TECHNIQUES D' ENQUÊTE

Catalogue 12-001

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

JUIN 1996

•

VOLUME 22

•

NUMÉRO 1

Statistique Canada Statistics Canada

Canada

TECHNIQUES D' ENQUÊTE

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

JUIN 1996 • VOLUME 22 • NUMÉRO 1

Publication autorisée par le ministre
responsable de Statistique Canada

© Ministre de l'Industrie, 1996

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 du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Juin 1996

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

N° 12-001-XPB au catalogue
Périodicité : semestrielle

ISSN 0714-0045

Ottawa

10102174 28
STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA

Statistique Canada Statistics Canada

Canada

800

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
D. Binder
G.J.C. Hole
F. Mayda (Directeur de la Production)
C. Patrick
R. Platek (Ancien président)
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*
J.-C. Deville, *INSEE*
J.D. Drew, *Statistique Canada*
J.-J. Droesbeke, *Université Libre de Bruxelles*
W.A. Fuller, *Iowa State University*
M. Gonzalez, *U.S. Office of Management and Budget*
R.M. Groves, *University of Maryland*
M.A. Hidiroglou, *Statistique Canada*
D. Holt, *Central Statistical Office, U.K.*
G. Kalton, *Westat, Inc.*
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, *George Washington University*
J. Sedransk, *State University of New York*
C.J. Skinner, *University of Southampton*
P.J. Waite, *U.S. Bureau of the Census*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

Rédacteurs adjoints J. Denis, M. Latouche, H. Mantel et D. Stukel, *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 faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes-ménages, 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-XPB 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 Statistique Canada, Division des opérations et de l'intégration, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research et la Société Statistique du Canada.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada
Volume 22, numéro 1, juin 1996

TABLE DES MATIÈRES

Dans ce numéro 1

D.G. STEEL, D. HOLT et M. TRANMER
Inférences au niveau unitaire à partir de données agrégées 3

D.A. BINDER
Méthodes de linéarisation pour les échantillons à une et deux phases: Une approche de type «recette» 17

W. YUNG et J.N.K. RAO
Linéarisation des estimateurs de variance jackknife dans un échantillonnage stratifié à degrés multiples 23

Y.P. CHAUBEY, F. NEBEBE et P.S. CHEN
Estimation des caractéristiques des petites régions au moyen d'un modèle gaussien inverse 33

L. RIZZO, G. KALTON et J.M. BRICK
Comparaison de quelques méthodes de correction de la non-réponse d'un panel 43

Y. DING et S.E. FIENBERG
Estimation de la population par échantillonnage multiple et sous-dénombrement lors du recensement en présence d'erreurs d'appariement 55

J.G. SLANTA et T.R. KRENZKE
Utilisation de la méthode de Lavallée et Hidiroglou pour le calcul des limites de stratification aux fins de l'enquête annuelle sur les dépenses en capital du Bureau of the Census 65

E.B. DAGUM
Nouvelle méthode visant à limiter le nombre d'ondulations indésirables et les corrections lors de l'estimation de la tendance-cycle au moyen du modèle X-11-ARMMI 77

Y. TILLÉ
Un algorithme de stratification mobile 85

A.G. de WAAL et L.C.R.J. WILLENBORG
Aperçu du contrôle statistique de la divulgation des microdonnées 95

Techniques d'enquête, juin 1996
Vol. 22, n° 1, pp. 1–2
Statistique Canada

Dans ce numéro

Ce numéro de *Techniques d'enquête* contient des articles traitant de divers sujets. Dans le premier, Steel, Holt et Tranmer examinent le problème de l'utilisation de données agrégées lors d'études portant sur les relations au niveau de l'individu ou du ménage. Ils proposent un modèle général simple qui s'efforce d'intégrer les effets géographiques d'agrégation. Ils décrivent ensuite l'effet de ce modèle sur l'estimation des moyennes et des matrices de covariances de la population, ainsi que sur l'analyse au niveau régional. De plus, en faisant intervenir des variables auxiliaires pour lesquelles certaines sources externes fournissent une estimation de la matrice de covariances au niveau unitaire, les auteurs proposent des méthodes qui donnent une estimation non biaisée des paramètres de niveau individuel, de façon à éviter l'effet d'agrégation géographique.

Binder propose une approche de type «recette» permettant d'obtenir une approximation, par série de Taylor, de la variance d'une grande variété d'estimateurs extraits d'enquêtes complexes. L'auteur présente plusieurs exemples utiles et de nouveaux résultats sur l'application de cette technique générale à l'échantillonnage à deux phases. Il justifie aussi la méthode proposée en montrant qu'elle est cohérente avec la formulation avancée antérieurement par Binder et Patak.

De leur côté, Yung et Rao proposent une approximation linéaire de l'estimateur de variance jackknife. La méthode du jackknife linéarisée garde les propriétés statistiques intéressantes de l'estimateur de variance jackknife habituel, mais nécessite des calculs beaucoup moins laborieux. Le nouvel estimateur de la variance que proposent les auteurs s'applique à un estimateur de régression généralisée d'un total et au ratio de deux estimateurs de régression généralisée. Dans le cadre d'une simulation s'inspirant des données de l'enquête sur la U.S. Current Population Survey, ces auteurs ont constaté que l'estimateur de variance jackknife, l'estimateur de variance jackknife linéarisé et l'estimateur de variance linéarisé habituel donnent de très bons résultats pour l'estimation d'un total obtenu par stratification a posteriori, tandis qu'une forme incorrecte de l'estimateur jackknife entraîne un biais important.

Chaubey, Nebebe et Chen envisagent l'utilisation d'un modèle gaussien inverse pour les données étalées vers la droite et élaborent un estimateur correspondant au moyen d'un modèle pour les totaux de domaines, estimateur qui se compose de variables explicatives pour une régression gaussienne inverse et d'estimateurs d'expansion du biais dû à la régression. Les auteurs proposent aussi une variante de l'estimateur qui attribue un poids réduit aux paramètres servant à corriger le biais. Cet estimateur ressemble à l'estimateur de régression modifié que proposaient Särndal et Hidiroglou. Les estimateurs suggérés fonctionnent raisonnablement bien dans le cadre d'une simulation reposant sur des données synthétiques relatives au revenu qui s'inspire de l'enquête de Statistique Canada sur le revenu, l'équipement ménager et les finances des ménages.

Rizzo, Kalton et Brick étudient l'usage des renseignements complémentaires pour compenser la non-réponse du panel par diverses techniques d'ajustement des poids. Se servant de données tirées du Survey of Income and Program Participation (SIPP) pour illustrer leur méthode de pondération, les auteurs abordent deux aspects importants, à savoir le choix des variables auxiliaires à utiliser lors de l'ajustement des poids servant à compenser la non-réponse et le choix de la technique d'ajustement proprement dite. Les variables auxiliaires appropriées sont choisies par une méthode de sélection combinée à un modèle de régression logistique. Les méthodes d'ajustement de la pondération des non-réponses examinées s'appuient sur des modèles de régression logistique, des algorithmes de recherche catégorique et une itération généralisée. Les auteurs analysent en détail la comparaison empirique des diverses méthodes.

