisant (4.1) et en nous rappelant de la contrainte (2.1), nous pouvons évidemment écrire

$$\pi_{ck} = E[n_c(s)/N_c] = n/N.$$

La probabilité de sélection conjointe de deux unités, k et k', de la même cellule c est

$$\pi_{ck,ck'} = E[n_c(s)\{n_c(s) - 1\}/\{N_c(N_c - 1)\}].$$

Supposons que $n_c(s) = I_c + \tilde{n}_c(s)$ où I_c est l'entier fixe $[nP_c]$ et $\tilde{n}_c(s) = 0$ ou 1.

Si $nP_c \leq 1$, alors $I_c = 0$ et $\pi_{ck,ck'} = 0$. Par conséquent, une condition nécessaire pour qu'une estimation non biaisée de la variance soit possible est que $nP_c > 1$ pour toutes les cellules c. D'autre part, si cette condition et vérifiée, il en résulte que $n_c(s) \geq 1$ pour toutes les cellules c, de sorte que la probabilité de sélection de toute paire d'unités de cellules différentes est aussi toujours positive. Bref, cette condition est nécessaire et suffisante pour qu'une estimation non biaisée de la variance soit possible.

Quand cette condition est vérifiée, nous obtenons

$$\pi_{ck,ck'} = I_c(I_c + 2r_c - 1)/[N_c(N_c - 1)] = A_c,$$

par exemple, où $r_c = E[\tilde{n}_c(s)] = nP_c - I_c$.

Les probabilités de sélection conjointe de paires d'unités de cellules différentes c et c' sont

$$\pi_{ck,ck'} = E[n_c(s)n_{c'}(s)/(N_cN_{c'})]$$
$$= [I_cI_{c'} + r_{c'}I_c + r_cI_{c'} + r_{cc'}]/(N_cN_{c'}) = B_{cc'}, \tag{5.1}$$

par exemple, où $r_{cc'} = E[\tilde{n}_c(s)\tilde{n}_{c'}(s)]$.

Par conséquent, un estimateur non biaisé de $V(\bar{y}(s))$, de la forme Sen-Yates-Grundy, peut être construit de la manière habituelle.

En pratique, toutefois, nous voulons tenir compte de situations où $nP_c \leq 1$ pour certaines cellules c. Dans ce cas, une hypothèse que nous pourrions faire, comme l'ont fait Bryant et coll. (1960, section 7), afin de trouver un estimateur de la variance, c'est de considérer que la variance de Y pour l'ensemble de la population dans chaque cellule c est une valeur constante, disons S^2.

Obtenons d'abord la variance de $\bar{y}(s)$ dans le cas général

$$V(\bar{y}(s)) = \frac{1}{2n^2}\sum_c\sum_{k\neq k'}\sum\left(\frac{n^2}{N^2} - A_c\right)(y_{ck} - y_{ck'})^2$$
$$+ \frac{1}{2n^2}\sum_{c\neq c'}\sum\sum_{k,k'}\left(\frac{n^2}{N^2} - B_{cc'}\right)(y_{ck} - y_{c'k'})^2.$$

Pourvu que $B_{cc'} > 0\ \forall\ c, c'$, nous pouvons estimer sans biais le second terme sous la forme

$$\frac{1}{2n^2}\sum_A\sum\sum_{k=1}^{n_c(s)}\sum_{k'=1}^{n_{c'}(s)}\left(\frac{\frac{n^2}{N^2} - B_{cc'}}{B_{cc'}}\right)(y_{ck} - y_{c'k'})^2,$$

où $A = \{c, c' : n_c(s) \geq 1, n_{c'}(s) \geq 1, c \neq c'\}$.

Le premier terme peut s'écrire ainsi

$$\frac{1}{2n^2}\sum_c\left(\frac{n^2}{N^2} - A_c\right)2N_c^2S^2.$$

Pour tout c tel que $n_c(s) \geq 2$

$$E\left(\sum_{\substack{k=1\\k\neq k'}}^{n_c(s)}\sum_{k'=1}^{n_c(s)}\frac{(y_{ck} - y_{ck'})^2}{2n_c(s)\{n_c(s) - 1\}}\,\middle|\, n_c(s)\right) = S^2.$$

Donc, pourvu qu'au moins un $n_c(s)$ soit ≥ 2, un estimateur non biaisé du premier terme est donné par

$$\frac{1}{2n^2D} \sum_{\{c:n_c(s)\geq 2\}} \left(\frac{n^2}{N^2} - A_c\right) 2N_c^2 \sum_{k=1}^{n_c(s)} \sum_{k'=1}^{n_c'(s)} \frac{(y_{ck} - y_{ck'})^2}{2n_c(s)\{n_c(s) - 1\}}$$

où D = le nombre de cellules c telles que $n_c(s) \geq 2$.

Le raisonnement ci-dessus exige que $B_{cc'} > 0$. Si

$$I_c = I_{c'} = 0,$$

alors, d'après (5.1), il faut que

$$r_{cc'} = \sum \tilde{n}_c(s)\tilde{n}_{c'}(s)p(s) > 0, \qquad (5.2)$$

qui est linéaire en $p(s)$. La contrainte (5.2) peut être traitée par programmation linéaire, si désiré. Une telle contrainte existera pour chaque paire c, c' telle que $I_c = I_{c'} = 0$.

6. CONCLUSION

Nous avons proposé une méthode fondée sur la programmation linéaire pour le traitement de stratifications multidimensionnelles, en nous appuyant sur les idées de Rao et Nigam (1990, 1992). Il s'agit d'une méthode simple, qui offre beaucoup de souplesse en permettant de choisir toute une gamme d'objectifs différents par l'entremise de la fonction de perte $w(s)$, et en permettant tout un éventail de contraintes, comme le fait d'exiger que les probabilités de sélection conjointes de toutes les cellules soient positives. Le principal obstacle pratique à l'application de la méthode est que son traitement informatique peut rapidement devenir coûteux, voire impossible, à mesure qu'augmente le nombre de cellules de la stratification multidimensionnelle. Certaines suggestions ont été faites en vue de réduire l'ampleur des calculs. Il serait utile que d'autres recherches soient entreprises dans ce domaine. Pour les cas où les exigences du traitement informatique seraient prohibitives, la méthode de Causey et coll. (1985) demeure une solution de rechange.

REMERCIEMENTS

Les auteurs tiennent à remercier Wesley Yung pour son aide touchant la programmation informatique, ainsi que J.N.K. Rao pour ses précieux commentaires aux premiers stades du travail. R.R. Sitter est subventionné par le Conseil de recherches en sciences naturelles et en génie du Canada. Nous exprimons nos remerciements aux arbitres et au rédacteur associé pour leurs suggestions et leurs commentaires utiles.

ANNEXE

Preuve de (4.7) pour la méthode proposée

Notons que

$$\mathrm{Cov}(n_{ij}(s), n_{i'j'}(s)) = \mathrm{E}(n_{ij}(s)n_{i'j'}(s)) - \mathrm{E}(n_{ij}(s))\mathrm{E}(n_{i'j'}(s)).$$

L'équation (2.1) indique que $\mathrm{E}(n_{ij}(s)) = nP_{ij}$. Par définition

$$\mathrm{E}(n_{ij}(s)n_{i'j'}(s)) = \sum_s n_{ij}(s)n_{i'j'}(s)p(s).$$

Donc

$$\sum_j \mathrm{E}(n_{ij}(s))\mathrm{E}(n_{i'j'}(s)) = n^2P_{i'j'} \sum_j P_{ij} = n^2P_{i'j'}P_{i\cdot}, \qquad (7.1)$$

et

$$\sum_j \mathrm{E}(n_{ij}(s)n_{i'j'}(s)) = \sum_j \sum_s n_{ij}(s)n_{i'j'}(s)p(s) = \sum_s p(s)n_{i'j'}(s) \sum_j n_{ij}(s). \qquad (7.2)$$

Supposons que la solution du problème d'optimisation linéaire (2.2) soit égale à zéro, avec $w(s)$ donné par (2.9). Dans ce cas, $\sum_j n_{ij}(s) = n_{i\cdot}(s) = nP_{i\cdot}$ et il résulte de (7.2) que

$$\sum_j \mathrm{E}(n_{ij}(s)n_{i'j'}(s)) = \sum_s p(s)n_{i'j'}(s)nP_{i\cdot} = nP_{i\cdot} \sum_s n_{i'j'}(s)p(s) = nP_{i\cdot}\mathrm{E}(n_{i'j'}(s)) = nP_{i\cdot}nP_{i'j'}. \qquad (7.3)$$

Les équations (7.1) et (7.3), ensemble, font en sorte que $\sum_j \mathrm{Cov}(n_{ij}(s), n_{i'j'}(s)) = 0$. On peut montrer de la même manière que

$$\sum_i \mathrm{Cov}(n_{ij}(s), n_{i'j'}(s)) = \sum_{i'} \mathrm{Cov}(n_{ij}(s), n_{i'j'}(s)) = \sum_{j'} \mathrm{Cov}(n_{ij}(s), n_{i'j'}(s)) = 0.$$

BIBLIOGRAPHIE

BRYANT, E.C., HARTLEY, H.O., et JESSEN, R.J. (1960). Design and estimation in two-way stratification. *Journal of the American Statistical Association*, 55, 105-124.

CAUSEY, B.D., COX, L.H., et ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.

GOODMAN, R., et KISH, L. (1950). Controlled selection-a technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.

JESSEN, R.J. (1970). Probability sampling with marginal constraints. *Journal of the American Statistical Association*, 65, 776-795.

RAO, J.N.K., et NIGAM, A.K. (1990). Optimal controlled sampling design. *Biometrika*, 77, 807-814.

RAO, J.N.K., et NIGAM, A.K. (1992). 'Optimal' controlled sampling: a unified approach. *Revue Internationale de Statistique*, 60, 89-98.

WATERTON, J.J. (1983). An exercise in controlled selection. *Applied Statistics*, 32, 150-164.

Techniques d'enquête, juin 1994
Vol. 20, n° 1, pp. 79–89
Statistique Canada

Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988

WAYNE A. FULLER, MARIE M. LOUGHIN et HAROLD D. BAKER[1]

RÉSUMÉ

Les auteurs appliquent une méthode de production de poids de régression à la Nationwide Food Consumption Survey réalisée en 1987-1988 par le Département de l'agriculture des États-Unis. Ils ont eu recours à l'estimation par régression à cause du taux de non-réponse élevé dans l'enquête. Les poids de régression sont des poids établis par la méthode des moindres carrés généralisés, modifiés de telle manière qu'ils soient tous positifs et que les plus élevés d'entre eux soient plus petits que les poids établis par les moindres carrés. Les auteurs montrent qu'en situation de non-réponse l'estimateur par régression peut avoir une erreur quadratique moyenne beaucoup moins élevée que l'estimateur direct ordinaire.

MOTS CLÉS: Poids non négatifs; cohérence.

1. INTRODUCTION

Souvent, quand on procède à un échantillonnage, on connaît la valeur moyenne des variables auxiliaires pour la population mais on ne se sert pas, pour le prélèvement de l'échantillon, des valeurs de ces variables pour les unités de cette population. Même si cette information n'est pas utilisée dans le plan d'échantillonnage, il est très souhaitable d'intégrer les moyennes de population dans la procédure d'estimation. Les méthodes d'estimation courantes qui utilisent de l'information supplémentaire sont l'estimation par quotient, la stratification a posteriori, l'estimation par régression et l'estimation par la méthode itérative du quotient. L'estimation par régression est la méthode la plus générale en ce sens qu'elle peut traiter des variables auxiliaires multidimensionnelles, continues ou discrètes. La stratification a posteriori peut être considérée comme un cas particulier de l'estimation par régression, où les variables de régression sont des variables indicatrices pour les strates formées a posteriori. La méthode itérative du quotient, aussi appelée ajustement proportionnel itératif, est limitée à de l'information supplémentaire sous forme de catégories discrètes. Sur cette méthode itérative du quotient, on pourra lire Deming et Stephan (1940), Stephan (1942), El-Badry et Stephan (1955), Ireland et Kullback (1968), Darroch et Ratcliff (1972), Brackstone et Rao (1979) et Oh et Scheuren (1987).

Watson (1937), Cochran (1942) et Jessen (1942) sont les premiers ouvrages où l'on traite d'estimation par régression. La théorie élémentaire de cette forme d'estimation se trouve dans Cochran (1977, chap. 7). De nombreux auteurs se sont intéressés à l'estimation par régression appliquée à des échantillons d'enquête: Mickey (1959), Fuller (1975), Royall et Cumberland (1981), Isaki et Fuller (1982), Wright (1983), Luery (1986), Alexander (1987), Bethlehem et Keller (1987), Copeland, Peitzmeier et Hoy (1987), Lemaître et Dufour (1987), Särndal, Swensson et Wretman (1989), Deville et Särndal (1992), Zieschang (1990) et Rao (1992).

Beaucoup de ces ouvrages présentent l'estimation par régression comme une méthode permettant de réduire la variance dans les échantillons probabilistes. En réalité, un des avantages de l'estimation par régression est qu'elle peut réduire le biais causé par la non-réponse ponctuelle. Voir, par exemple, Little et Rubin (1987, p. 55), pour le cas particulier des cellules d'ajustement, et Bethlehem (1988), pour l'estimateur par régression généralisé.

Dans l'application décrite plus bas, la non-réponse nous amène à recourir à l'estimation par régression, et c'est cette expérience que nous décrivons dans la section 3. Dans la section 2, nous présentons l'estimateur par régression ordinaire ainsi qu'une version modifiée de celui-ci qui produit des poids positifs. Enfin, dans la section 4, nous appliquons la méthode de pondération à la Nationwide Food Consumption Survey.