Ding et Fienberg examinent les modèles de l'erreur d'appariement dont on peut se servir pour estimer la population totale à partir de l'appariement probabiliste de deux échantillons ou plus. Leur modèle est principalement destiné à être appliqué à un recensement par échantillonnage multiple, c'est-à-dire à un recensement doublé d'échantillons secondaires. Ils illustrent l'utilité de leurs méthodes en les appliquant à une analyse des données issues de la répétition générale du recensement de 1988 à St. Louis, pour laquelle trois échantillons ont été appariés: celui du recensement proprement dit, l'échantillon de l'enquête postcensitaire et la liste administrative complémentaire.

Dans un article portant sur la stratification optimale, Slanta et Krenzke discutent de l'utilisation de la méthode Lavallée-Hidiroglou. Cette méthode itérative vise à déterminer des bornes de stratification qui minimisent la taille de l'échantillon pour un coefficient de variation donné. Dans un contexte pratique, les auteurs présentent les difficultés rencontrées avec la méthode et montrent comment elles ont été résolues.

Dagum propose une nouvelle méthode pour estimer les tendances sous-jacentes à partir des données désaisonnalisées. Son approche comporte deux étapes. On extrapole d'abord les données désaisonnalisées au moyen d'un modèle ARMMI. Ensuite, on applique un filtre Henderson à 13 termes à la série élargie en se servant de strictes limites sigma pour identifier et remplacer les valeurs extrêmes. L'auteur compare sa méthode à la méthode type au moyen de données issues de plusieurs séries économiques chronologiques et constate ainsi que la nouvelle méthode engendre moins d'irrégularités dans la tendance estimative tout en permettant une identification aussi rapide des changements de sens et en exigeant généralement de plus petites corrections.

Tillé propose un algorithme généralisant la méthode de sélection-rejet permettant de constituer un échantillon aléatoire simple sans remise. Un cas particulier de cet algorithme baptisé «algorithme de stratification mobile» est discuté. Il permet d'obtenir un effet de stratification lissé en utilisant comme variable de stratification le numéro d'ordre des unités d'observation. Cet algorithme permet de contourner le délicat problème du découpage d'une variable continue en strates.

De Waal et Willenborg passent en revue les recherches récentes sur le contrôle statistique de la divulgation dans les fichiers de microdonnées, selon le point de vue du bureau de statistique néerlandais. Les auteurs développent des modèles en fonction de la probabilité qu'on puisse identifier un enregistrement particulier et de la probabilité qu'un enregistrement quelconque dans un fichier de microdonnées puisse être identifié de nouveau. Un nouveau codage global et la suppression locale sont des méthodes qui permettent de réduire les risques de divulgation. Les auteurs parviennent à la conclusion que beaucoup de recherche et de travaux de développement restent à faire sur le plan de la méthodologie avant qu'on parvienne à des logiciels efficaces.

Enfin, c'est le coeur lourd que je dois signaler le récent décès de Maria Gonzalez, emportée par un infarctus alors qu'elle passait ses vacances à Porto Rico, en février dernier. Parmi ses nombreuses contributions à la statistique, ces dernières années, Maria était notamment devenue rédactrice associée de *Techniques d'enquête*. À ce titre, nous avons considérablement apprécié les efforts qu'elle a déployés pour améliorer la qualité et le champ d'intérêt du périodique. Nous la regretterons longtemps. Sa notice nécrologique, rédigée par Elizabeth et Fritz Scheuren, a paru dans le numéro d'avril de *Amstat News*.

Le rédacteur en chef

Techniques d'enquête, juin 1996
Vol. 22, n° 1, pp. 3–15
Statistique Canada

Inférences au niveau unitaire à partir de données agrégées

D.G. STEEL, D. HOLT et M. TRANMER[1]

RÉSUMÉ

Les données ne sont souvent disponibles que sous la forme de moyennes de groupes ou de régions. Or, on sait pertinemment qu'une analyse statistique articulée sur les données de ce genre aboutit fréquemment à des résultats très différents de ceux obtenus lorsqu'on analyse les données correspondantes sur les individus ou les ménages. Croire que les résultats d'une analyse de niveau régional sont applicables au niveau individuel, c'est risquer de commettre l'erreur écologique. Les effets de l'agrégation ou les effets écologiques résultent en partie du fait qu'une région n'est pas constituée d'un assemblage aléatoire d'êtres humains ou de ménages. Les paramètres socioéconomiques varient considérablement d'une région à l'autre. On doit intégrer la structure de la population au modèle statistique utilisé pour l'analyse si on veut bien saisir les conséquences de l'agrégation. Les auteurs proposent un modèle général simple pour y parvenir et décrivent l'effet de ce modèle sur l'estimation des moyennes et des matrices des covariances de la population. Par ailleurs, ils montrent comment obtenir une estimation non biaisée des paramètres au niveau de l'individu à partir des données agrégées, de façon à éviter l'erreur écologique. Les méthodes qu'ils préconisent supposent l'identification des «variables de groupement» qui caractérisent le processus qui a mené à la structure de la population ou, du moins, les différences entre régions. On doit pour cela trouver une estimation de la matrice des covariances pour les variables d'agrégation, au niveau unitaire, d'une source quelconque. L'analyse des données du recensement de 1991 du Royaume-Uni a permis d'identifier les principales variables d'agrégation et de mesurer l'efficacité des méthodes de correction envisagées pour estimer les matrices des covariances et les coefficients de corrélation. Les résultats de ces travaux concourent à l'élaboration d'une stratégie pour l'analyse des données agrégées.

MOTS CLÉS: Agrégation; erreur écologique; groupement; sélection; composantes de la variance.

1. INTRODUCTION

Les chercheurs qui souhaitent étudier les relations au niveau de l'individu éprouvent souvent des difficultés parce qu'ils doivent utiliser des données agrégées, par exemple une moyenne ou un total régional. Idéalement, on devrait se servir des données au niveau de l'unité recueillies lors du sondage ou du recensement, mais ces données sont inaccessibles parce qu'il faut en préserver la confidentialité ou parce qu'elles ne viennent pas d'une enquête ou d'un recensement récents. Les régimes administratifs nous procurent des renseignements sur diverses variables, par exemple le chômage, la santé et la morbidité. Malheureusement, ces données sont fournies habituellement pour des agrégats, notamment par région, toujours pour des raisons de confidentialité. Le recensement fournit lui aussi des données régionales. C'est pourquoi les recherches sociales et épidémiologiques portent encore couramment sur des données collectives.

Soit une population au sein de laquelle chaque sujet est associé à un vecteur de variables intéressantes, dont la répartition a pour moyenne μ_y et pour matrice des covariances Σ_{yy}. Nous aimerions savoir quels liens existent entre ces variables, comme les représentent les corrélations, les coefficients de régression et les grandes composantes qu'on peut tirer de la matrice des covariances Σ_{yy}, sujet principal de nos inférences. Ces variables pourraient comprendre, par exemple, une batterie de tests d'accomplissement dans le cadre d'une étude sur l'éducation, l'incidence d'une maladie donnée et une suite de variables explicatives dans le contexte d'une étude épidémiologique, voire une série de mesures de privation dans l'optique d'une étude sociologique. Nous supposons ici qu'on ne peut obtenir les données au niveau de l'individu. Néanmoins, la région peut être subdivisée en plus petits ensembles comme des districts de recensement (DR) et pour chacun de ces sous-ensembles g ou pour un échantillon de régions, on peut calculer le vecteur des valeurs moyennes $\bar{y}_g$ pour les variables étudiées de l'échantillon n_g d'où sont extraites ces dernières.