2. ESTIMATEUR PAR RÉGRESSION

Afin de présenter l'estimateur par régression que nous allons utiliser dans cette étude, nous supposons qu'un échantillon de n unités est prélevé et que la probabilité de sélection pour l'unité i est π_i. Pour les besoins de notre exposé, nous posons comme condition suffisante que π_i soit proportionnelle aux probabilités de sélection. Il peut s'agir d'un échantillon stratifié à deux degrés et l'unité peut être soit l'unité primaire d'échantillonnage, soit l'unité d'observation. Dans notre application, il s'agit de l'unité d'observation. Supposons qu'un vecteur de moyennes de population de dimension k, désigné par $\bar{X} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k)$, est connu, qu'un vecteur $(y_i, x_{i1}, x_{i2}, \ldots, x_{ik})$ est observé pour chaque unité de l'échantillon et qu'on

[1] Wayne A. Fuller, Marie M. Loughin et Harold D. Baker, Iowa State University.

cherche à obtenir un estimateur de la moyenne de y. Nous supposons que le premier élément de x_i est égal à un pour tous les i. Par conséquent, le premier élément de $\bar{X}$ est aussi égal à un. Le vecteur $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ est parfois appelé vecteur des variables de contrôle. Un estimateur par régression de la moyenne de y est

$$\bar{y}_r = \bar{X}\hat{\beta}, \tag{2.1}$$

où

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i' \pi_i^{-1} x_i \right)^{-1} \sum_{i=1}^{n} x_i' \pi_i^{-1} y_i, \tag{2.2}$$

$\sum x_i' \pi_i^{-1} x_i$ étant supposée régulière. Cette définition de l'estimateur par régression est conforme à celle donnée par Mickey (1959), qui proposait de limiter l'usage du terme "estimateur par régression" aux estimateurs qui sont invariants en position et en échelle. L'estimateur (2.1) peut aussi s'écrire sous la forme

$$\bar{y}_r = \sum_{i=1}^{n} w_i y_i, \tag{2.3}$$

où

$$w_i = \bar{X}\left(\sum_{i=1}^{n} x_i' \pi_i^{-1} x_i \right)^{-1} x_i' \pi_i^{-1}, \tag{2.4}$$

les poids ayant la propriété suivante:

$$\sum_{i=1}^{n} w_i x_i = \bar{X}. \tag{2.5}$$

Les poids définis en (2.4) sont assez faciles à calculer et, une fois qu'ils l'ont été, ils peuvent servir à l'estimation de n'importe quelle variable y. Si on remplace le vecteur x_j par le vecteur

$$(1, z_j) = (1, x_{j2} - \bar{X}_2, x_{j3} - \bar{X}_3, \ldots, x_{jk} - \bar{X}_k), \tag{2.6}$$

on peut exprimer l'estimateur sous la forme

$$\bar{y}_r = \bar{y}_\pi + (\bar{Z} - \bar{z}_\pi)\hat{\beta}_z = \bar{y}_\pi - \bar{z}_\pi \hat{\beta}, \tag{2.7}$$

où $\bar{Z} = 0$ est la moyenne de population de z_j, $\bar{z}_\pi = \bar{x}_\pi - \bar{X}$,

$$(\bar{y}_\pi, \bar{z}_\pi) = \left(\sum_{i=1}^{n} \pi_i^{-1} \right)^{-1} \sum_{i=1}^{n} \pi_i^{-1} (y_i, z_i)$$

et

$$\hat{\beta}_z = \left[\sum_{j=1}^{n} (z_j - \bar{z}_\pi)' \pi_i^{-1} (z_j - \bar{z}_\pi) \right]^{-1} \sum_{j=1}^{n} (z_j - \bar{z}_\pi)' \pi_i^{-1} y_j .$$

Dans l'expression (2.7), $\bar{y}_r$ représente l'ordonnée à l'origine dans la régression de y par rapport à z. Par conséquent, la théorie exposée par Fuller (1975) sur les coefficients de régression s'applique à l'estimateur par régression de la moyenne. Si l'effectif de la population est connu et désigné par N, le total estimé pour la population est $N\bar{y}_r$.

En définissant une série de populations et d'échantillons, on peut montrer que l'estimateur (2.1) est un estimateur convergent de la moyenne de y. Voir, par exemple, Fuller (1975). L'estimateur de la variance de l'estimateur par régression est une fonction des probabilités conjointes. Considérons un échantillon à deux degrés stratifié et remplaçons l'indice inférieur i par l'indice triple ℓjt. Alors, en omettant le facteur de correction pour population finie, nous avons l'estimateur de variance suivant:

$$\hat{V}\{\bar{y}_r\} = (n - k)^{-1} n \sum_{\ell=1}^{L} (n_\ell - 1)^{-1} n_\ell \sum_{j=1}^{n_\ell} (d_{\ell j.} - d_{\ell ..})^2, \tag{2.8}$$

où

$$d_{\ell j.} = \sum_{t=1}^{m_{\ell j}} w_{\ell jt} (y_{\ell jt} - x_{\ell jt}\hat{\beta}),$$

$$d_{\ell ..} = n_\ell^{-1} \sum_{j=1}^{n_\ell} d_{\ell j.},$$

n_ℓ étant le nombre d'unités primaires-échantillon dans la strate ℓ, $m_{\ell j}$, le nombre d'éléments de l'échantillon dans l'unité primaire d'échantillonnage j de la strate ℓ, $\hat{\beta}$, le vecteur de coefficients défini en (2.2), n, l'effectif de l'échantillon et $w_{\ell jt}$, le poids de l'élément t dans l'unité primaire d'échantillonnage j de la strate ℓ. Le facteur $n - k$ est utilisé par analogie avec le diviseur contenu dans l'estimateur sans biais de la variance de l'erreur pour la régression ordinaire. Lorsque le vecteur des variables de contrôle est exprimé sous la forme définie en (2.6), l'équation (2.8) correspond à la variance estimée du premier élément de $\hat{\beta}$, l'ordonnée à l'origine estimée. L'estimateur (2.8) avait été proposé dans Hidiroglou, Fuller et Hickman (1976) et sa convergence, vérifiée dans Fuller (1975). Voir aussi Särndal, Swensson et Wretman (1989).

Les estimateurs construits avec les poids (2.4) conviennent bien pour de grands échantillons, mais sont peut-être moins intéressants avec de petits échantillons. Comme les poids sont des fonctions linéaires de x_i, il se peut que certains d'entre eux soient négatifs. Or, des poids négatifs peuvent faire que des estimations de paramètres positifs seront négatives. Husain (1969) a exploré des méthodes pour construire des poids de régression non négatifs. Huang (1978), pour sa part, a élaboré un programme informatique pour produire de tels poids. Huang et Fuller (1978)

ont décrit la méthode de production de poids et montré que la distribution de l'estimateur modifié pour de grands échantillons est identique à celle de l'estimateur par régression ordinaire. Voir aussi Goebel (1976).

L'algorithme de Huang (1978) est une méthode itérative fondée sur le principe des moindres carrés généralisés. Cet algorithme a pour but de produire un ensemble de poids w_i, $i = 1, 2, \ldots, n$, qui satisfont (2.5) et qui diffèrent peu des poids initiaux, l'écart étant une fonction du poids initial. L'algorithme de Huang vise à calculer des poids w_i qui satisfont l'inéquation

$$(1 - M) \max_{1 \le i \le n} w_i \pi_i^{-1} \le (1 + M) \min_{1 \le i \le n} w_i \pi_i^{-1},$$

où le paramètre M, $0 < M \le 1$, est défini par l'utilisateur, sa valeur se situant le plus souvent dans l'intervalle [0.8, 1.0]. Si les poids de régression primaires définis en (2.4) ne satisfont pas aux conditions, on calcule une deuxième série de poids. Ceux-ci sont des poids de régression pondérés, où chaque observation est affectée d'un facteur de contrôle: facteur peu élevé pour les observations qui ont un poids primaire faible ou élevé; facteur relativement grand pour les observations qui ont un poids primaire voisin de π_i^{-1}. Les poids de la deuxième série sont soumis à une vérification et, s'ils ne satisfont pas aux conditions, on modifie les facteurs de contrôle, et ainsi de suite. L'algorithme est décrit en annexe.

La pondération de contrôle utilisée dans l'algorithme de Huang a beaucoup en commun avec les méthodes de réduction d'influence et les méthodes de régression robustes, c'est-à-dire que dans l'estimateur final, la pondération de contrôle a pour effet de réduire la part des observations éloignées de la moyenne dans l'estimation du vecteur-pente. Voir Hampel (1978), Krasker (1980) et Mallows (1983). Parmi les ouvrages récents portant sur ce type d'estimateurs pour des échantillons d'enquête, notons Deville et Särndal (1992), Akkerboom, Sikkel et van Herk (1991), Hulliger (1993) et Singh (1993).

Il n'est pas toujours possible de créer des poids qui en même temps répondent aux critères et satisfassent l'équation (2.5). Par exemple, si toutes les observations de x_{t2} excèdent la moyenne, il n'y aura aucun ensemble de poids positifs dont la somme sera égale à un et qui satisferont aussi $\sum_{t=1}^{n} x_{t2} w_t = \bar{X}_2$. Le programme de production de poids s'interrompra donc si, après un nombre déterminé d'itérations, il n'est pas possible d'obtenir des poids qui répondent aux critères établis.

Dans certains cas, il est souhaitable de limiter les poids à des entiers non négatifs, par exemple lorsqu'on établit des estimations de totaux et que la population renferme des unités bien définies, comme des personnes. L'utilisation de poids entiers non négatifs donne alors des estimations plus satisfaisantes en ce sens qu'elles sont palpables. On peut créer des poids en nombre entier de sorte qu'il ne soit pas nécessaire d'effectuer un arrondissement lorsqu'on construit des tableaux. Avec de tels poids, tous les tableaux à multiples entrées seront automatiquement cohérents.

Le programme de Huang contient une fonction qui permet d'arrondir les poids en nombre réel sans en modifier la somme. Après l'arrondissement, l'équation (2.5) ne sera en général plus parfaitement exacte. Nous avons observé qu'en itérant l'algorithme de Huang en se servant des poids primaires en nombre entier comme poids initiaux, on pouvait obtenir des poids en nombre entier qui satisfaisaient à peu de choses près l'équation (2.5). La question de l'arrondissement est traitée dans Cox (1987), Cox et Ernst (1982) et Fagan, Greenberg et Hemmig (1988).

3. ESTIMATION PAR RÉGRESSION EN SITUATION DE NON-RÉPONSE

À l'origine, la théorie de l'estimation par régression supposait que l'échantillon étudié était un échantillon probabiliste de la population. Or, on reconnaît depuis longtemps déjà que l'estimation par régression peut servir à réduire le biais qui découle des faiblesses de la méthode de collecte des données. La plus notable de ces faiblesses est la non-réponse. Dans tous les grands échantillons constitués de personnes, il y a des individus qui ne donnent pas l'information voulue. Si les non-répondants n'ont pas les mêmes caractéristiques que les répondants, les estimations directes établies à l'aide des données recueillies seront biaisées. Si l'on possède de l'information supplémentaire, l'estimation par régression est une méthode pour réduire le biais. La grandeur de la réduction dépendra du rapport entre les variables de contrôle, les variables étudiées et les probabilités de réponse. Voir Little et Rubin (1987) pour une analyse générale de ces questions.

Posons π_i^* comme la probabilité conjointe de sélection et de réponse, exprimée comme le produit de π_i par la probabilité conditionnelle d'observer l'unité étant donné que celle-ci a été échantillonnée. Alors,

$$E\left\{ \sum_{i=1}^{n} x_i' \pi_i^{-1} x_i \mid \xi_N \right\} = \sum_{i=1}^{N} x_i' \pi_i^{-1} \pi_i^* x_i \qquad (3.1)$$

et

$$E\left\{ \sum_{i=1}^{n} x_i' \pi_i^{-1} y_i \mid \xi_N \right\} = \sum_{i=1}^{N} x_i' \pi_i^{-1} \pi_i^* y_i, \qquad (3.2)$$

où l'espérance dépend de la population finie donnée ξ_N et n est la taille effective de l'échantillon. En situation de non-réponse, le ratio $p_i = \pi_i^* \pi_i^{-1}$ est la probabilité de réponse pour l'individu i. Par conséquent, dans des conditions comme celles définies par Fuller (1975),

$$\operatorname*{plim}_{\substack{n \to \infty \\ N \to \infty}} (\hat{\beta} - \gamma) = 0, \qquad (3.3)$$

où $\hat{\beta}$ est défini en (2.2) et

$$\gamma = \left(\sum_{i=1}^{N} x_i' \pi_i^{-1} \pi_i^* x_i \right)^{-1} \sum_{i=1}^{N} x_i' \pi_i^{-1} \pi_i^* y_i. \qquad (3.4)$$

Ensuite,

$$\bar{Y} = \bar{X}\gamma + \bar{A}, \tag{3.5}$$

où $\bar{A} = N^{-1}\sum_{i=1}^{N} a_i$ et $a_i = y_i - x_i\gamma$. Donc, l'estimateur par régression (2.1) sera un estimateur convergent de $\bar{Y}$ si $\text{plim}_{N\to\infty}\bar{A} = 0$. La probabilité limite de $\bar{A}$ sera nulle si la population finie est un échantillon aléatoire tiré d'une population infinie pour laquelle le modèle linéaire

$$y_i = x_i\beta + e_i, \quad E\{e_i\} = 0$$

est valide pour tous les i.

La moyenne $\bar{A}$ est nulle lorsque $\overset{*}{\pi}_i = \pi_i$ pour tous les i et qu'un élément de x_i est égal à un pour tous les i car dans ces conditions

$$\gamma = \beta = \left(\sum_{i=1}^{N} x_i'x_i\right)^{-1}\sum_{i=1}^{N} x_i'y_i \tag{3.6}$$

et $\sum_{i=1}^{N}(y_i - x_i\beta) = 0$. Une condition suffisante pour que $\bar{A}$ soit nulle est l'existence d'un vecteur ligne c qui soit tel que

$$cx_i' = \overset{*}{\pi}_i^{-1}\pi_i = p_i^{-1}, \tag{3.7}$$

pour $i = 1, 2, \ldots, N$. Donc, si le rapport de la probabilité nominale à la probabilité vraie est une fonction linéaire des variables de contrôle, l'estimateur par régression est un estimateur convergent de la moyenne de y, la limite s'appliquant à des suites comme le définit Fuller (1975). Pour que l'équation (3.7) soit satisfaite, il faudrait, par exemple, que les éléments de x_i soient des variables fictives qui définissent des sous-groupes et que les probabilités de réponse soient les mêmes à l'intérieur de chaque sous-groupe. On décrit parfois cette situation en disant que des éléments sont absents de façon aléatoire dans chaque sous-groupe. Dans l'analyse empirique, nous nous servons de l'hypothèse que $\bar{A} = 0$ comme hypothèse de travail.