Le but de l'analyse est d'obtenir une matrice des covariances Σ_{yy} couvrant ces petites régions. L'inférence n'est pas conditionnelle à l'appartenance à une petite région, mais se rapporte à la distribution marginale entre ces régions. La situation est différente avec l'estimation d'une petite région. Dans ce cas, l'inférence vise la distribution conditionnelle dans la région en question. Pareil objectif est tout à fait légitime, mais diffère de celui qui nous intéresse. Il se peut qu'on puisse recourir aux mêmes modèles, mais l'inférence n'a pas la même visée. Notre formulation nous permet néanmoins d'ajouter les variables spécifiques à certains groupes aux variables intéressantes, s'il y a lieu. Ainsi, si on associe à chaque individu un ensemble de moyennes DR pour la région où il se trouve, il est alors possible de les inclure au vecteur y qui nous intéresse. Pareille approche se prête aux analyses de régression qui utilisent les moyennes des petites régions comme variables explicatives.

[1] D.G. Steel, Department of Applied Statistics, University of Wollongong, NSW 2522, Australia; D. Holt et M. Tranmer, Department of Social Statistics, University of Southampton, S017 1BJ, United Kingdom.

Les premiers travaux sur l'analyse des données agrégées remontent à Gehlke et Biehl (1934). D'éminents chercheurs comme Yule et Kendall (1950), Robinson (1950), Blalock (1964), Openshaw et Taylor (1979) et, plus récemment, Arbia (1989) ont sensiblement contribué à l'étude du problème. Le fait que les unités régionales utilisées présentent rarement une importance particulière (elles sont construites pour des raisons d'économie, de facilité ou de commodité administrative) soulève aussi des difficultés. Par ailleurs, les résultats de l'analyse effectuée au niveau du groupe dépendent de l'échelle des unités, soit de leur taille moyenne et du jeu de limites retenu. Plusieurs études empiriques ont fait ressortir ces aspects, notamment celles de Clark et Avery (1976), de Perle (1977), d'Openshaw (1984) et de Fotheringham et Wong (1991). Les travaux n'ont malheureusement pas débouché sur une théorie permettant une application générale, ni sur des méthodes pratiques en vertu desquelles on pourrait tirer des inférences fiables de niveau unitaire à partir des résultats des analyses effectuées au niveau du groupe.

On attribue les effets de l'agrégation au fait que la population des unités géographiques ne s'est pas constituée au hasard. En règle générale, les individus d'une même région ont tendance à se ressembler davantage, car ils n'ont pas choisi de vivre là au hasard ou parce qu'ils s'y trouvent consécutivement à l'action de forces analogues, voire parce qu'ils ont des échanges mutuels. Il existe donc des différences socioéconomiques entre les régions, différences qui se confondent avec les effets individuels dans l'analyse statistique des données agrégées sur les régions concernées. Nous proposons d'abord un modèle général simple qui s'efforce d'intégrer ces effets, puis nous examinons les conséquences de son utilisation et les implications d'une telle méthode pour l'analyse au niveau régional. Par ailleurs, nous suggérons des méthodes qui, dans certaines circonstances, donneront une estimation non biaisée des paramètres de niveau individuel à partir des données agrégées, donc permettront d'éviter l'erreur écologique. Ces méthodes font intervenir des variables auxiliaires pour lesquelles certaines sources donnent une matrice des covariances de niveau unitaire pour l'échantillon. Cette approche a été appliquée aux données du recensement de 1991 du Royaume-Uni et une stratégie a été développée en vue de l'analyse des données agrégées.

2. MODÈLES POUR LES EFFETS RÉGIONAUX

Soit une population de N sujets ayant chacun un vecteur y de caractéristiques à étudier. La population se compose de M groupes et la variable aléatoire c_i indique la région à laquelle la i-ème unité de population appartient. On dénombre N_g individus dans la g-ième région.

Supposons que μ_y et Σ_{yy} soient les paramètres de la superpopulation. Compte tenu de ces prémisses, on obtient la théorie statistique qui suit. Quelques aspects relatifs au plan d'échantillonnage seront toutefois examinés à la fin de la partie 2.

Nous partons de l'hypothèse qu'il existe un ensemble de données d'échantillon s de taille n et que ces données individuelles ont été agrégées pour former un ensemble de m moyennes régionales utilisables pour l'analyse. Il est donc possible de calculer les statistiques suivantes au niveau régional:

moyenne de la g-ième région:

$$\bar{y}_g = \frac{1}{n_g} \sum_{i \in g,s} y_i \tag{2.1}$$

moyenne totale de l'échantillon:

$$\bar{y} = \frac{1}{n} \sum_{g \in s} n_g \bar{y}_g = \frac{1}{n} \sum_{i \in s} y_i \tag{2.2}$$

matrice des covariances de l'échantillon au niveau régional:

$$\bar{S}_{yy} = \frac{1}{m-1} \sum_{g \in s} n_g (\bar{y}_g - \bar{y})(\bar{y}_g - \bar{y})'. \tag{2.3}$$

On peut définir des statistiques analogues au niveau unitaire, mais l'analyste ne pourra s'en servir. Par exemple, $S_{yy} = 1/(n-1) \sum_{i \in s} (y_i - \bar{y})(y_i - \bar{y})'$ est la matrice des covariances de l'échantillon au niveau unitaire.

2.1 Groupement aléatoire

Bien que les groupes géographiques soient rarement le fait du hasard, pareille hypothèse constitue un bon point de départ quand on s'intéresse à l'analyse écologique. Si les groupes se forment de façon aléatoire, maintes analyses effectuées au niveau du groupe seront valables, même si leur utilité est réduite. Steel et Holt (1995) étudient les propriétés de diverses statistiques comme la moyenne, la variance et les coefficients de régression et de corrélation dans une situation de ce genre. Quand les groupes se constituent au hasard, c'est-à-dire $y \perp c$, alors

$$E[\bar{y}_g \mid s,c] = \mu_y \tag{2.4}$$

$$V(\bar{y}_g \mid s,c) = \frac{1}{n_g} \Sigma_{yy}. \tag{2.5}$$

Les propriétés fondamentales des statistiques au niveau unitaire et au niveau du groupe sont faciles à déduire:

$$\mathrm{Cov}(\bar{y}_g, \bar{y}_h \mid s,c) = \mathbf{0} \quad g \neq h \tag{2.6}$$

$$E[\bar{y} \mid s,c] = \mu_y \tag{2.7}$$

$$E[S_{yy} \mid s,c] = \Sigma_{yy} \tag{2.8}$$

$$E[\bar{S}_{yy} \mid s,c] = \Sigma_{yy}. \tag{2.9}$$

Ces propriétés s'appliquent si on peut négliger l'échantillonnage étant donné les paramètres discriminants du groupe, bref le plan d'échantillonnage peut dépendre des groupes mais pas de y ou d'une variable quelconque associée à y, sous réserve de c. Par exemple, on pourrait utiliser le recensement ou un échantillon aléatoire simple de groupes et d'unités appartenant à ces groupes.

On peut recourir à des statistiques non pondérées au niveau du groupe en posant $n_g = 1$ dans les équations (2.2) et (2.3). On obtient ainsi des estimateurs inefficaces. Le degré d'inefficacité dépendra de la distribution des groupes d'échantillon des groupes. La pondération selon les tailles d'échantillon des groupes est importante. Cela fait, on peut procéder aux inférences habituelles en apportant les corrections appropriées aux degrés de liberté. La variabilité dépend du nombre de régions plutôt que du nombre d'observations et on adapte les intervalles de confiance et les épreuves en conséquence.

2.2 Modèle des composantes de la variance

Une façon simple d'illustrer la corrélation positive entre les membres d'un groupe normalement observée dans une population consiste à utiliser un modèle des composantes de la variance, ce qui, dans le cas d'une analyse à plusieurs variables, correspond à

$$y_i = \mu_y + v_g + \epsilon_i \qquad i \in g$$

où v_g et ϵ_i sont des composantes aléatoires indépendantes au niveau du groupe et au niveau de l'individu, respectivement, avec une espérance mathématique nulle, $V(\epsilon_i \mid c) = \Sigma_{\epsilon\epsilon}$ et $V(v_g \mid c) = \Delta_{yy}$.

Modèle A:

$$E[y_i \mid c] = \mu_y \tag{2.10}$$

$$V(y_i \mid c) = \Sigma_{\epsilon\epsilon} + \Delta_{yy} = \Sigma_{yy} \tag{2.11}$$

$$\begin{aligned} \mathrm{Cov}(y_i, y_j \mid c) &= \Delta_{yy} \quad \text{si} \quad c_i = c_j \quad i \neq j \\ &= 0 \text{ sinon.} \end{aligned} \tag{2.12}$$

La notation $V(\cdot \mid c)$ signifie que la matrice des covariances dépend de l'étiquette c du groupe, donc détermine l'appartenance à un groupe commun. On estime cependant qu'elle est inconditionnelle pour les effets aléatoires au niveau du groupe. Par conséquent, $V(y_i \mid c)$ inclut la variance totale, à la fois pour la matrice des covariances de groupe $\Sigma_{\epsilon\epsilon}$ et la matrice des covariances de groupe Δ_{yy}.

On obtient facilement les propriétés des moyennes de l'échantillon au niveau du groupe à partir du modèle A si on peut négliger l'échantillonnage étant donné c,

$$E[\bar{y}_g \mid s,c] = \mu_y \tag{2.13}$$

$$V(\bar{y}_g \mid s,c) = \frac{1}{n_g}(\Sigma_{yy} + (n_g - 1)\Delta_{yy}) \tag{2.14}$$

$$\mathrm{Cov}(\bar{y}_g, \bar{y}_h \mid s,c) = 0 \quad g \neq h. \tag{2.15}$$

Les propriétés des statistiques au niveau unitaire et au niveau du groupe sont les suivantes:

$$E[\bar{y} \mid s,c] = \mu_y \tag{2.16}$$

$$E[S_{yy} \mid s,c] = \Sigma_{yy} - \frac{\bar{n}^0 - 1}{n - 1}\Delta_{yy} \tag{2.17}$$

$$E[\bar{S}_{yy} \mid s,c] = \Sigma_{yy} + (\bar{n}^* - 1)\Delta_{yy} \tag{2.18}$$

où $\bar{n} = n/m$, $\bar{n}^0 = 1/n \sum_{g\in s} n_g^2 = \bar{n}(1 + C_n^2)$, $\bar{n}^* = \bar{n}(1 - C_n^2/(m-1))$ et $C_n^2 = 1/m \sum_{g\in s}(n_g - \bar{n})^2/\bar{n}^2$ est le carré du coefficient de variation des tailles de groupe de l'échantillon. Notons que le coefficient de Δ_{yy} est $0(m^{-1})$ dans (2.17) et $0(\bar{n})$ dans (2.18), ce qui montre bien comment une petite erreur d'analyse au niveau unitaire prend de l'ampleur après agrégation. Nous approfondirons ces résultats à la partie 2.4.

2.3 Modèles de groupement

Les chercheurs qui s'intéressent à l'analyse écologique ont échafaudé des modèles qui tiennent compte du mécanisme de création des groupes. Ils supposent qu'un processus de synthèse attribue les unités à tel ou tel groupe, selon un vecteur de variables de groupement z_i, de façon stochastique ou déterministe. Blalock (1964) recourt implicitement à cette approche dans son analyse, tandis qu'Hannan et Burstein (1974), Litchman (1974), Langbein et Litchman (1978), Smith (1977) et Blalock (1979, 1985) s'en servent de façon explicite. Steel (1985) parle de modèles de groupement, car il suppose que les groupes naissent au terme d'un processus quelconque qui fait intervenir les variables au niveau des relations à l'étude. La formation du groupe est perçue comme une distorsion, aussi les relations intéressantes sont-elles définies auparavant. On précise souvent dans la discussion sur les modèles contextuels que les effets contextuels apparents peuvent résulter de tels facteurs. La version multivariée du modèle est la suivante:

Modèle B:

$$E[y_i \mid z,c] = \mu_{y.z} + \beta_{yz}' z_i \tag{2.19}$$

$$V(y_i \mid z,c) = \Sigma_{yy.z} \tag{2.20}$$

$$\mathrm{Cov}(y_i, y_j \mid z,c) = 0 \quad i \neq j. \tag{2.21}$$

Dans ce modèle, l'espérance conditionnelle de y_i dépend seulement de la valeur des variables auxiliaires de la i-ième unité et est indépendante du groupe à laquelle l'unité appartient ou de la valeur des variables auxiliaires des autres unités de la population. La covariance conditionnelle entre deux unités quelconques est zéro. Ce modèle est valable pour les modèles de groupement dans lesquels le mécanisme de

formation du groupe se caractérise par les variables auxiliaires z_i. Les variables auxiliaires peuvent être considérées comme des variables déterminant à quel groupe appartient telle ou telle unité. Sous un angle plus général, les variables auxiliaires correspondent aux principales variables de niveau individuel qui ne sont pas distribuées au hasard ni entre les groupes à cause des processus de sélection ou de migration auxquels la population est assujettie. On peut aussi inclure au modèle les variables contextuelles sous la forme de variables auxiliaires de valeur identique pour toutes les unités du groupe.

Si le vecteur des variables auxiliaires a une distribution marginale dont la moyenne est μ_z et la matrice des covariances, Σ_{zz} la moyenne marginale et la matrice des covariances de y seront respectivement $\mu_y = \mu_{y.z} + \beta'_{yz}\mu_z$ et $\Sigma_{yy} = \Sigma_{yy.z} + \beta'_{yz}\Sigma_{zz}\beta_{yz}$. Les propriétés des moyennes de l'échantillon au niveau du groupe sont faciles à tirer du modèle B:

$$E[\bar{y}_g \mid s,z,c] = \mu_y + \beta'_{yz}(\bar{z}_g - \mu_z) \tag{2.22}$$

$$V(\bar{y}_g \mid s,z,c) = \frac{1}{n_g}\Sigma_{yy.z} \tag{2.23}$$

$$\mathrm{Cov}(\bar{y}_g,\bar{y}_h \mid s,z,c) = 0 \qquad g \neq h. \tag{2.24}$$

Les statistiques au niveau du groupe auront donc les propriétés qui suivent:

$$E[\bar{y} \mid s,z,c] = \mu_y + \beta'_{yz}(\bar{z} - \mu_z) \tag{2.25}$$

$$E[S_{yy} \mid s,z,c] = \Sigma_{yy} + \beta'_{yz}(S_{zz} - \Sigma_{zz})\beta_{yz} \tag{2.26}$$

$$E[\bar{S}_{yy} \mid s,z,c] = \Sigma_{yy} + \beta'_{yz}(\bar{S}_{zz} - \Sigma_{zz})\beta_{yz} \tag{2.27}$$

où S_{zz} et $\bar{S}_{zz}$ sont définis de façon analogue à S_{yy} et $\bar{S}_{yy}$ de l'équation (2.3) et de la phrase suivant cette dernière.