Dans un problème de régression, il y a certaines hypothèses qui ne peuvent être vérifiées à l'aide des données de l'échantillon. Dans la régression par les moindres carrés ordinaires par exemple, les résidus $\hat{e}_i = y_i - x_i\hat{\beta}$ ne sont pas corrélés avec x_i et ne peuvent donc pas servir à vérifier l'hypothèse selon laquelle les erreurs vraies sont non corrélées avec x. Par conséquent, dans une enquête avec non-réponse, on doit trouver des variables de contrôle qui sont corrélées avec y ou que l'on croit être corrélées avec les probabilités de réponse. Toutefois, un ensemble particulier de variables de contrôle ne suffit pas pour affirmer que l'estimation par régression a contribué à supprimer le biais.

Dans la pratique, il est souvent possible de trouver des variables x qui sont corrélées avec la probabilité de réponse ou avec les variables y. Par exemple, dans la Nationwide Food Consumption Survey de 1987-1988, on a enregistré un faible taux de réponse chez les ménages à revenu élevé. Par conséquent, l'utilisation de variables ayant trait au revenu du ménage dans un estimateur par régression est censée réduire le biais de l'estimation de la moyenne pour des caractéristiques qui sont corrélées avec le revenu.

L'erreur contenue dans $\hat{\beta}$, comme estimateur de γ, peut être définie approximativement par la formule

$$\hat{\beta} - \gamma \doteq G^{-1}T^{-1}\sum_{i=1}^{n} x_i'\pi_i^{-1}a_i,$$

où a_i est défini en (3.5),

$$T = \sum_{i=1}^{N}\pi_i^{-1}\overset{*}{\pi}_i$$

et

$$G = T^{-1}\sum_{i=1}^{N} x_i'\pi_i^{-1}\overset{*}{\pi}_i x_i.$$

Suivant des hypothèses raisonnables,

$$\hat{T} = \sum_{i=1}^{n}\pi_i^{-1}$$

et

$$\hat{G} = \hat{T}^{-1}\sum_{i=1}^{n} x_i'\pi_i^{-1}x_i$$

sont des estimateurs convergents de T et G. Par conséquent, on peut estimer la variance de l'estimateur par régression en estimant la variance de $\sum_{i=1}^{n} x_i'\pi_i^{-1}a_i$. Si nous supposons que les probabilités conditionnelles de réponse relatives à une unité primaire d'échantillonnage sont indépendantes de celles relatives à toutes les autres et qu'au moins une unité est observée dans chaque u.p.é. échantillonnée, alors l'estimateur (2.8) est toujours indiqué comme estimateur de la variance de l'estimation de la moyenne de y.

L'estimateur (2.8) peut aussi convenir lorsque les poids de régression sont calculés autrement que par (2.4). Par exemple, définissons les poids par l'expression

$$w_{gi} = \bar{X}\left[\sum_{i=1}^{n} x_i'\pi_i^{-1}g_i x_i\right]^{-1} x_i'\pi_i^{-1}g_i,$$

où les g_i sont fonction des x_i. Supposons que

$$\text{plim}\,\hat{\beta}_g = \gamma_g,$$

où

$$\hat{\beta}_g = \left[\sum_{i=1}^{n} x_i'\pi_i g_i x_i\right]^{-1}\sum_{i=1}^{n} x_i'\pi_i^{-1}g_i y_i.$$

Supposons aussi que

$$\underset{N\to\infty}{\text{plim}}\, N^{-1}\sum_{i=1}^{N}(y_i - x_i\gamma_g) = 0.$$

Alors, l'expression (2.8), où l'on a substitué w_{gi} à w_{ljt}, est un estimateur convergent de la variance de l'estimateur. Nous nous servirons de l'estimateur (2.8) dans nos analyses empiriques.

La formule (2.8) définit les deux effets de l'estimation par régression pour la variance de l'estimation de la moyenne. L'effet de corrélation réduit cette variance tandis que l'accroissement de la somme des carrés des poids l'augmente. Pour comprendre ces effets, considérons un échantillon aléatoire simple. Si la variable y est corrélée avec x, la corrélation tend à réduire la variance de l'estimateur par régression par rapport à celle de l'estimateur simple car

$$E\{(y_i - x_i\beta)^2\} \leq E\{[y_i - E(y_i)]^2\}.$$

Si la moyenne empirique des variables de contrôle diffère de la moyenne de ces variables pour la population, alors

$$\sum_{i=1}^{n} w_i^2 > n^{-1},$$

où n^{-1} est la somme des carrés des poids pour un échantillon aléatoire simple.

Lorsque l'on compare la variance de la moyenne de l'échantillon avec celle de l'estimateur par régression, il ne faut pas oublier que le recours à l'estimation par régression pour des échantillons affectés par la non-réponse a pour but notamment de produire un estimateur moins biaisé que l'estimateur direct. C'est pourquoi, dans la section suivante, nous comparons un estimateur de l'erreur quadratique moyenne de l'estimateur simple avec un estimateur de la variance de l'estimateur par régression.

4. APPLICATION À LA NATIONWIDE FOOD CONSUMPTION SURVEY

La Nationwide Food Consumption Survey de 1987-1988 a été effectuée par le Human Nutrition Information Service du Département de l'agriculture des É.-U. L'échantillon initial était un échantillon stratifié autopondéré d'unités primaires d'échantillonnage (u.p.é.) spatiales réparties dans les 48 États contigus. Les u.p.é. étaient divisées en unités secondaires appelées "segments aréolaires". Les ménages des segments aréolaires participaient à une interview en personne. Les opérations sur le terrain étaient confiées à un entrepreneur lié par contrat au Human Nutrition Information Service et se sont déroulées entre avril 1987 et août 1988.

Environ 37% des unités de logement reconnues occupées ont fourni toute l'information voulue sur la consommation d'aliments au sein du ménage. La taille effective de l'échantillon était de 4,495 ménages. Vu le faible taux de réponse, le Human Nutrition Information Service a décidé de recourir à la pondération par régression pour l'estimation. L'organisme a estimé les chiffres de population pour toutes les caractéristiques sauf le degré d'urbanisation en se fondant sur les données de la Current Population Survey de mars 1987; voir Bureau of the Census (1987). Les chiffres de population pour les catégories du degré d'urbanisation ont été fournis par l'entrepreneur. Dans notre analyse, nous considérons les chiffres de population estimés comme des chiffres officiels.

Tableau 1

Caractéristiques des ménages dans l'échantillon et la population

Caractéristique	Catégorie	Effectif dans l'échantillon de ménages	Proportion de l'échantillon de ménages	Proportion de la population
Période d'interview	Printemps	1,828	40.7	25.0
	Été	678	15.1	25.0
	Automne	717	16.0	25.0
	Hiver	1,272	28.3	25.0
Région	Nord-Est	905	20.1	21.2
	Midwest	1,172	26.1	24.7
	Sud	1,567	34.9	34.4
	Ouest	851	18.9	19.6
Degré d'urbanisation	Noyau urbain	1,064	23.7	31.2
	Banlieue	2,122	47.2	46.0
	Région non métropolitaine	1,309	29.1	22.9
Revenu du ménage en % du seuil de pauvreté	< 131%	1,041	23.2	20.0
	131-300%	1,564	34.8	32.2
	301-500%	1,108	24.6	25.9
	> 500%	782	17.4	21.8
Coupons alimentaires	Oui	314	7.0	7.4
	Non	4,181	93.0	92.6
Propriétaire du logement	Oui	2,998	66.7	64.1
	Non	1,497	33.3	35.9
Origine raciale du chef du ménage	Noir	519	11.5	11.1
	Non noir	3,976	88.5	88.9
Âge du chef du ménage	< 25	338	7.5	7.9
	25-39	1,588	35.3	36.1
	40-59	1,369	30.5	30.5
	60-69	660	14.7	13.0
	70+	540	12.0	12.6
Chef du ménage	Homme et femme	3,057	68.0	60.8
	Femme seulement	1,044	23.2	26.0
	Homme seulement	394	8.8	13.2
Femme chef de ménage a travaillé	Oui	1,792	39.9	41.5
	Non	2,703	60.1	58.5
Exactement un adulte dans le ménage	Oui	1,211	26.9	29.7
	Non	3,284	73.1	70.3
Exactement deux adultes dans le ménage	Oui	2,616	58.2	54.2
	Non	1,879	41.8	45.8
Présence d'enfants de moins de 7 ans	Oui	1,009	22.4	20.1
	Non	3,486	77.6	79.9
Présence d'enfants de 7 à 17 ans	Oui	1,309	29.1	26.5
	Non	3,186	70.9	73.5
Taille du ménage	(Moyenne)		2.731	2.642
Taille du ménage, au carré	(Moyenne)		9.546	9.125

Le tableau 1 présente les caractéristiques de la population et de l'échantillon de ménages. L'échantillon initial montrait un déséquilibre en ce qui concerne la période d'interview: près de 41% des interviews avaient eu lieu au trimestre de

printemps et environ 16% dans chaque cas au trimestre d'été et au trimestre d'automne. Des interviews ont eu lieu aux trimestres de printemps et d'été de 1987 et 1988.

L'échantillon n'était pas non plus équilibré en ce qui a trait à la caractéristique "degré d'urbanisation". Les ménages vivant dans les noyaux urbains étaient sous-représentés dans l'échantillon (24%, contre 31% dans la population), alors que les ménages vivant dans les régions non métropolitaines étaient surreprésentés (29%, contre 23% dans la population).

Les ménages à revenu élevé étaient sous-représentés alors que c'était le contraire pour les ménages ayant à leur tête et un homme et une femme (68%, contre 61% dans la population). Les ménages avec enfants étaient, eux, surreprésentés. Pour plusieurs autres caractéristiques socio-démographiques, le déséquilibre était assez peu prononcé.

Le personnel du Human Nutrition Information Service est d'avis que les caractéristiques énumérées dans le tableau 1 ont un rapport avec les habitudes de consommation alimentaire. C'est pourquoi on s'est servi de variables basées sur ces caractéristiques dans la pondération par régression. Pour mettre en marche le programme de production de poids, on a transformé chacune des variables nominales du tableau 1 en un ensemble de variables indicatrices. Par exemple, on a créé trois variables pour la caractéristique "revenu du ménage en pourcentage du seuil de pauvreté". Ces trois variables sont les suivantes:

$Z_{t1} = 1$ si la proportion pour le t-ième ménage est inférieure à 131%
$= 0$ autrement,

$Z_{t2} = 1$ si la proportion pour le t-ième ménage se situe entre 131 et 300%
$= 0$ autrement,

$Z_{t3} = 1$ si la proportion pour le t-ième ménage se situe entre 301 et 500%
$= 0$ autrement.

Vingt-cinq variables indicatrices ont été créées de cette manière. En outre, la taille du ménage et le carré de la taille du ménage servaient de variables continues.

Les vingt-sept variables ont été utilisées dans le programme de Huang pour produire les poids de régression. On a fixé à 0.9 la valeur du paramètre M du programme de production de poids et on a arrondi les poids à l'entier naturel le plus près, ces poids étant exprimés en milliers. La somme des poids est de 88,942, chiffre qui correspond au nombre de ménages (en milliers) dans la population. Le poids moyen est de 19.787, le poids le moins élevé, de 6, et le poids le plus élevé, de 47. Le poids le plus élevé équivaut donc à 2.38 fois le poids moyen. La somme des carrés des poids est de 2,317,930. Le produit du carré du poids moyen par la taille de l'échantillon est de 1,759,884. Par conséquent, s'il y a absence de corrélation entre une variable quelconque et les 27 variables, la variance d'une estimation calculée à l'aide de poids sera environ 1.32 fois plus élevée que la variance d'un estimateur non pondéré ordinaire.

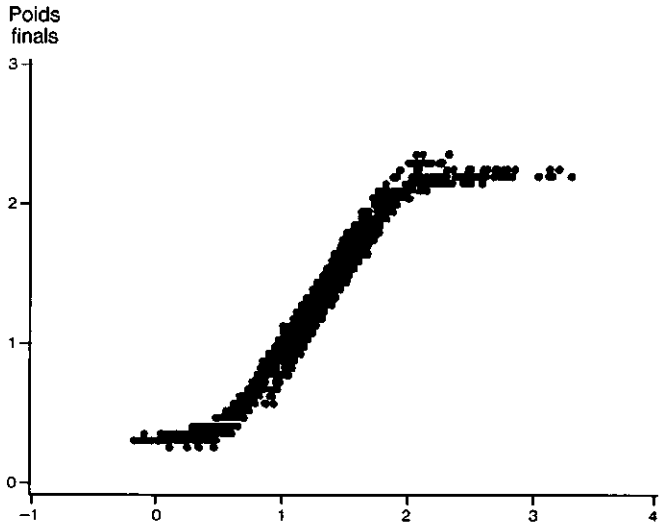


Figure 1. Poids finals exprimés en fonction des poids par les moindres carrés ordinaires, les deux séries étant divisées par le poids moyen.

La figure 1 contient un graphique où sont représentés, sur un axe, les poids de régression calculés au moyen de l'algorithme de Huang et, sur l'autre axe, les poids par les moindres carrés ordinaires. Comme il y a 4,495 ménages, beaucoup de points ne sont pas apparents. Les deux séries de poids sont normalisées, ayant été divisées par le poids moyen. La moyenne pour des poids ayant subi cette transformation est donc égale à un. Comme 27 variables de contrôle sont utilisées, les poids calculés par l'algorithme de Huang forment une nuée de points le long d'une courbe en S tracée en fonction des poids initiaux. S'il n'y avait qu'une variable de contrôle, les points formeraient une courbe en S au sens strict. Les poids initiaux se rattachant à des observations à la gauche du zéro étaient négatifs.