2.4 Modèle combiné

Jusqu'à présent, on peut dire que les deux modèles examinés expliquent les effets de groupe chacun de leur côté. Il est possible de les combiner en un modèle plus réaliste qui intégrera les deux effets de groupe et les composantes résiduelles de la variance:

Modèle C:

$$E[y_i \mid z,c] = \mu_{y.z} + \beta'_{yz}z_i \tag{2.28}$$

$$V(y_i \mid z,c) = \Sigma_{yy.z} \tag{2.29}$$

$$\begin{aligned}\mathrm{Cov}(y_i,y_j \mid z,c) &= \Delta_{yy.z} \quad \text{si} \quad c_i = c_j \quad i \neq j \\ &= 0 \quad \text{sinon.}\end{aligned} \tag{2.30}$$

Le nouveau modèle accepte les mécanismes de création des groupes que caractérisent les variables auxiliaires z_i. On y retrouve aussi les corrélations résiduelles à l'intérieur d'un groupe qui résultent des effets aléatoires attribués aux variables aléatoires inconnues du niveau du groupe, après exclusion des variables de groupement.

Voici les propriétés des moyennes de l'échantillon au niveau du groupe provenant du modèle C, lorsqu'on néglige l'échantillonnage, étant donné (z,c),

$$E[\bar{y}_g \mid s,z,c] = \mu_y + \beta'_{yz}(\bar{z}_g - \mu_z) \tag{2.31}$$

et

$$V(\bar{y}_g \mid s,z,c) = \frac{1}{n_g}(\Sigma_{yy.z} + (n_g - 1)\Delta_{yy.z}) \tag{2.32}$$

$$\mathrm{Cov}(\bar{y}_g,\bar{y}_h \mid s,z,c) = 0 \qquad g \neq h \tag{2.33}$$

$$E[\bar{y} \mid s,z,c] = \mu_y + \beta'_{yz}(\bar{z} - \mu_z) \tag{2.34}$$

$$E[S_{yy} \mid s,z,c] = \Sigma_{yy} + \beta'_{yz}(S_{zz} - \Sigma_{zz})\beta_{yz} - \frac{\bar{n}^0 - 1}{n - 1}\Delta_{yy.z} \tag{2.35}$$

$$E[\bar{S}_{yy} \mid s,z,c] = \Sigma_{yy} + \beta'_{yz}(\bar{S}_{zz} - \Sigma_{zz})\beta_{yz} + (\bar{n}^* - 1)\Delta_{yy.z}. \tag{2.36}$$

Les équations (2.17) et (2.18) montrent comment l'agrégation amplifie les effets aléatoires au niveau du groupe dans le modèle des composantes de la variance A. Dans l'équation (2.17), le coefficient de Δ_{yy} est $0(m^{-1})$, tandis que dans (2.18), il correspond à $0(\bar{n})$. Avec le modèle combiné C, les équations (2.35) et (2.36) montrent comment l'inclusion des variables de groupement permet de scinder le biais en deux éléments additifs: le premier associé aux variables de groupement, aux liens de ces dernières avec les variables d'intérêt et à leur effet cumulatif, et le second faisant intervenir $\Delta_{yy.z}$, soit les composantes résiduelles de la variance, après prise en compte des variables de groupement. Notons que les coefficients de $\Delta_{yy.z}$ dans les équations (2.35) et (2.36) restent $0(m^{-1})$ et $0(\bar{n})$ respectivement, comme aux équations (2.17) et (2.18). Les composantes résiduelles de la variance devraient néanmoins être plus faibles, en général. L'équation (2.29) part de l'hypothèse que la variance résiduelle de c est constante.

L'hypothèse que l'échantillonnage est négligeable pour (z,c) signifie que le plan d'échantillonnage peut dépendre des variables auxiliaires et des paramètres discriminants du groupe. On peut donc effectuer une stratification selon z et procéder à un échantillonnage en grappes ou à plusieurs degrés en fonction des groupes.

La matrice $\bar{S}_{yy}$ pondérée au niveau du groupe permet d'estimer Σ_{yy}. Le premier terme de biais de (2.36) vient de l'effet des variables de groupement. Il est égal à zéro si $\beta_{yz} = \mathbf{0}$, ou presque égal à zéro si $\bar{S}_{zz} \doteq \Sigma_{zz}$. La condition $\beta_{yz} = \mathbf{0}$ est ferme et signifie que les variables intéressantes n'ont aucun lien avec les variables de groupement. L'effet de l'agrégation sur la covariance de l'échantillon pour deux variables quelconques dépend des liens des deux variables en question avec les variables du groupement z_i. On devrait s'attendre à ce que les effets cumulatifs soient plus importants quand elles sont plus étroitement liées aux variables de groupement. En raison de la condition $\bar{S}_{zz} \doteq \Sigma_{zz}$ les variables z échappent aux effets de la sélection et de l'agrégation. Néanmoins, il est peu probable que ces conditions s'appliquent dans la réalité, de sorte que beaucoup de variables auront un biais. L'erreur d'échantillonnage et de groupement attribuable aux variables auxiliaires est mesurée par $S_{zz} - \Sigma_{zz}$ pour l'estimateur de niveau unitaire, et par $\bar{S}_{zz} - \Sigma_{zz}$ pour l'estimateur au niveau du groupe. L'expression $\bar{S}_{zz} - \Sigma_{zz}$ traduit l'effet net de l'échantillonnage et de l'agrégation sur les variables auxiliaires.

Le deuxième terme de biais (2.36) est nul si $\Delta_{yy.z} = \mathbf{0}$. Il n'existe donc aucune corrélation résiduelle à l'intérieur du groupe pour les variables y, sous réserve des variables de groupement. La chose est peu probable en pratique, mais il est souhaitable d'identifier les variables de groupement qui intègrent le plus possible les effets cumulatifs en rendant le terme résiduel le plus faible possible.

Les effets du groupement et de l'échantillonnage selon z et l'effet attribuable à la corrélation résiduelle à l'intérieur du groupe s'additionnent; c'est ce qui se produit avec les types de corrélations intérieures au groupe plus complexes, pourvu que la linéarité du modèle soit préservée. Si z suit un simple modèle des composantes de la variable comme le modèle A,

$$E[\bar{S}_{zz} \mid s,c] = \Sigma_{zz} + (\bar{n}^* - 1)\Delta_{zz}$$

$$E[\bar{S}_{yy} \mid s,c] = \Sigma_{yy} + (\bar{n}^* - 1)\beta'_{yz}\,\Delta_{zz}\,\beta_{yz} + \Delta_{yy.z} \qquad (2.37)$$

et les covariances des variables intéressantes à l'intérieur du groupe seront constituées de l'élément résultant des covariances des variables auxiliaires à l'intérieur du groupe et des composantes résiduelles. Le côté droit de l'équation (2.37) est tiré de (2.18) puisque z suit sans conditions le modèle des composantes de la variance, si y en fait autant. Le modèle fondamental a pour but de trouver les variables auxiliaires qui permettront de réduire au maximum les covariances résiduelles ou conditionnelles $\Delta_{yy.z}$ à l'intérieur d'un groupe, ou de les éliminer, dans le meilleur des cas.

2.5 Correction des effets cumulatifs

Peu de propositions utiles ont été avancées sur la manière dont les analyses de niveau régional pourraient être corrigées afin de donner une estimation raisonnable des relations au niveau unitaire. Duncan et Davis (1953) ont examiné la fourchette éventuelle des coefficients de corrélation au moyen d'un tableau carré (2 × 2) dont les marges étaient connues. Les limites résultantes sont souvent trop importantes pour présenter une utilité pratique. De son côté, Goodman (1959) a énoncé les conditions précises dans lesquelles l'analyse écologique pourrait servir à tirer des inférences sur les relations de niveau individuel avec un modèle de régression. Langbein et Litchman (1978) ont examiné quelques méthodes applicables aux groupements articulés sur les variables dépendantes, quand on connaît les variances de niveau unitaire des variables dépendantes et indépendantes du modèle de régression. Malheureusement, aucune de ces approches n'est assez générale pour permettre la résolution du problème.

Si on examine le biais pour $\bar{S}_{yy}$ dans (2.36), on constate qu'en ajoutant $\beta'_{yz}(\Sigma_{zz} - \bar{S}_{zz})\beta_{yz}$ à $\bar{S}_{yy}$, le terme de biais attribuable aux variables de groupement disparaît. L'équation (2.31) implique que

$$E[\bar{B}_{yz} \mid s,z,c] = \beta_{yz} \qquad (2.38)$$

où $\bar{B}_{yz} = \bar{S}_{zz}^{-1}\,\bar{S}_{zy}$.

Si on disposait de la matrice des covariances S_{zzs_0} de z, pour l'échantillon unitaire s_0 tiré de m_0 groupes, l'estimateur corrigé

$$\hat{\Sigma}_{yy}(z) = \bar{S}_{yy} + \bar{B}'_{yz}(S_{zzs_0} - \bar{S}_{zz})\bar{B}_{yz} \qquad (2.39)$$

éliminerait le biais d'agrégation résultant des variables de groupement z, pourvu que S_{zzs_0} s'approche de Σ_{zz}. Il se peut que la source de S_{zzs_0} soit largement indépendante des données utilisées dans $\bar{S}_{yy}$ et $\bar{B}_{yz}$. Steel (1985) montre que l'estimateur corrigé (2.39) peut correspondre à l'estimateur du maximum de vraisemblance de Σ_{yy} (après substitution de $m - 1$ par m, *etc.*, comme on le fait d'habitude). Si la normalité de la répartition de (y,z) s'applique, s_0 devient un simple échantillon aléatoire de la population et $\Delta_{yy.z} = \mathbf{0}$. L'estimateur corrigé correspond à l'ajustement de Pearson (1903) que Holt, Smith et Winter (1980) envisagent dans leur analyse de régression et que Smith et Holmes (1989) utilisent dans leur analyse à plusieurs variables. Dans les deux cas, on corrige les statistiques obtenues à partir des données de niveau unitaire issues d'un échantillon prélevé en fonction des variables auxiliaires. En ce qui nous concerne, la correction est appliquée aux statistiques tirées des moyennes des régions, et les variables auxiliaires utilisées pour effectuer la correction comprennent les variables de groupement ainsi que les variables du plan d'échantillonnage. L'estimateur corrigé de μ_y est

$$\hat{\mu}_y(z) = \bar{y} + \bar{B}'_{yz}(\bar{z}_{s_0} - \bar{z}) \qquad (2.40)$$

où $\bar{z}_{s_0}$ est la moyenne de s_0.

D'après (2.34) et (2.38), on constate que

$$E[\hat{\mu}_y(z) \mid s,z,s_0,c] = \mu_y + \beta'_{yz}(\bar{z}_{s_0} - \mu_z). \qquad (2.41)$$

Par ailleurs, selon Steel (1985), (2.36) et (2.38) impliquent que

$$E[\hat{\Sigma}_{yy}(z) \mid s,z,s_0,c] = \Sigma_{yy} + \beta'_{yz}(S_{zzs_0} - \Sigma_{zz})\beta_{yz} + (\bar{n}^* - 1)\Delta_{yy.z} + 0(m^{-1}) \quad (2.42)$$

pourvu que $\text{tr}(\bar{S}_{zz}^{-1} S_{zzs_0})$ et $\bar{n}\,\text{tr}\big((\bar{S}_{zz}^{-1} S_{zzs_0} - I)\bar{S}_{zz}^{-1} \bar{S}_{zz}^{(2)}\big)$ soient bornés, $\bar{S}_{zz}^{(2)}$ étant défini de la même façon que $\bar{S}_{zz}$, sauf que n_g est remplacé par $n_g^2/\bar{n}$.

Quand on compare (2.42) à (2.35), on se rend compte que la composante du biais attribuable aux variables de groupement a été ramenée au biais lié à l'utilisation de S_{yys_0} si elle est disponible. L'estimateur corrige les effets cumulatifs résultant de z et ramène l'effet du plan d'échantillonnage associé à s à celui lié à s_0.

Supposons que le plan d'échantillonnage utilisé pour produire s_0 et les valeurs des variables auxiliaires viennent d'une superpopulation, de telle sorte que

$$E[\bar{z}_{s_0} \mid s_0,c] = \mu_z + 0(m_0^{-1}) \quad (2.43)$$

$$E[S_{zzs_0} \mid s_0,c] = \Sigma_{zz} + 0(m_0^{-1}) \quad (2.44)$$

où m_0 représente le nombre de groupes dans s_0.

Alors,

$$E[\hat{\mu}_y(z) \mid s,s_0,c] = \mu_y + 0(m_0^{-1}) \quad (2.45)$$

$$E[\hat{\Sigma}_{yy}(z) \mid s,s_0,c] = \Sigma_{yy} + (\bar{n}^* - 1)\Delta_{yy.z} + 0(\tilde{m}^{-1}) \quad (2.46)$$

où

$$\tilde{m} = \min(m,m_0).$$

Les conditions (2.43) et (2.44) s'appliquent si les valeurs de groupe de la population z viennent d'un modèle dont les composantes de la variance sont similaires à celles du modèle A et si le plan d'échantillonnage de s_0 ne dépend que du groupement, pas des variables auxiliaires. L'échantillonnage aléatoire simple, l'échantillonnage par grappes à probabilité de sélection identique, voire l'échantillonnage à plusieurs degrés respectent cette condition. On pourrait aussi se servir des données du recensement pour que s_0 constitue la population finie dans son ensemble.