Pour comparer les estimations calculées à l'aide de poids avec les estimations non pondérées, nous utilisons les variables suivantes:

Y_1 = nombre total redressé de repas pris à l'extérieur du foyer (repas à l'extérieur);

Y_2 = valeur monétaire totale des aliments consommés à la maison (aliments à la maison);

Y_3 = taille du ménage en nombre de personnes équivalent à 21 repas (repas-personnes);

Y_4 = indicateur pour désigner les ménages qui tiennent la maison (tenue de la maison).

La taille du ménage en nombre de personnes équivalent à 21 repas est le nombre total redressé de repas préparés avec les réserves alimentaires du ménage dans les 7 derniers jours, divisé par 21. La variable "repas-personnes" est la somme de deux termes: le premier est la somme des proportions des repas pris à la maison par chaque membre du ménage durant la semaine d'interview; le second, le nombre de repas servis à des invités, à des pensionnaires ou à des employés durant la semaine d'interview, divisé par 21. Autrement dit,

$$\begin{matrix}\text{repas-personnes}\\ \text{pour le}\\ j\text{-ième ménage}\end{matrix} = \sum_i (h_{ij} + a_{ij})^{-1} h_{ij} + (21)^{-1} b_j,$$

où h_{ij} = nombre de repas pris à la maison par l'individu i dans le ménage j durant la semaine d'interview, a_{ij} = nombre de repas pris à l'extérieur du foyer par l'individu i dans le ménage j durant la semaine d'interview et b_j = nombre de repas servis à des personnes qui ne sont pas membres du ménage dans le ménage j durant la semaine d'interview.

Le nombre total redressé de repas pris à l'extérieur du foyer est la somme des proportions des repas pris à l'extérieur par chaque membre du ménage durant la semaine d'interview, multipliée par 21. En reprenant la notation utilisée plus haut pour les repas-personnes, nous avons

$$\begin{matrix}\text{repas à l'extérieur}\\ \text{pour le}\\ j\text{-ième ménage}\end{matrix} = 21 \sum_i (h_{ij} + a_{ij})^{-1} a_{ij}.$$

La valeur totale des aliments consommés à la maison est égale à la somme des dépenses d'alimentation et de la valeur monétaire des aliments produits à la maison et des aliments reçus gratis qui ont été consommés durant la semaine d'enquête. Les dépenses d'alimentation ont été évaluées au prix que l'on disait avoir payé, abstraction faite du moment de l'achat et la taxe de vente exclue. En ce qui concerne les aliments dont le prix n'était pas déclaré, les aliments produits à la maison et les aliments reçus en cadeau ou en guise de paiement, ils ont été évalués au prix moyen la livre qu'avaient payé les ménages visés par l'enquête pour de la nourriture semblable dans la même région et à la même période de l'année.

Un ménage qui tient la maison est un ménage où au moins une personne a un nombre redressé de dix repas ou plus avec les réserves alimentaires du ménage dans les sept jours précédant l'interview. En règle générale, les études de consommation alimentaire des ménages portent uniquement sur des ménages qui tiennent la maison.

Les moyennes de variables calculées à l'aide de données non pondérées figurent dans la colonne "moyenne non pondérée" du tableau 2. Trois moyennes sont calculées pour chacune des trois variables suivantes: "repas à l'extérieur", "aliments à la maison" et "repas-personnes". Deux moyennes sont calculées pour les deux sous-populations définies par la variable indicatrice "tenue de la maison". La troisième, qui correspond à la rubrique "ensemble de la population" dans le tableau, est la moyenne estimée pour la population totale. L'erreur-type de l'estimation figure entre parenthèses sous chaque valeur estimée. Les estimations non pondérées et les erreurs types correspondantes ont été calculées à l'aide de PC CARP; voir Fuller et coll. (1986). Le calcul a tenu compte du fait qu'il s'agissait d'un échantillon en grappes stratifié aréolaire.

Comme il s'agit d'un échantillon à deux degrés, les variances estimées sont plus grandes que la variance d'un échantillon aléatoire simple qui contiendrait le même nombre de ménages. Le rapport entre la variance d'une estimation pour un plan donné et la variance de l'estimation pour un échantillon aléatoire simple de même taille est appelé "effet du plan". En ce qui concerne les moyennes estimées à l'aide de données non pondérées pour l'ensemble de la population, l'effet du plan estimé est d'environ 2.5 pour les repas à l'extérieur et les repas-personnes, d'environ 4.1 pour les aliments à la maison et d'environ 1.5 pour la tenue de la maison.

Tableau 2

Propriétés de divers estimateurs

Variable	Moyenne non pondérée	Moyenne pondérée	Écart	Efficacité relative de la régression
Repas à l'extérieur				
Ménages tenant maison	7.75 (0.22)	7.93 (0.17)	−0.18 (0.09)	2.52
Ménages ne tenant pas maison	18.31 (1.14)	18.12 (1.19)	0.19 (0.68)	0.92
Ensemble de la population	8.27 (0.22)	8.57 (0.22)	−0.30 (0.12)	2.56
Aliments à la maison				
Ménages tenant maison	61.10 (1.14)	59.56 (0.98)	1.54 (0.41)	3.65
Ménages ne tenant pas maison	25.99 (1.25)	26.39 (1.46)	−0.40 (1.00)	0.73
Ensemble de la population	59.37 (1.12)	57.49 (0.91)	1.88 (0.39)	5.60
Repas-personnes				
Ménages tenant maison	2.42 (0.03)	2.33 (0.01)	0.09 (0.01)	89.00
Ménages ne tenant pas maison	0.51 (0.03)	0.49 (0.03)	0.02 (0.02)	1.00
Ensemble de la population	2.33 (0.03)	2.22 (0.01)	0.11 (0.01)	129.00
Tenue de la maison (%)	95.06 (0.40)	93.77 (0.58)	1.29 (0.10)	5.30

La colonne intitulée "moyenne pondérée" contient les estimations calculées au moyen des poids de régression. Les erreurs types correspondantes ont été calculées sur PC CARP au moyen de la formule (2.8), les π_i^{-1} étant remplacés par les poids de régression. Pour calculer la variance, il fallait faire une régression pour chaque variable y. Les moyennes estimées pour les sous-populations sont des rapports d'estimateurs par régression. On a calculé la variance de ces moyennes estimées en calculant au moyen de la formule (2.8) la variance des écarts aléatoires de Taylor pour les rapports en question. Les estimations non pondérées et les estimations pondérées ont des erreurs types semblables en ce qui concerne les repas à l'extérieur et les aliments à la maison. Cependant, pour les repas-personnes, l'erreur type de l'estimation par régression de la moyenne de population est d'environ un tiers de l'erreur type de l'estimation non pondérée. L'erreur type de l'estimateur par régression est moins élevée parce que la variable "repas-personnes" est fortement corrélée avec les variables de taille du ménage utilisées comme variables de contrôle dans la régression.

Les valeurs estimées des coefficients de corrélation multiple carrée entre les variables du tableau 2 et les 27 variables de contrôle sont de 0.29 pour les repas à l'extérieur, de 0.44 pour les aliments à la maison, de 0.82 pour les repas-personnes et de 0.12 pour la tenue de la maison. Si les moyennes d'échantillon des variables de contrôle étaient presque égales aux moyennes de population, l'erreur type de l'estimation par régression de la moyenne pour les repas à l'extérieur serait d'environ 84 centièmes ($(1 - 0.29)^{1/2}$) de l'erreur type de l'estimation non pondérée. En fait, l'erreur type estimée de l'estimation par régression équivaut à peu près à 97 centièmes de l'erreur type de l'estimation non pondérée. L'écart est dû au fait que $\sum_{i=1}^{n} w_i^2$ est beaucoup plus grande que n^{-1} parce que l'échantillon accuse un déséquilibre pour un certain nombre de caractéristiques. Notons que

$$0.97 \doteq [(0.71)(1.32)]^{1/2},$$

où $0.71 = (1 - 0.29)$ est un moins le carré du coefficient de corrélation et $1.32 = n\sum_{i=1}^{n} w_i^2$. En ce qui concerne la variable "tenue de la maison", la situation est plus particulière. La corrélation n'est pas forte et, semble-t-il, les écarts par rapport à la droite de régression les plus prononcés sont associés à des poids élevés. La variance estimée de l'estimateur par régression est d'environ le double de celle de l'estimateur non pondéré.

Le tableau 2 contient aussi la valeur estimée de l'écart entre l'estimation non pondérée et l'estimation pondérée. L'écart entre le total estimé non pondéré et le total estimé pondéré est

$$\sum_{t=1}^{n} Nn^{-1}y_t - \sum_{t=1}^{n} w_t y_t = \sum_{t=1}^{n} (n^{-1}N - w_t)y_t.$$

L'écart entre les moyennes estimées est égal au quotient de la différence de totaux par la taille de la population. Pour calculer la variance de l'écart entre les moyennes, notons que l'hypothèse d'un écart nul équivaut à l'hypothèse d'une corrélation nulle entre w_t et y_t. Nous avons donc effectué la régression non pondérée de y_t par rapport à w_t et nous avons calculé la variance du coefficient de régression selon le plan au moyen de PC CARP. Les erreurs types des écarts estimés du tableau 2 sont telles que la "statistique t" pour l'hypothèse de l'écart nul est égale à la "statistique t" pour le coefficient de w_t dans la régression de y_t par rapport à w_t.

Pour les quatre caractéristiques, l'écart entre l'estimation pondérée et l'estimation non pondérée de la moyenne de population est significatif aux niveaux habituels. Par conséquent, si on suppose que les estimateurs par régression sont non biaisés, les estimateurs non pondérés, eux, comportent un biais appréciable.

En ce qui regarde les estimations des moyennes de sous-population, les résultats sont partagés. L'écart entre l'estimation pondérée et l'estimation non pondérée est significatif pour les trois moyennes relatives à la sous-population des ménages tenant la maison – qui est en fait la population étudiée – alors qu'il ne l'est pas pour les trois moyennes relatives à la sous-population des ménages qui ne tiennent pas la maison. L'échantillon ne compte que 222 ménages de ce genre. Par conséquent, la variance de l'écart entre l'estimation pondérée et l'estimation non pondérée est beaucoup plus grande pour les ménages qui ne tiennent pas la maison que pour les autres.

L'écart entre les deux estimations de la moyenne de population est fonction de l'écart entre les deux estimations de la moyenne de sous-population et de l'écart entre les deux estimations de la proportion de ménages de l'une et l'autre catégories. Voilà pourquoi l'écart pour l'"ensemble de la population" peut être plus grand que l'écart pour l'une et l'autre sous-populations.

La dernière colonne du tableau 2 contient le rapport de l'erreur quadratique moyenne estimée de l'estimateur non pondéré à la variance de l'estimateur par régression. L'erreur quadratique moyenne estimée de l'estimateur non pondéré a été calculée par la formule

$$\widehat{\text{EQM}}_u = \hat{V} + \max\{0, (\text{Diff})^2 - (\text{e.t. diff})^2\},$$

où $\hat{V}$ est la variance estimée de l'estimation non pondérée, "Diff", l'écart entre les deux estimations dans le tableau 2, et "e.t. diff", l'erreur type de l'écart dans le tableau 2. La variance estimée $\hat{V}$ de l'estimateur non pondéré est définie par la formule (2.8), où w_{tjt} est constant et où $x_{tjt}\hat{\beta}$ est remplacé par $\bar{y}_{t..}$. Le second terme de l'erreur quadratique moyenne estimée est le carré du biais estimé. Suivant l'hypothèse que l'estimateur par régression est non biaisé, l'espérance de $(\text{Diff})^2$ est égale à la somme du carré du biais et de la variance de l'écart. Donc, suivant l'hypothèse que l'estimateur par régression est non biaisé, l'espérance mathématique de $(\text{Diff})^2$ est le carré du biais plus la variance de la différence. L'erreur quadratique moyenne estimée de l'estimateur pondéré correspond à la variance de l'estimateur pondéré calculée comme le carré de l'erreur type dans le tableau 2.

Si l'on considère les quatre caractéristiques pour lesquelles des moyennes de population ont été estimées, l'efficacité relative estimée de l'estimateur par régression (par rapport à un estimateur simple) varie de 2.5 à 129. La valeur la moins élevée se rapporte à la variable "repas à l'extérieur". Les variances des deux estimateurs sont semblables mais, à cause du biais estimé, l'erreur quadratique moyenne (EQM) de l'estimation par régression pour cette variable équivaut à environ 40% de l'EQM de l'estimation non pondérée. Pour la variable "aliments à la maison", l'EQM de l'estimation par régression équivaut à moins de 20% de celle de l'estimation non pondérée; pour les deux autres variables, le rapport correspondant est d'environ 1% (repas-personnes) et 20% (tenue de la maison). Dans tous les cas, le carré du biais est une composante très importante de l'erreur quadratique moyenne estimée.

Comme on observe un biais relativement faible dans les estimations non pondérées touchant la sous-population des ménages qui ne tiennent pas la maison, on considère que pour cette sous-population les estimations non pondérées sont un peu plus efficaces que les estimations par

régression. Or, cette dernière ne représente qu'environ 6% de la population et on observe peu de corrélation entre les écarts par rapport à la moyenne de sous-population et les variables de contrôle. Par ailleurs, on considère que les estimations par régression sont beaucoup plus efficaces que les estimations non pondérées pour la sous-population des ménages qui tiennent la maison. L'efficacité relative des estimations par régression pour cette sous-population se rapproche de celle des estimations pour l'ensemble de la population.

Même si l'on tient compte du fait que les chiffres de population tirés de la Current Population Survey ne sont pas des chiffres officiels, il est clair que l'estimation par régression de moyennes de population engendre des gains d'efficacité appréciables. Et, bien que l'on considère qu'un estimateur par régression de la moyenne d'une petite sous-population est moins efficace qu'un estimateur non pondéré du même paramètre, la perte d'efficacité est faible par rapport aux gains d'efficacité notables estimés pour les autres variables.

REMERCIEMENTS

Cette étude a été rendue possible en partie grâce à un contrat de soutien à la recherche (n° 58-3198-9-032) passé avec le Human Nutrition Information Service du Département de l'agriculture des É.-U. Les auteurs remercient Phil Kott, Patricia Guenther et les arbitres pour leurs commentaires utiles.