Le biais attribuable aux variables de groupement peut donc être corrigé, pourvu qu'on dispose d'une matrice des covariances quelconque pour z au niveau unitaire. La raison pour laquelle on recourt à une telle approche est de créer une situation où les effets prédominants du groupe seraient attribués à la sélectivité ou au groupement, par le truchement des variables de groupement. La correction relative aux variables auxiliaires supprime l'effet de la corrélation apparente à l'intérieur du groupe associée à ces variables. Même corrigé cependant, l'estimateur inclut un élément de biais en raison de $\Delta_{yy.z}$ et si z n'atténue pas assez les corrélations à l'intérieur du groupe, le biais demeure appréciable. Cette approche dépend donc du choix des variables auxiliaires qui réduiront les corrélations à l'intérieur du groupe.

Si le plan d'échantillonnage de s_0 et le modèle de superpopulation de z interdisent l'application de (2.43) et de (2.44), on peut remplacer $\bar{z}_{s_0}$ et S_{zzs_0} par les estimateurs $\hat{\mu}_{zs_0}$ et $\hat{\Sigma}_{zzs_0}$ pour obtenir des estimateurs corrigés $\hat{\mu}_y(z)$ et $\hat{\Sigma}_{yy}(z)$. L'espérance mathématique des estimateurs corrigés est donnée par (2.41) et (2.42), où il suffit de remplacer $\bar{z}_{s_0}$ par $\hat{\mu}_{zs_0}$ et S_{zzs_0} par $\hat{\Sigma}_{zzs_0}$. Plusieurs possibilités peuvent être retenues pour $\hat{\mu}_{zs_0}$ et $\hat{\Sigma}_{zzs_0}$ à partir de l'échantillon s_0. Smith et Holmes (1989) ont examiné toute une gamme d'estimateurs envisageables, en fonction du modèle et du plan d'échantillonnage. Supposons, par exemple, que le plan d'échantillonnage utilisé pour obtenir s_0 comporte une stratification d'après les valeurs du vecteur des variables de taille $\boldsymbol{x}$. Appelons la probabilité d'inclusion de l'unité de population i dans l'échantillon, Π_i. Le poids associé à cette probabilité serait $w_i = (\Pi_i)^{-1}$. L'estimateur de μ_z pondéré pour la probabilité est $\bar{z}_{s_0}^* = \sum_{i\in s_0} w_i z_i$, et celui de Σ_{zz} est $S_{zzs_0} = \sum_{i\in s_0} w_i z_i z_i' - w_0^{-1} \bar{z}_{s_0}^* \bar{z}_{s_0}^{*\prime}$ où $w_0 = \sum_{i\in s_0} w_i$.

Les estimateurs corrigés de Pearson pour μ_z et Σ_{zz} sont $\bar{z}_{s_0} + \boldsymbol{B}'_{zxs_0}(\bar{x}_u - \bar{x}_{s_0})$ et $S_{zzs_0} + \boldsymbol{B}'_{zxs_0}(S_{xxu} - S_{xxs_0})\boldsymbol{B}_{zxs_0}$ respectivement. Dans ce cas, $\bar{x}_u$ et S_{xxu} correspondent au vecteur des moyennes et à la matrice des covariances des variables du plan d'échantillonnage de $\boldsymbol{x}$ pour la population finie, et $\boldsymbol{B}_{zxs_0} = S_{xxs_0}^{-1} S_{xzs_0}$.

On pourrait aussi utiliser les estimateurs de Pearson corrigés et pondérés pour la probabilité, à savoir $\bar{z}_{s_0}^* + \boldsymbol{B}'_{zxs_0}(\bar{x}_u - \bar{x}_{s_0}^*)$ et $S_{zzs_0}^* + \boldsymbol{B}'_{zxs_0}(S_{xxu} - S_{xxs_0}^*)\boldsymbol{B}_{zxs_0}^*$.

Ici $\bar{x}_{s_0}^*$ et $S_{xxs_0}^*$ correspondent respectivement à $\bar{z}_{s_0}^*$ et $S_{zzs_0}^*$ et où $\boldsymbol{B}_{zxs_0}^* = S_{xxs_0}^{*-1} S_{xzs_0}^*$. Puisque, jusqu'à présent, l'approche repose essentiellement sur l'utilisation d'un modèle, il serait préférable de prendre les estimateurs μ_y et Σ_{zz}. En pratique néanmoins, les données dont on pourrait se servir pour la correction comprendraient les estimateurs connus de la moyenne et de la covariance, pondérés pour la probabilité et tirés de l'échantillon s_0, qui est indépendant de s. Par conséquent,

$$E_{p_0}[\hat{\mu}_{zs_0} \mid z,c] = \bar{z}_u$$

$$E_{p_0}[\hat{\Sigma}_{zzs_0} \mid z,c] = S_{zzu}$$

où $\bar{z}_u$ et S_{zzu} correspondent au vecteur des moyennes et à la matrice des covariances des variables auxiliaires de la population finie, et où E_{p_0} représente l'espérance mathématique d'un échantillon répété avec le plan d'échantillonnage utilisé pour obtenir s_0, c'est-à-dire la distribution de randomisation. De (2.41) et (2.42), on peut déduire que

$$E[\hat{\mu}_y(z) \mid s,z,c] = \mu_y + \beta_{yz'}(\bar{z}_u - \mu_z)$$

$$E[\hat{\Sigma}_{yy}(z) \mid s,z,c] = \Sigma_{yy} + \beta_{yz'}(S_{zzu} - \Sigma_{zz})\beta_{yz} + (\bar{n}^* - 1)\Delta_{yy.z} + 0(m^{-1}).$$

Ces espérances mathématiques sont reprises dans le modèle statistique qui génère les valeurs y et la répartition par randomisation associée à s_0. En réalité $\bar{z}_u$ et S_{zzu} se rapprochent beaucoup de μ_z et Σ_{zz} respectivement.

3. IDENTIFICATION DES VARIABLES DE GROUPEMENT

Dans la partie qui précède, nous avons présenté un ensemble de variables auxiliaires z caractérisant les variations régionales. Il a servi à corriger l'analyse agrégée et à réduire le biais attribuable à l'agrégation. Avec des variables auxiliaires idéales, $\Delta_{yy.z}$ serait égal à zéro et la méthode de correction supprimerait totalement le biais dû à l'agrégation. En pratique cependant, on ignore quelles variables auxiliaires permettraient d'obtenir $\Delta_{yy.z} = \mathbf{0}$. C'est pourquoi nous devons nous en tenir à des variables pour lesquelles on connaît la moyenne régionale, compte tenu des données analysées, et pour lesquelles il est possible d'estimer la matrice des covariances au niveau unitaire $\hat{\Sigma}_{zz}$. On pourra recourir notamment aux données démographiques et aux variables sur l'habitation de base que fournit couramment le recensement. Toutefois, il se peut que ces variables ne caractérisent pas exactement le processus de groupement, donc n'expliquent pas les différences entre régions aussi bien qu'on le souhaiterait.

3.1 Stratégie d'analyse

Dans la pratique, les variables de groupement ne sont pas connues. Il faut donc élaborer une stratégie en vue d'identifier les variables d'ajustement pour lesquelles on possède une estimation de la matrice des covariances au niveau unitaire, et qui expliqueront les effets de groupement. Voici une façon d'y arriver:

1) Identifier un ensemble de variables couvrant le même domaine que les variables auxquelles on s'intéresse, mais pour lesquelles il existe des données de niveau régional et de niveau unitaire pour une période quelconque du passé (un recensement antérieur, par exemple).