ANNEXE
PROGRAMME DE PRODUCTION DE POIDS

Dans cette annexe, nous décrivons le programme de production de poids de régression de Huang et Fuller (1978). Ce programme offre la possibilité de définir un poids minimum et un poids maximum. La version originale du programme n'offrait pas cette possibilité. Pour une analyse de méthodes de production de poids connexes, voir Singh (1993).

Supposons connues les moyennes de population $(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k)$ des k variables auxiliaires $(X_1, X_2, \ldots, X_k)$. Supposons aussi qu'il existe un échantillon de n observations et posons

$$X = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1k} \\ X_{21} & X_{22} & \ldots & X_{2k} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{nk} \end{pmatrix}, \qquad \text{(A.1)}$$

où X_{ij} est la valeur observée de la variable j pour l'individu i.

Outre le tableau d'observations d'échantillon et les moyennes de population, deux facteurs initiaux, v_i et $g_i^{(0)}$, $i = 1, 2, \ldots, n$, sont nécessaires au calcul. Le facteur v_i est en général inversement proportionnel à la probabilité de sélection. La valeur implicite de $g_i^{(0)}$ est l'unité (c.-à-d. $g_i^{(0)} = 1$). Dans le cas d'échantillons stratifiés ou de données avec variances inégales, l'utilisateur peut choisir d'autres valeurs pour $g_i^{(0)}$. (Voir Huang 1978 ou Goebel 1976.) Les paramètres du programme sont la taille de l'échantillon (n), la taille de la population (N), le paramètre M, le nombre maximum d'itérations (LI), la limite supérieure du rapport du poids au poids moyen (U_B) et la limite inférieure du rapport du poids au poids moyen (L_B). Il faut que $0 \leq L_B < 1 < U_B$. Dans notre description, nous supposons que $\sum_{i=1}^{n} v_i = n$. Le programme normalise les v_i de sorte que leur somme soit égale à n.

Le programme peut servir à créer des poids en vue de l'estimation de moyennes ou de totaux. Les poids pour les totaux équivalent aux poids pour les moyennes multipliés par N. En ce qui concerne les moyennes, le programme vise à créer des poids w_i qui soient tels que

$$\sum_{i=1}^{n} w_i(1, X_i) = (1, \bar{X}), \qquad \text{(A.2)}$$

$$L_B < n w_i < U_B, \qquad \text{(A.3)}$$

$$(1 - M) \max_{1 \leq i \leq n} w_i v_i \leq (1 + M) \min_{1 \leq i \leq n} w_i v_i, \qquad \text{(A.4)}$$

pour $i = 1, 2, \ldots, n$.

Le programme est itératif. En l'occurrence, une itération consiste à calculer les poids par les moindres carrés généralisés, un facteur de contrôle h_i étant appliqué à chaque observation. Le facteur h_i est le produit de v_i par g_i, g_i étant, après la première itération, une fonction - décrite par une courbe en cloche - de la distance (exprimée en une mesure appropriée) entre l'observation et la moyenne de population. À chaque itération, les poids satisfont l'équation (A.2) mais peuvent ne pas satisfaire (A.3) ou (A.4).

Il ne sera pas toujours possible de créer des poids qui satisfont les conditions spécifiées dans le nombre d'itérations voulu. Si l'échantillon est tel qu'il n'est pas possible de respecter la contrainte, le programme produit les poids de la dernière itération. Dans le cas d'une variable x unique, si le critère ne peut être respecté, il y aura deux poids: un pour les valeurs supérieures à la moyenne de population et un autre pour les valeurs qui lui sont inférieures.

Pour décrire l'algorithme, posons

$$Z_{ij} = X_{ij} - \bar{X}_j,$$

$$Z = \begin{pmatrix} Z_{11} & Z_{12} & \ldots & Z_{1p} \\ \vdots & \vdots & & \vdots \\ Z_{n1} & Z_{n2} & \ldots & Z_{np} \end{pmatrix},$$

$$V = \text{diag}(v_1, v_2, \ldots, v_n),$$

$$J_n = (1, 1, \ldots, 1)',$$

$$A^{(0)} = Z'H^{(0)}Z,$$

$$G^{(0)} = \text{diag}(g_1^{(0)}, \ldots, g_n^{(0)})$$

et

$$H^{(0)} = VG^{(0)}.$$

L'algorithme consiste dans l'itération de trois opérations.

1. Le calcul initial est pour $\alpha = 0$. À l'itération α, le vecteur des poids de régression, désigné par $w^{(\alpha)}$, est

$$w^{(\alpha)} = [1 + n\bar{u}_v^{(\alpha)}]^{-1} V(n^{-1}J_n + u^{(\alpha)}) = (w_1^{(\alpha)}, \ldots, w_n^{(\alpha)})', \quad \text{(A.5)}$$

où

$$u^{(\alpha)} = G^{(\alpha)}Z(A^{(\alpha)})^{\dagger}(\bar{X} - \bar{x}_v) = (u_1^{(\alpha)}, \ldots, u_n^{(\alpha)})',$$

$$\bar{x}_v = \left(\sum_{i=1}^{n} v_i\right)^{-1} \sum_{i=1}^{n} v_i x_i,$$

$(A^{(\alpha)})^{\dagger}$ est un inverse généralisé symétrique de $A^{(0)}$,

$$n\bar{u}_v^{(\alpha)} = \max\{J_n'Vu^{(\alpha)}, n^{-1} - 1\}, \quad \text{(A.6)}$$

et

$$A^{(\alpha)} = Z'H^{(\alpha)}Z.$$

2. Le programme vérifie si les poids obtenus à l'étape 1 satisfont les critères.

   (a) Est-ce que $| nu_i^{(\alpha)} | \leq M$ pour tous i ?

   (b) Est-ce que

$$L_B \leq nw_i^{(\alpha)} \leq U_B$$

   pour tous i ?

   Si le critère (a) ou (b) n'est pas respecté pour n'importe quel i et que le programme n'a pas complété LI itérations, on passe à l'étape 3. Si les critères (a) et (b) sont satisfaits ou si le programme a complété LI itérations, les poids obtenu à cette étape sont produits pour analyse.

3. Les facteurs de contrôle $h_i^{(\alpha)}$, $i = 1, 2, \ldots, n$, sont modifiés. Posons

$$H^{(\alpha)} = H^{(\alpha-1)}G^{(\alpha)},$$

où

$$G^{(\alpha)} = \text{diag}(g_1^{(\alpha)}, g_2^{(\alpha)}, \ldots, g_n^{(\alpha)}),$$

$$g_i^{(\alpha)} = 1 \qquad 0 \leq d_i^{(\alpha)} < 0.5$$

$$= 1 - 0.8(d_i^{(\alpha)} - 0.5)^2 \qquad 0.5 \leq d_i^{(\alpha)} \leq 1$$

$$= 0.8(d_i^{(\alpha)})^{-1} \qquad d_i^{(\alpha)} > 1,$$

$$d_i^{(\alpha)} = 1.33[D_i^{(\alpha-1)}]^{-1} n \mid u_i^{(\alpha-1)} \mid,$$

$$D_i^{(\alpha-1)} = \min\{M, C_{\text{Li}}^{(\alpha-1)}\} \qquad \text{si } w_i^{(\alpha-1)} < v_i$$

$$= \min\{M, C_{\text{Bi}}^{(\alpha-1)}\} \qquad \text{si } w_i^{(\alpha-1)} \geq v_i,$$

$$C_{\text{Li}}^{(\alpha-1)} = \max\{| v_i^{-1}(1 + n\bar{u}_v^{(\alpha-1)})L_B - 1 |, 0.1\, M\},$$

$$C_{\text{Bi}}^{(\alpha-1)} = \max\{| v_i^{-1}(1 + n\bar{u}_v^{(\alpha-1)})U_B - 1 |, 0.1\, M\}.$$

On passe à l'étape 1 pour calculer de nouveaux poids de régression.

La constante 1.33 dans la définition de $d_i^{(\alpha)}$ et la constante 0.8 dans la définition de $g_i^{(\alpha)}$ ont été choisies pour accélérer la convergence. Les facteurs de contrôle $g_i^{(\alpha)}$ servent à sous-pondérer les observations suivant une distance par rapport à la moyenne de population.

La définition de $w^{(\alpha)}$ en (A.5) représente une autre manière d'écrire le vecteur des poids par les moindres carrés généralisés définis en (2.4) lorsque $\pi_i^{-1} = h_i^{(\alpha)}$.

BIBLIOGRAPHIE

AKKERBOOM, J.C., SIKKEL, D., et van HERK, H. (1991). Robust weighting of financial survey data. Article présenté à la réunion de l'Institut International de Statistique, Cairo, Egypt.

ALEXANDER, C.H. (1987). A model based justification for survey weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 183-188.

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

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

BRACKSTONE, G.J., et RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā*, Séries C, 97-114.

BUREAU OF THE CENSUS (1987). Current Population Survey, March 1987: Technical Documentation. Washington, D.C.

COCHRAN, W.G. (1942). Sampling theory when the units are of unequal sizes. *Journal of the American Statistical Association*, 37, 199-212.

COCHRAN, W.G. (1977). *Sampling Techniques*, 3ième Éd. New York: John Wiley.

COPELAND, K.R., PEITZMEIER, F.K., et HOY, C.E. (1987). Méthode alternative pour ajuster les estimations de la Current Population Survey aux chiffres de population. *Techniques d'enquête*, 13, 183-191.

COX, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.

COX, L.H., et ERNST, L.R. (1982). Controlled rounding. *INFOR*, 20, 423-432.

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

DEMING, W.E., et STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.

DARROCH, J.N., et RATCLIFF, D. (1972). Generalized iterative sealing for log-linear models. *Annals of Mathematical Statistics*, 43, 1470-1480.

EL-BADRY, M.A., et STEPHAN, F.F. (1985). On adjusting sample tabulations to census counts. *Journal of the American Statistical Association*, 50, 738-762.

FAGAN, J.T., GREENBERG, B.V., et HEMMIG, B. (1988). Controlled rounding of three dimensional tables. Statistical Research Division Report Census/SRD/RR-88/02. U.S. Bureau of the Census, Washington, D.C.

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

FULLER, W.A., KENNEDY, W., SCHNELL, D., SULLIVAN, G., et PARK, H.J. (1986). PC CARP. Statistical Laboratory, Iowa State University, Ames Iowa.

GOEBEL, J.J. (1976). Application of an iterative regression technique to a national potential cropland survey. *Proceedings of the Social Statistics Section, American Statistical Association*, 350-353.

HAMPEL, F.R. (1978). Optimally bounding the gross-error-sensitivity and the influence of position in factor space. *Proceedings of the Statistical Computing Section, American Statistical Association*, 59-64.

HIDIROGLOU, M.A. (1974). Estimation of regression parameters for finite populations. Thèse de doctorat non-publiée, Iowa State University, Ames, Iowa.

HIDIROGLOU, M.A., FULLER, W.A., et HICKMAN, R.D. (1976). SUPER CARP, Statistical Laboratory, Iowa State University, Ames, Iowa.

HUANG, E.T. (1978). Nonnegative regression estimation for sample survey data. Thèse de doctorat non-publiée, Iowa State University, Ames, Iowa.

HUANG, E.T., et FULLER, W.A. (1978). Nonnegative regression estimation for survey data. *Proceedings of the Social Statistics Section, American Statistical Association 1978*. 300-303.

HULLIGER, B. (1993). Robustification of the Horvitz-Thompson estimator. Document présenté à la 49ième session de l'Institut International de Statistique. Livre 1, 583-584.

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

IRELAND, C.T., et KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 169-188.

JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. Iowa Experiment Station Research Bulletin, 304.

KRASKER, W.A. (1980). Estimation in linear regression models with disparate data points. *Econometrica*, 48, 1333-1346.

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

LITTLE, R.J.A., et RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.

LUERY, D. (1986). Weighting survey data under linear constraints on the weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 325-330.

MALLOWS, C.L. (1983). Discussion of Huber: Mimimax aspects of bounded-influence regression. *Journal of the American Statistical Association*, 78, 77.

MICKEY, M.R. (1959). Some finite population unbiased ratio and regression estimators. *Journal of the American Statistical Association*, 54, 594-612.

OH, H.L, et SCHEUREN, F. (1987). Variante de la méthode itérative du quotient. *Techniques d'enquête*, 13, 221-232.

RAO, J.N.K. (1992). Estimating totals and distribution functions using auxiliary information at the estimation stage. Document présenté au Workshop on Users of Auxiliary Information in Surveys, Örebo, Sweden, Octobre, 1992.

ROYALL, R.M., et CUMBERLAND, W.G. (1981). The finite-population linear regression estimator and estimators of its variance - an empirical study. *Journal of the American Statistical Association*, 76, 924-930.

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

SAS INSTITUTE INC. (1989). SAS/STAT User's Guide, Version 6 Fourth Edition, Volume 1. Cary, NC: SAS Institute Inc.

SINGH, A.C. (1993). On weight adjustment in survey sampling. Document non publié. Statistique Canada, Ottawa, Canada.

STEPHAN, F.F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. *Annals of Mathematical Statistics*, 13, 166-178.

WATSON, D.J. (1937). The estimation of leaf areas. *Journal of Agricultural Science*, 27, 474-483.

WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.

ZIESCHANG, K.D. (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association*, 85, 986-1001.

Techniques d'enquête, juin 1994
Vol. 20, n° 1, pp. 91–98
Statistique Canada

Estimation du taux de sans-abri dans les régions rurales: une étude des régions non urbaines de l'Ohio

ELIZABETH A. STASNY, BEVERLY G. TOOMEY et RICHARD J. FIRST[1]

RÉSUMÉ

Depuis quelques temps, on consacre beaucoup d'effort pour dénombrer les sans-abri et en définir les caractéristiques. Cependant, on s'est concentré jusqu'à maintenant sur les sans-abri des régions urbaines. Dans cet article, nous décrivons les efforts qui ont été faits en vue d'estimer le taux de sans-abri dans les comtés non urbains de l'Ohio. Les méthodes qui servent à repérer les sans-abri et même la définition de la clochardise sont différentes dans les régions rurales, où les institutions qui accueillent les sans-abri sont moins nombreuses. De plus, l'emploi d'estimateurs ordinaires basés sur des échantillons d'enquête peut poser des difficultés dans une analyse axée sur les régions rurales, étant donné que ces estimateurs exigent généralement de grandes populations, de grands échantillons et de faibles fractions de sondage. Nous décrivons l'enquête qui a été faite dans les régions non urbaines de l'Ohio pour dénombrer les sans-abri et nous présentons l'étude de simulation qui a été effectuée dans le but d'évaluer l'utilité des estimateurs ordinaires de la proportion d'une population basés sur un échantillon en grappes stratifié.

MOTS CLÉS: Estimateur biaisé; estimateur par régression; petit échantillon; échantillon en grappes stratifié; simulation.

1. INTRODUCTION

La notion de sans-abri évoque souvent dans notre esprit l'idée des "marginaux de rues" et des "clochardes". Nous nous représentons des personnes qui dorment sur des bancs de parc ou des grilles à air ou dans des lieux de refuge pour sans-abri. Cependant, ces images stéréotypées viennent des grandes villes et ne décrivent pas nécessairement avec exactitude les conditions de vie des sans-abri en région rurale.

Bon nombre des études portant sur les sans-abri ont été réalisées dans les grands centres urbains. Par exemple, l'étude de 1987 du Urban Institute visait à dénombrer les sans-abri de 20 grandes villes américaines. Rossi a effectué une autre grande étude du même genre à Chicago. (Pour avoir un aperçu des méthodes d'enquête utilisées dans ces études et d'autres études qui ont servi à dénombrer des populations de sans-abri, prière de se référer à Burt et Taeuber (1991).)

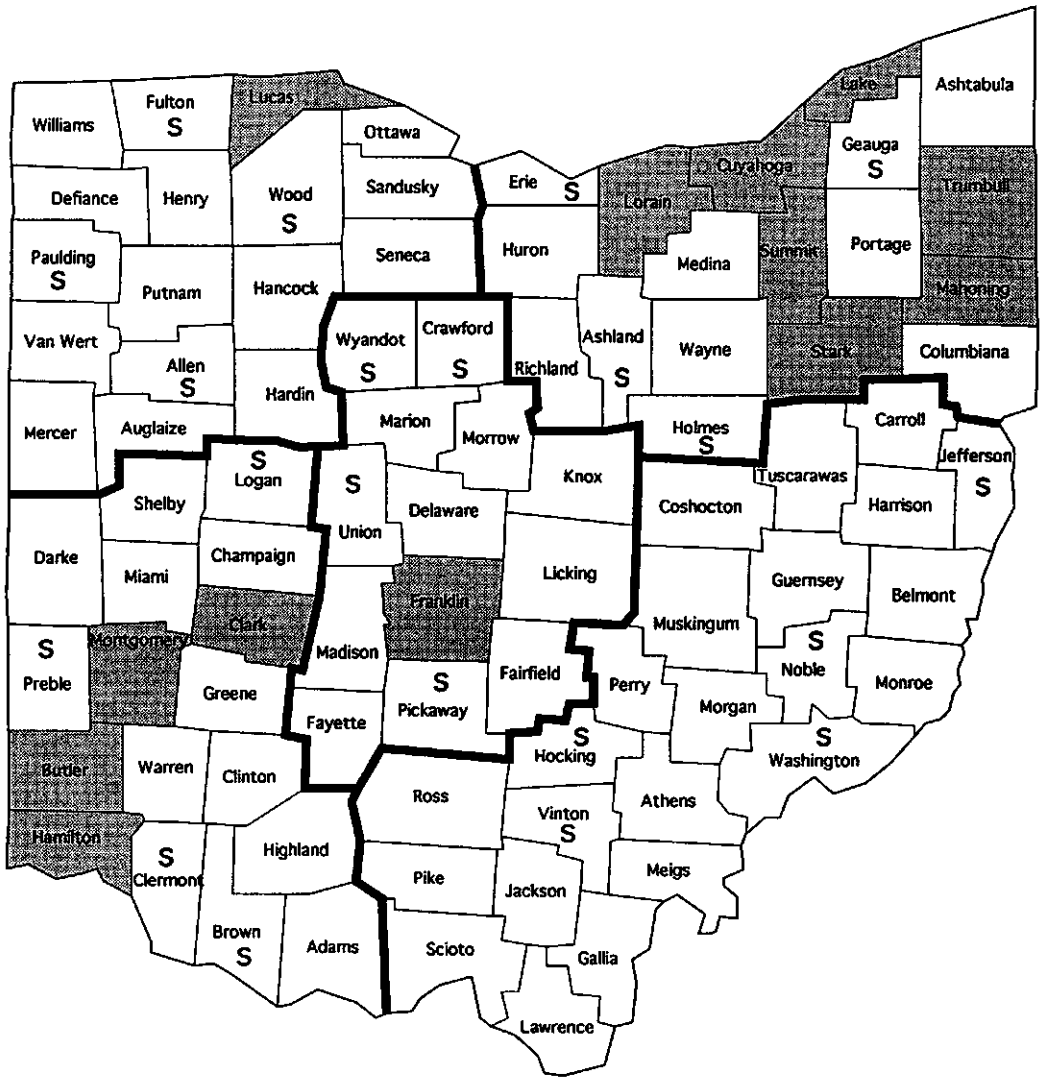
À l'occasion du recensement de la population de 1990 aux États-Unis, on a tenté de dénombrer les sans-abri par l'opération S-Night (Shelter and Street Night). On avait dressé spécialement à cette fin une liste des lieux de refuge des sans-abri à la grandeur du pays. Le plus haut dirigeant élu de plus de 39,000 municipalités rurales et urbaines devait fournir une liste des endroits (refuges, lieux publics, *etc.*) où les sans-abri passent la nuit. Les recenseurs ont procédé au dénombrement des sans-abri dans une seule nuit, celle du 20 mars. Il convient de souligner que l'opération S-Night avait pour but essentiellement de dénombrer les sans-abri; les données du recensement contiennent relativement peu d'information sur les caractéristiques des sans-abri. Pour plus de détails sur l'opération S-Night, prière de se référer à Taeuber et Siegel (1990).

Contrairement aux enquêtes faites dans les régions urbaines et à l'opération S-Night, l'enquête que nous décrivons ici avait pour but de repérer et de dénombrer les sans-abri des régions non urbaines, où qu'ils soient, et de recueillir de l'information qui permettrait de définir les caractéristiques de ces personnes. Dans la section qui suit, nous décrivons le plan de sondage de l'enquête de 1990 sur les sans-abri des régions rurales de l'Ohio. Nous donnons notre définition de la clochardise en région rurale et nous décrivons les méthodes utilisées pour repérer et interviewer les sans-abri. Dans la section 3, nous présentons les estimations du taux de sans-abri en région rurale que nous avons calculées à l'aide de l'estimateur ordinaire de la proportion basé sur un échantillon en grappes stratifié. Comme ces estimations sont susceptibles d'être biaisées, nous présentons aussi les résultats d'une étude de simulation qui avait pour but d'évaluer la taille probable du biais. Dans la section 4, nous considérons un estimateur par régression pour calculer le taux de sans-abri et nous comparons cet estimateur à celui utilisé dans la section 3. Enfin, la section 5 renferme les conclusions de notre étude.

2. DESCRIPTION DE L'ENQUÊTE

Le territoire de l'Ohio est divisé en 88 comtés, dont 13 sont des comtés urbains qui comprennent de grandes villes et 75 sont définis comme des comtés ruraux ou non urbains. Parmi ces 75 comtés, qui font l'objet de notre étude, il y a des comtés entièrement ruraux, des comtés qui ne sont pas voisins de comtés urbains et qui ont un chef-lieu moyennement peuplé, ainsi que des comtés suburbains (c.-à-d. voisins de comtés dans lesquels se trouvent de grandes régions métropolitaines).

[1] Elizabeth A. Stasny, Department of Statistics; Beverly G. Toomey et Richard J. First, College of Social Work, The Ohio State University, Columbus, Ohio 43210.



Nota: Les comtés ombrés sont les comtés urbains qui étaient exclus de l'étude. La lettre "S" désigne un comté de l'échantillon. Les traits gras définissent les cinq strates géographiques: Nord-Est, Nord-Ouest, Centre, Sud-Est et Sud- Ouest.

Figure 1. Carte de l'Ohio et ses comtés

Le plan de sondage qui a servi à l'enquête de 1990 a été choisi expressément pour faciliter la comparaison avec une étude qui a été réalisée en 1984 sur les sans-abri des régions rurales de l'Ohio (Roth et coll. 1985). Dans cette étude, on avait groupé les comtés de l'Ohio en cinq régions (Nord-Est, Nord-Ouest, Centre, Sud-Est et Sud-Ouest) et on avait prélevé 16 comtés ruraux au moyen d'un plan d'échantillonnage aléatoire stratifié. Les 21 comtés qui ont été échantillonnés pour l'enquête de 1990 comprenaient les 16 comtés de l'étude de 1984 plus un comté tiré au hasard dans chacune des cinq régions. (Notons que l'analyse des données de l'enquête de 1990 nous permet de croire que la stratification du territoire de l'Ohio ne contribue pas à améliorer l'estimation du taux de sans-abri en région rurale.) Dans la figure 1, on peut voir une carte de l'Ohio où sont indiqués les cinq régions, les comtés urbains et les comtés qui font partie de l'échantillon.

Dans les sous-sections suivantes, nous décrivons brièvement la méthodologie de l'enquête de 1990. Pour une description plus complète, prière de consulter First et coll. (1993) ainsi que Toomey et coll. (1993).

2.1 Personnel d'enquête

Nous avons tenté de recenser tous les sans-abri des 21 comtés de l'échantillon. Comme il n'existe pas nécessairement de lieux de refuge ou de rencontre pour les sans-abri dans les régions non urbaines, l'enquête s'est étalée sur une période de six mois et a fait appel à un réseau de conseillers pour trouver les sans-abri. La période d'enquête

a débuté avec la première semaine complète de février 1990. Dans chaque comté de l'échantillon, un réseau d'aiguillage s'occupait d'identifier et de trouver les sans-abri. Chaque réseau était sous la surveillance d'un coordonnateur local. Des enquêteurs principaux surveillaient le travail du coordonnateur de comté et des employés du bureau central. Par des appels téléphoniques aux deux semaines et des visites sur le terrain, les enquêteurs principaux contrôlaient la collecte des données afin d'assurer l'uniformité des procédés et la qualité des opérations.

Les conseillers et les intervieweurs, qui avaient été choisis à cause de leur bonne connaissance du comté dans lequel ils travaillaient, s'occupaient d'identifier les gens qui répondaient à la définition d'un sans-abri. Parmi les conseillers, on retrouvait des pasteurs, des employés d'hôpital, des dirigeants d'organismes de bienfaisance, des représentants élus, des animateurs de la communauté, des barmans, des commis d'hôtel, des employés de laverie automatique, de même que des fournisseurs de services professionnels tels que les membres des départements de santé communautaire, les bibliothécaires, les conseillers agricoles, les employés des postes, les ministres du culte, les rangers, les groupes d'action communautaire, les travailleurs sociaux de cas, les travailleurs de santé mentale et les agents de la paix. Quant aux intervieweurs, ils étaient une centaine à réaliser les interviews auprès des sans-abri, après avoir assisté à une séance de formation de quatre heures et s'être vu remettre un guide des opérations sur le terrain. Les interviews se sont déroulées dans des bureaux, des petits restaurants, des chambres de motel, des automobiles, des parcs publics, des granges, des laveries automatiques, des bars et sous des viaducs de chemin de fer. Les intervieweurs avaient été informés des ressources communautaires existantes et avaient la compétence nécessaire pour orienter éventuellement des répondants vers les services appropriés. De plus, les intervieweurs disposaient d'un certain budget au cas où ils seraient appelés à dépanner un répondant (moins de 600 $ ont été dépensés de cette manière). Il fallait limiter l'aide offerte par l'intermédiaire des intervieweurs pour ne pas inciter des gens à se faire passer pour des sans-abri.

2.2 Définition de la clochardise en région rurale

On s'est servi de questions filtre pour identifier les sans-abri. La définition de la clochardise utilisée pour cette étude différait nécessairement quelque peu de celle qui avait servi pour les études en région urbaine. En effet, il existe moins de lieux de refuge pour les itinérants dans les régions rurales. Un répondant était reconnu comme un sans-abri s'il prétendait ne pas avoir de domicile fixe et si, la nuit précédant l'interview, il avait dormi soit 1) à la belle étoile, soit 2) dans un refuge pour sans-abri, soit 3) dans un hôtel ou un motel bon marché, pourvu que la durée de séjour réelle ou prévue fût de 45 jours ou moins, ou 4) dans un autre lieu particulier, pourvu que la durée de séjour réelle ou prévue fût de 45 jours ou moins. La quatrième catégorie comprenait les gens qui séjournaient dans des hangars, des granges, de vieux autobus ou de vieilles roulottes sans eau ni électricité, pourvu qu'ils n'en étaient pas les propriétaires et qu'ils ne payaient pas de loyer pour demeurer là. Étaient reconnues aussi comme des sans-abri les personnes qui résidaient temporairement chez des amis ou des parents, pourvu qu'elles étaient arrivées depuis moins de 46 jours, qu'elles ne faisaient pas partie à proprement parler du ménage et qu'elles prévoyaient quitter dans un délai de 45 jours. Les personnes qui séjournaient dans des refuges pour femmes battues, des hôpitaux, des prisons, des camps de travailleurs migrants, *etc.* n'étaient pas comptées parmi les sans-abri à moins qu'elles n'eussent aucun endroit où aller au moment de quitter l'établissement.

On peut comparer notre définition de la clochardise avec celle utilisée dans les études sur les sans-abri en région urbaine. C'est dans la Stewart B. McKinney Homeless Assistance Act (1987) que l'on trouve les éléments de la définition de la clochardise au sens des études sur les sans-abri en région urbaine. Selon cette loi, un ou une sans-abri est "une personne qui n'a pas d'endroit fixe, régulier et convenable où passer la nuit et qui doit compter surtout sur 1) les établissements supervisés à caractère public ou privé qui ont pour mission d'offrir des services d'hébergement temporaire (c.-à-d. établissements de bien-être, habitations collectives et logements de transition pour les personnes souffrant de troubles psychiques); 2) les institutions qui offrent des services d'hébergement temporaire aux personnes qui doivent être placées en établissement; ou 3) des endroits publics ou privés qui ne sont pas considérés normalement comme des lieux d'hébergement pour la nuit" (TRADUCTION). Comme le laisse entendre l'article de Rossi et coll. (1987), de cette définition découle la notion de "sans-abri au sens littéral". La définition ci-dessus ne recouvre pas, par exemple, les sans-abri qui demeurent temporairement chez des amis ou des parents. Or, nous avons choisi de tenir compte de cette catégorie de sans-abri dans notre étude. Selon notre analyse, environ le tiers des personnes qui ont été dénombrées lors de notre enquête n'auraient pas pu l'être si nous avions utilisé la définition de la clochardise pour les régions urbaines. On ne peut dire de combien augmenterait le chiffre estimé des sans-abri dans les régions urbaines si on tenait compte des personnes qui demeurent temporairement chez des amis ou des parents.

2.3 Période d'interview

La période de six mois utilisée pour dénombrer les sans-abri dans les régions rurales tranche avec la période d'une journée qui est utilisée normalement dans les enquêtes faites en région urbaine. Dans une analyse de sept enquêtes ayant pour objet les sans-abri, Burt et Taeuber (1991) remarquent que la période d'interview pouvait correspondre à une seule journée (en soirée notamment) ou pouvait s'étendre sur une ou deux semaines et les interviews avaient lieu à un seul endroit. Dans la plupart de ces enquêtes, on cherchait les sans-abri à interviewer dans les refuges, les soupes populaires, les bâtiments désaffectés ou les autres endroits semblables. Comme, en région rurale, les refuges pour sans-abri et les soupes populaires sont plus rares, il est plus difficile de trouver les sans-abri, ce pourquoi une période d'enquête plus longue est recommandée.

Pour faciliter la comparaison avec les enquêtes qui durent une journée ou une semaine, on a demandé aux sans-abri qui ont participé à notre enquête depuis combien de temps ils étaient dans cette condition. Grâce à cette information, nous avons pu déterminer, pour chaque comté de l'échantillon, le nombre de personnes qui étaient des sans-abri dans la première semaine de l'enquête, soit la première semaine complète de février 1990.

Dans la section 3, nous présentons des estimations du taux de sans-abri pour la période d'une semaine et la période de six mois. Le taux pour la période de six mois comprend quiconque répondait à la définition du sans-abri à un moment quelconque dans la période d'interview. Le taux pour la période d'une semaine comprend les personnes interviewées au cours des six mois qui disaient avoir été des sans-abri durant la première semaine complète de février.

Pour éviter de compter un répondant deux fois durant la période de six mois, on a attribué à chaque participant un numéro d'identification personnel formé de la date de naissance de la personne, d'un code pour le sexe et des trois premières lettres du nom de famille. Un seul enregistrement en double était présent dans la base de données, et il a été supprimé. (Nous ne disposons pas de renseignements sur les enregistrements en double observés sur le terrain.) Compte tenu des mesures qui ont été prises pour limiter les doubles comptes, nous croyons que si le processus de collecte des données devait comporter un biais, ce serait un biais par défaut (c.-à-d. sous-estimation du nombre de sans-abri en région rurale). Durant la période d'interview de six mois, 1,100 adultes et 480 enfants ont été identifiés comme des sans-abri dans les 21 comtés de l'échantillon.

2.4 Questionnaire d'enquête

Si les réponses données aux questions filtre par une personne permettaient d'identifier celle-ci comme un sans-abri, on demandait à cette personne de répondre à un questionnaire qui avait pour but de recueillir des données sur l'individu même et sur son expérience de vie. Sur les 1,100 adultes qui avaient été identifiés comme des sans-abri, 919 ont répondu entièrement au questionnaire. Bien que l'objet premier de cet article soit l'estimation du nombre de sans-abri en région rurale, nous décrivons brièvement ci-dessous le questionnaire qui a servi à recueillir de l'information dans le but de connaître les caractéristiques des sans-abri.

Le questionnaire intégral comprenait trois sections. Dans la première, on trouvait des questions sur les caractéristiques démographiques et l'expérience de vie (par exemple, motifs de la clochardisation, recours à des services de santé mentale et à d'autres services sociaux, antécédents de travail, consommation de drogues et d'alcool, structure de la famille, et bien-être général). La deuxième section du questionnaire comprenait dix échelles (par exemple, dépression-anxiété, désorientation-déficience de la mémoire, arriération-carence affective) tirées du tableau des états psychiatriques de Spitzer et coll. (1970) (Psychiatric Status Schedule). La troisième section servait à faire le bilan de l'interview; remplie par l'intervieweur, elle contenait des renseignements tels que le lieu de l'interview et les caractéristiques du répondant (exemple: sexe, comportements anormaux), ainsi qu'une appréciation des réponses fournies. Les conclusions tirées des données de la troisième section sont résumées dans First et coll. (1994).

3. ESTIMATIONS DU TAUX DE SANS-ABRI

3.1 Définition de l'estimateur

Les estimations régionales du taux de sans-abri en région rurale ont été calculées à l'aide de l'estimateur ordinaire de la proportion basé sur un échantillon en grappes stratifié avec tailles de grappe inégales. Dans ce cas-ci, la grappe correspond au comté, la taille de la grappe, à la population du comté et la taille de la strate, à la population de la région. L'estimateur est défini de la façon suivante:

Pour la région i, le taux estimé de sans-abri est r_i, où

$$r_i = \frac{\text{nombre de sans-abri dans les comtés ruraux échantillonnés de la région } i}{\text{population totale des comtés ruraux échantillonnés de la région } i}.$$

Le taux estimé de sans-abri pour l'État est donc:

$$r_{\text{État}} = \frac{\sum_i [r_i \times \text{population des comtés ruraux de la région } i],}{\text{population totale des comtés ruraux de l'Ohio}},$$

où la sommation s'étend aux cinq régions géographiques illustrées dans la figure 1. Les chiffres de population pour les 75 comtés non urbains sont tirés du recensement de 1990.

Le tableau 1 contient les taux estimés de sans-abri (nombre de sans-abri par tranche de 10,000 habitants) pour la période d'une semaine et la période de six mois.

Tableau 1

Taux estimés de sans-abri dans les régions rurales de l'Ohio (par tranche de 10,000 habitants)

Région	Taux pour la période d'une semaine (du 4 au 10 février 1990)	Taux pour la période de six mois (février à juillet 1990)
État	5.68 (0.99)	14.00 (2.09)
Nord-Est	3.44 (0.79)	12.00 (2.19)
Nord-Ouest	5.21 (3.51)	12.77 (5.18)
Centre	5.85 (1.86)	12.11 (3.05)
Sud-Est	6.89 (1.93)	15.90 (5.91)
Sud-Ouest	7.25 (2.44)	16.78 (5.32)

Nota: Les erreurs types figurent entre parenthèses à côté de chaque estimation.

Comme l'estimateur ci-dessus est le rapport de deux variables aléatoires, soit le nombre de sans-abri et l'effectif de grappes échantillonnées, il est entaché d'un biais (voir, par exemple, Cochran (1977)). Le biais diminue à mesure que la taille de l'échantillon (en l'occurrence, le nombre de comtés échantillonnés) augmente. Étant donné la faible taille de notre échantillon, nous reconnaissons que nos estimations peuvent être biaisées. Par ailleurs, la fraction de sondage que nous avons utilisée est grande parce que le nombre de comtés ruraux est peu élevé. Nous allons donc chercher à évaluer la taille probable du biais contenu dans nos estimations. (Notons que les erreurs types qui figurent dans le tableau 1 pourraient aussi être inexactes à cause de la faible taille de l'échantillon.)

3.2 Étude de simulation

Nous avons eu recours à une étude de simulation afin d'évaluer la taille probable du biais contenu dans nos estimations. Tout d'abord, nous avons généré cinq "populations" avec, dans chaque cas, le nombre de sans-abri pour chacun des 75 comtés non urbains de l'Ohio. Pour les 21 comtés qui faisaient partie de l'échantillon, le nombre observé de sans-abri tenait lieu de chiffre officiel; pour les 54 autres comtés, les chiffres ont été produits aléatoirement de la manière décrite ci-dessous. Notons que les chiffres simulés sont des chiffres pour la période de six mois.

Nous avons créé la première population en générant le logarithme naturel du taux de sans-abri à partir d'une distribution normale unique. Nous avons choisi le logarithme du taux parce que l'histogramme des taux observés pour les 21 comtés de l'échantillon est très asymétrique alors que l'histogramme des logarithmes des taux est à peu près symétrique. La moyenne des logarithmes des taux observés est 2.465, avec un écart type de 0.7154. Nous avons donc généré aléatoirement (à l'aide du progiciel S) les logarithmes des taux de sans-abri à partir d'une distribution normale ayant cette moyenne et cet écart type. Après que les logarithmes ont été générés pour les 54 comtés qui ne faisaient pas partie de l'échantillon, ils ont servi, avec les chiffres de population du recensement de 1990 pour chaque comté, à établir le nombre simulé des sans-abri pour les 54 comtés.

La deuxième population simulée a été créée de la même manière que la première, sauf qu'il y avait une distribution normale pour chacune des cinq régions géographiques de l'Ohio. Les paramètres de cette distribution, pour chaque région, étaient la moyenne et l'écart type des logarithmes des taux de sans-abri pour les comtés de l'échantillon. Là encore, les logarithmes simulés ont servi à établir le nombre de sans-abri pour les 54 comtés ruraux non échantillonnés.

La troisième population a été générée au moyen d'une régression du taux de sans-abri (par tranche de 10,000 habitants) par rapport au pourcentage de personnes âgées pour chaque comté de l'échantillon. (Le choix de cette variable explicative est déterminé par l'usage d'un estimateur par régression, comme on l'explique dans la section 4.) Le modèle de régression ajusté est

$$\widehat{\text{taux}} = -9.02 + 2.32\,[\%\text{ de personnes âgées}],$$

avec $R^2 = 0.197$, $\sqrt{\text{EQM}} = 9.03$, et valeur $p = 0.044$ pour le test F global appliqué à la droite de régression. On a créé la population en estimant, pour chaque comté non échantillonné, le taux de sans-abri à partir du pourcentage de personnes âgées, puis en ajoutant un terme d'erreur normale aléatoire. Comme le graphique des résidus de la droite de régression donnait à penser que la variance résiduelle est plus élevée pour les comtés où la proportion des personnes âgées est plus forte, on a généré les termes d'erreur aléatoire à partir de deux distributions normales différentes, selon que la proportion des personnes âgées dans un comté était inférieure ou supérieure à 10%. Dans l'un et l'autre cas, c'est l'écart type des résidus qui servait de paramètre pour la distribution normale.

La quatrième population a été générée au moyen d'une régression du taux de sans-abri (par tranche de 10,000 habitants) par rapport au pourcentage de personnes âgées pour chaque comté de l'échantillon et aux indicateurs des régions de l'État dans lesquelles se trouvent les comtés. Si nous désignons par I_{NE}, I_{NW}, I_C, et I_{SE} les variables indicatrices pour les régions Nord-Est, Nord-Ouest, Centre et Sud-Est respectivement, le modèle de régression ajusté est

$$\widehat{\text{taux}} = -10.40 + 3.23\,[\%\text{ de personnes âgées}] - 6.47 I_{NE} - 8.55 I_{NW} - 8.64 I_C - 14.25 I_{SE},$$

avec $R^2 = .407$ (R^2 corrigé $= .210$), $\sqrt{\text{EQM}} = 8.73$, et valeur $p = 0.127$ pour le test F global appliqué à la droite de régression. On a créé la population en estimant, pour chaque comté non échantillonné, le taux de sans-abri au moyen de l'équation de régression, puis en ajoutant un terme d'erreur normale aléatoire. Là encore, le graphique des résidus de la droite de régression donnait à penser que la variance résiduelle est plus élevée pour les comtés où la proportion des personnes âgées est plus forte. On a donc généré les termes d'erreur aléatoire à partir de deux distributions normales différentes, selon que la proportion des personnes âgées dans un comté était inférieure ou supérieure à 10%. Dans l'un et l'autre cas, c'est l'écart type des résidus qui, encore une fois, servait de paramètre pour la distribution normale.

La cinquième population simulée diffère quelque peu des quatre autres. Elle a été générée à l'aide d'un modèle de régression qui permet de prévoir directement le nombre de sans-abri à partir de la taille de la population du comté. Le modèle de régression ajusté est:

$$\widehat{\text{nombre de sans-abri}} = 13.23 + 0.001154\,[\text{population}],$$

avec $R^2 = 0.386$, $\sqrt{\text{EQM}} = 54.29$, et valeur $p = 0.003$ pour le test F global appliqué à la droite de régression. On a créé la population en estimant, pour chaque comté non échantillonné, le nombre de sans-abri au moyen de l'équation de régression ajustée, puis en ajoutant un terme d'erreur normale aléatoire. Comme le graphique des résidus donnait à penser que la variance résiduelle est plus

élevée pour les comtés plus fortement peuplés, on a généré les termes d'erreur aléatoire à partir de deux distributions normales différentes, selon que la population du comté était inférieure ou supérieure à 30,000. Dans l'un et l'autre cas, c'est l'écart type des résidus qui servait de paramètre pour la distribution normale.

Une fois les cinq populations obtenues, nous nous sommes servis de chacune d'elles pour évaluer la taille du biais contenu dans les estimations du taux de sans-abri en région rurale. Comme nous avions créé une "population" complète, nous pouvions calculer dans les cinq cas le "vrai" taux de sans-abri pour tout l'État et pour les cinq régions géographiques.

Dans la simulation, des échantillons formés de 21 comtés ruraux ont été prélevés selon le plan d'échantillonnage stratifié utilisé dans l'enquête de 1990, c'est-à-dire tirage aléatoire sans remise de quatre comtés dans chacune des régions suivantes: Nord-Est, Nord-Ouest, Centre et Sud-Ouest, et tirage de cinq comtés dans la région Sud-Est. Ensuite, les taux estimés de sans-abri ont été calculés pour les cinq régions et l'État en général au moyen des formules définies dans la section 3.1. Enfin, les estimations ainsi obtenues ont été comparées aux taux de sans-abri pour la population simulée afin d'évaluer le biais contenu dans les estimations. Cette opération en trois étapes (échantillonnage, calcul d'estimations et évaluation du biais) a été répétée un million de fois avec remise pour chaque population simulée. (Le nombre d'échantillons possibles dépasse 7.15×10^{15}.) La même série de nombres aléatoires a servi au tirage des échantillons pour les cinq populations. Les résultats de la simulation figurent dans le tableau 2.

Tableau 2

Biais contenu dans le taux estimé de sans-abri (nombre de sans-abri par tranche de 10,000 habitants) pour cinq populations simulées (données basées sur 1,000,000 d'échantillons simulés)

	Population				
	1	2	3	4	5
ÉTAT	0.0406	0.1308	0.2618	0.2433	0.2547
	(2.056)	(1.759)	(2.144)	(1.807)	(1.605)
RÉGION					
N-E	−0.0406	−0.0379	0.1538	0.0317	0.0993
	(3.333)	(2.923)	(3.748)	(4.034)	(1.937)
N-O	−0.0578	−0.2948	0.0529	0.0254	0.3234
	(3.591)	(3.194)	(3.474)	(3.460)	(4.249)
C	−0.2442	0.2700	0.3974	0.1362	0.1901
	(3.122)	(3.762)	(3.426)	(2.260)	(2.869)
S-E	−0.1034	−0.0279	−0.1132	−0.1798	0.0427
	(6.512)	(4.298)	(6.600)	(3.892)	(3.973)
S-O	0.6184	0.8093	0.9196	1.277	0.6716
	(4.215)	(4.990)	(4.610)	(5.173)	(4.274)

Nota: L'écart type de la distribution d'échantillonnage simulée de l'estimateur figure entre parenthèses au-dessous de chaque valeur.

On voit d'après le tableau 2 que le biais de l'estimation du taux de sans-abri pour l'État équivaut environ au centième de cette valeur estimée. (Rappelons que le taux estimé de sans-abri pour la période de six mois est environ 14 par tranche de 10,000 habitants pour l'ensemble de l'État. Les taux correspondants pour les populations simulées varient entre 13 et 15.) En ce qui concerne les régions, le biais équivaut aussi à peu près au centième de l'estimation régionale même si cette valeur repose sur un échantillon beaucoup plus petit. Ces résultats donnent à penser que le biais contenu dans nos estimations d'enquête devrait être relativement faible.

Comme le laissait prévoir le faible nombre de comtés dans l'échantillon, la variance de la distribution d'échantillonnage de l'estimateur est assez élevée. L'écart type des estimations obtenues par l'étude de simulation est environ 10 fois plus élevé que le biais. (Pour chacune des cinq simulations, les écarts types du million d'estimations sont du même ordre de grandeur que l'erreur type de l'estimation indiquée dans le tableau 1.) Ces observations permettent de croire que le biais des estimations d'enquête sera plutôt négligeable par rapport à l'erreur type de ces estimations.

Finalement, nous avons évalué la forme de la distribution d'échantillonnage de notre estimateur en examinant des histogrammes basés sur le million d'estimations calculées dans chacune des cinq simulations. Les histogrammes paraissaient symétriques, en forme de cloche, et ressemblaient étrangement à des histogrammes de données normales, ce qui nous permet de croire que les intervalles de confiance basés sur l'approximation normale seront assez précis.

4. ESTIMATEUR PAR RÉGRESSION

Des organismes comme le Bureau of the Census mettent à notre disposition de nombreuses données sur la conjoncture économique des comtés. Nous espérions pouvoir utiliser une partie de cette information dans un estimateur par régression afin d'améliorer nos estimations du taux de sans-abri. Nous avons donc cherché à définir un modèle de régression qui mettrait en relation le nombre de sans-abri dans un comté, ou le taux de sans-abri, et divers prédicteurs qui seraient susceptibles, selon nous, d'expliquer la clochardise. Ces prédicteurs étaient les suivants: population du comté, variation relative de la population entre 1980 et 1990, taux de chômage, proportion des personnes âgées dans la population, dépenses au titre de l'aide sociale, gains hebdomadaires moyens, proportion des maisons de rapport dans le parc immobilier, loyer médian, taux de pauvreté, pourcentage des ménages qui ont à leur tête une femme, pourcentage des terres réservées à l'agriculture, valeur moyenne des exploitations agricoles, revenu moyen par exploitation, rapport des emplois dans le secteur manufacturier aux emplois dans le secteur agricole, indicateur des codes de Beale système de classification ayant pour objet le degré de ruralité (voir Thomas 1977), et indicateurs régionaux.

Aucune de ces variables, ni aucune combinaison de celles-ci, ne s'est révélée un bon prédicteur du nombre de sans-abri ou du taux de clochardise. La seule variable qui soit digne de mention est la proportion des personnes âgées dans la population, qui a d'ailleurs servi à générer la troisième population simulée décrite dans la section 3.2, mais cette variable explique moins de 20% de la variation du taux de clochardise. Aucune autre variable ne pouvait expliquer convenablement la clochardise et nous n'avons pu trouver d'autre modèle de régression qui soit acceptable. C'est pourquoi nous nous sommes servis de la proportion des personnes âgées dans l'estimateur par régression du taux de sans-abri en région rurale pour l'État. Notons que cette variable est un prédicteur plausible du taux de sans-abri car l'existence de conditions économiques difficiles dans un comté rural semble être un motif de migration pour les jeunes; les personnes âgées, elles, ne bougent pas et leur poids démographique dans le comté augmente graduellement. C'est pourquoi la proportion de personnes âgées dans la population d'un comté peut être un indicateur d'une conjoncture difficile et d'une émigration interne. Cependant, nous ne pouvons exclure l'hypothèse que le lien observé dans nos données entre cette variable et le taux de sans-abri soit uniquement l'effet du hasard. Nous reconnaissons aussi que les erreurs qui surviennent immanquablement dans la collecte des données sur le terrain (par ex., effet de l'intervieweur, quantité de services offerts, facteurs géographiques) peuvent atténuer le lien qui devrait exister théoriquement entre le taux de sans-abri et certaines variables.

L'estimateur par régression combiné (voir, par exemple, Cochran (1977)) a donné un taux de sans-abri en région rurale (pour l'État) de 14.85 par tranche de 10,000 habitants, avec une erreur type de 1.64. Cette valeur se rapproche de l'estimation qui figure dans le tableau 1 pour la période de six mois, c'est-à-dire 14.00, avec une erreur type de 2.09. Comme l'estimateur par régression est lui aussi un estimateur biaisé, dont le biais diminue à mesure que la taille de l'échantillon augmente, nous avons effectué de nouveau une étude de simulation pour évaluer ce biais.

Tableau 3

Comparaison d'estimateurs du taux de sans-abri pour l'État (par tranche de 10,000 habitants) (données basées sur 1,000,000 d'échantillons tirés à répétition dans deux populations simulées)

	Estimateur de la section 3.1		Estimateur par régression	
	Population		Population	
	3	4	3	4
Biais moyen	0.2618	0.2433	1.7115	0.8360
Écart type	2.144	1.807	1.820	1.246
EQM	4.664	3.325	6.242	2.250

Nous nous sommes servis des troisième et quatrième populations simulées de la section 3.2 pour réaliser cette étude de simulation parce que ces populations avaient été générées à l'aide d'un modèle de régression qui faisait intervenir la proportion des personnes âgées. Comme pour la simulation de la section 3.2, nous avons calculé le biais de l'estimateur pour 1 million d'échantillons en grappes stratifiés tirés avec remise dans chaque population. La même série de nombres aléatoires a servi à la formation des échantillons pour les deux populations. Les résultats de l'étude de simulation pour l'estimateur de la section 3.1 et l'estimateur par régression sont résumés dans le tableau 3.

On remarque que l'estimateur par régression a un biais moyen plus élevé que l'estimateur de proportion basé sur un échantillon en grappes stratifié. En revanche, l'écart type de la distribution d'échantillonnage de l'estimateur par régression semble légèrement plus petit que celui de la distribution d'échantillonnage de l'autre estimateur pour chacune des deux populations simulées. Pour ce qui est de l'erreur quadratique moyenne (EQM), l'estimateur par régression est supérieur à l'estimateur de proportion dans un cas mais moins efficace dans l'autre. Par conséquent, les résultats du tableau 3 ne nous permettent pas de déterminer clairement quel estimateur conviendrait le mieux.

Puisque l'estimateur par régression ne représente pas une amélioration notable par rapport à l'estimateur de proportion et que son biais semble en moyenne plus élevé, et puisqu'il se peut que le choix de la variable "proportion des personnes âgées" (parmi tant d'autres) ait été uniquement l'effet du hasard, nous avons décidé d'utiliser l'estimateur de la section 3 pour le calcul du taux de sans-abri en région rurale.

5. CONCLUSIONS

Les données nationales sur la clochardise le plus souvent citées se trouvent dans Burt et Cohen (1989); ce rapport estime le taux de sans-abri en région urbaine à 37.4 par tranche de 10,000 habitants dans les villes de plus de 100,000 habitants et à 9 par tranche de 10,000 à l'extérieur des SMA (Standard Metropolitan Areas). L'étude qui a été faite sur les sans-abri dans les régions non urbaines de l'Ohio donne un taux de clochardise d'environ 14 par tranche de 10,000 pour une période de six mois et un taux de 5.68 par tranche de 10,000 pour une période d'une semaine.

Les résultats de notre étude de simulation permettent de croire que le biais de l'estimateur ordinaire de la proportion basé sur un petit échantillon en grappes est probablement négligeable si on le compare surtout à l'erreur type de l'estimation. Le biais des estimations relatives aux cinq régions géographiques de l'Ohio est, lui aussi, relativement faible. D'après les résultats de l'étude de simulation, il n'y a pas lieu de croire que les biais et les erreurs statistiques gâtent les principaux résultats de l'enquête sur les sans-abri en région rurale.

L'analyse de régression que nous avons effectuée afin d'estimer le nombre de sans-abri dans les comtés échantillonnés donne à penser qu'il est difficile d'expliquer le nombre de sans-abri dans des comtés non urbains à l'aide de variables économiques et démographiques dont on pourrait croire qu'elles ont un rapport avec la clochardise. La difficulté vient peut-être du fait que les comtés sont tellement différents les uns des autres, à cause de leur situation géographique par rapport aux agglomérations et des caractéristiques économiques propres à cette situation, qu'il est impossible de trouver une formule convenable pour la stratification des comtés non urbains de l'Ohio. L'utilisation d'un échantillon stratifié selon un critère géographique ne semble pas réduire la variance de l'estimation et notre analyse de régression n'indique pas d'autre variable de stratification possible. On pourrait bien observer la même chose pour d'autres États, quoiqu'il soit possible d'effectuer une stratification selon une variable quelconque pour l'ensemble des États-Unis par exemple.

REMERCIEMENTS

Cette étude a été rendue possible en partie grâce à une subvention du National Institute of Mental Health (n° R01MH46111) et grâce aux départements de la santé et de la santé mentale de l'Ohio. Les auteurs tiennent à souligner l'excellent travail de Mme H.C. Tsai, qui a effectué la plupart des calculs nécessaires à cette étude. Ils remercient aussi deux arbitres pour les commentaires utiles qu'ils ont formulés sur une version antérieure de cet article.

BIBLIOGRAPHIE

COCHRAN, W.G. (1977). *Sampling Techniques*, (3e édition). New York: Wiley.

BURT, M.R., et COHEN, B. (1989). *America's Homeless: Numbers, Characteristics and Programs that Serve Them*. Urban Institute Report 89-3. Washington, DC: Urban Institute Press.

BURT, M.R., et TAEUBER, C. M. (1991). Overview of seven studies that counted or estimated homeless populations dans *Enumerating Homeless Persons: Methods and Data Needs*. Proceedings of Census Bureau/ Interagency Council/ Department of Housing and Urban Development Conference, November 29-30, 1990, édité par C.M. Taeuber, Washington, D.C.: US Department of Commerce, 30-76.

FIRST, R.J., TOOMEY, B.G., RIFE, J.C., et STASNY, E.A. (1994). Outside of the City: A Statewide Study of Homelessness in Nonurban/Rural Areas. Final report for NIMH Grant #R01MH46111. Columbus, OH: College of Social Work, The Ohio State University.

ROSSI, P.H., WRIGHT, J.D., FISCHER G.A., et WILLIS, G. (1987). The urban homeless: estimating composition and size. *Science*, 235, 1336-1341.

ROTH, D., BEAN, J., LUST, N., et SAVEANU, T. (1985). *Homelessness in Ohio: A Study of People in Need*. Ohio Department of Mental Health, Columbus, Ohio.

SPITZER, R., ENDICOTT, J., et COHEN, J. (1970). The psychiatric status schedule: A technique for evaluation psychopathology and impairment in role functioning. *Archives of General Psychiatry*, 23, 41-55.

TAEUBER, C.M., et SIEGEL, P.M. (1990). Counting the Nation's Homeless Population in the 1990 Census. Document présenté au Conference on Enumerating Homeless Persons: Methods and Data Needs, Washington, D.C., November 29, 1990.

THOMAS, D.W. (1977). Beale Code Revisions Based on Those Devised by Fuguitt and Beale. The Rural Turnaround in Ohio: 1970-1975. E.S.S. #560. Columbus, OH: Department of Agricultural Economics and Rural Sociology, The Ohio State University.

TOOMEY, B.G., FIRST, R.J., GREENLEE, R., et CUMMINS, L. (1993). Counting the rural homeless population: Methodological dilemmas. *Social Work Research and Abstracts*, 24, 23-27.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 19, No. 1 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only; entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.