2) Ajouter aux variables qui précèdent d'autres variables (variables démographiques et variables sur l'habitation, par exemple) susceptibles de remplacer les variables z qui présentent une étroite association avec les variations régionales. On aura aussi besoin d'une estimation des matrices des covariances de niveau régional et de niveau unitaire pour la même période.

3) Analyser les données pour déterminer les variables qui expliquent le mieux les effets de niveau régional sur les variables d'intérêt. Nous reviendrons plus tard à cette analyse, baptisée analyse des VGC.

4) De (3), trouver la série de variables de correction qu'on pourrait extraire de l'ensemble de données courant et pour lesquelles on pourrait obtenir la matrice des covariances courante de niveau unitaire d'une source quelconque.

5) Il est possible qu'on découvre une estimation de la variance ou de la covariance de niveau unitaire pour certaines variables qui nous intéressent dans les tables existantes, par exemple. Ensuite, il faut encore calculer les effets de l'agrégation $\bar{Q}_{aa} = \bar{s}_{aa}/s_{aa}$ ou $\bar{Q}_{ab} = \bar{s}_{ab}/s_{ab}$.

6) Utiliser les variables identifiées en (4) pour corriger l'analyse agrégée des variables d'intérêt et vérifier les effets de l'agrégation corrigés correspondant à (5), après correction, afin de s'assurer que l'ajustement a bien donné les résultats escomptés.

3.2 Variables de groupement idéales

Nous examinerons d'abord l'ensemble idéal de variables de groupement utilisables pour la correction de façon à déterminer quelle analyse (VGC) convient le mieux aux données agrégées, conformément à la démarche décrite plus haut.

Supposons que pour l'ensemble complet des variables qui nous intéressent, on connaisse la matrice des variances-covariances de niveau régional $\bar{S}_{yy}$ et la matrice des variances-covariances de niveau unitaire S_{yys_1} pour l'échantillon s_1. Bien sûr, si la chose était faisable dans la réalité, le problème d'agrégation ne se poserait pas puisqu'on pourrait écarter $\bar{S}_{yy}$ et simplement prendre S_{yys_1} comme estimation de Σ_{yy}. Néanmoins, il y a trois bonnes raisons pour étudier pareille situation. Tout d'abord, il vaut la peine d'éclaircir la structure du groupement qui établit la relation entre $\bar{S}_{yy}$ et S_{yys_1}. En second lieu, il se pourrait que $\bar{S}_{yy}$ et S_{yys_1} soient connus pour un moment quelconque dans le temps, le jour du recensement par exemple, mais que l'analyse approfondie d'une version plus récente de $\bar{S}_{yy}$ doive reposer sur des données intercensitaires pour lesquelles on ignore S_{yys_1}. Si la structure du groupement persiste dans le temps, comme il est raisonnable de le penser, l'analyse de $\bar{S}_{yy}$ et de S_{yys_1} au jour du recensement pourrait faciliter l'analyse intercensistaire en permettant l'identification des variables clés qui expliquent la majeure partie des effets de l'agrégation. Ces possibilités sont à la base même de la stratégie décrite à la partie 3.1. Troisièmement, si les variables de y couvrent de nombreuses variables socio-économiques et démographiques, comme cela se produit avec le recensement, les variables clés à l'origine des effets de groupement des variables à l'étude pourraient aussi expliquer une bonne partie des effets de groupement des autres variables socio-économiques et démographiques. Soulignons que les deux échantillons s et s_1 peuvent être identiques, sans que cela soit une condition. Ainsi, s pourrait venir d'une source administrative, en fait une sorte de recensement fournissant des données agrégées sur les régions, tandis que s_1 pourrait comprendre les données individuelles d'un sondage sans marqueur géographique.

Pour faciliter l'identification des variables importantes associées au groupement, Steel (1985) suggère de calculer les valeurs propres $\hat{\theta}_1, \ldots, \hat{\theta}_p$ de $S_{yys_1}^{-1}\bar{S}_{yy}$ et la matrice $\hat{D}_y = [\hat{d}_1, \ldots, \hat{d}_p]$ de telle sorte que

$$\hat{D}_y' \bar{S}_{yy} \hat{D}_y = \text{diag}(\hat{\theta}_k) \quad \text{et} \quad \hat{D}_y' S_{yys_1} \hat{D}_y = I.$$

Les variables définies par la transformation

$$\hat{u}_i = \hat{D}_y' y_i$$

ont un ratio maximal pour la variance des groupes par rapport à l'échantillon, une corrélation nulle pour

l'échantillon au niveau unitaire et au niveau de groupe, et une variance égale à 1 pour l'échantillon au niveau unitaire. Ces variables sont baptisées variables de groupement canoniques (VGC) de l'échantillon. Les VGC de l'échantillon présentent les plus fortes corrélations à l'intérieur du groupe. Notons que $\text{tr}(S_{yys_1}^{-1}\bar{S}_{yy}) = \sum_k \hat{\theta}_k$ peut correspondre à l'effet d'agrégation pour des variables multiples.

La matrice $\hat{D}_y$ existe même si S_{yys_1} et $\bar{S}_{yy}$ reposent sur des échantillons différents, tant et aussi longtemps que le premier est défini positif et le second, semi-défini positif. Par ailleurs, la variance des VGC n'est pas négative. Toutefois, si s et s_1 diffèrent, il se pourrait que la variance maximale d'une VGC dépasse $(N-1)/(M-1)$, soit l'effet d'agrégation le plus élevé. Dans ce cas, la VGC comprendrait implicitement une composante négative pour la variance à l'intérieur du groupe. La chose ne présente guère d'importance puisque nous cherchons à identifier les variables de groupement importantes. En principe cependant, la variance fautive pourrait être fixée au maximum théorique. On obtient les VGC de l'échantillon à partir des vecteurs propres de $A_{yy} = S_{yys_1}^{-1}\bar{S}_{yy}$. Si s et s_1 sont le même échantillon, A_{yy} est le coefficient de régression de l'échantillon pour la régression des moyennes de groupe sur les valeurs de niveau unitaire pour l'échantillon. En fait, les VGC représentent les valeurs canoniques qui relient les données de niveau unitaire de l'échantillon à celles du niveau du groupe, tandis que $\hat{\theta}_k$ sont les corrélations canoniques de l'échantillon.

Une fois les VGC calculées, il est possible d'exprimer l'écart entre la matrice des covariances de l'échantillon au niveau du groupe et au niveau unitaire de la façon suivante:

$$\bar{S}_{yy} - S_{yys_1} = \sum_k (\hat{\theta}_k - 1)\hat{\phi}_k\hat{\phi}_k'$$

où $\hat{\phi}_k$ correspond au vecteur des covariances de l'échantillon entre la k-ième VGC et les variables originales. On peut donc diviser l'écart entre la matrice des covariances du niveau du groupe et celle du niveau unitaire en k composantes orthogonales, soit une pour chaque VGC.

En ce qui concerne la covariance entre y_{ia} et y_{ib}, l'écart entre la covariance de l'échantillon au niveau du groupe $\bar{s}_{ab}$ et celle du niveau unitaire s_{ab} ($\bar{s}_{ab}$ et s_{ab} étant respectivement des éléments de $\bar{S}_{yy}$ et de S_{yys_1}) est donnée par

$$\bar{s}_{ab} = s_{ab} + (s_{aa}s_{bb})^{1/2}\sum_k (\hat{\theta}_k - 1)\hat{\rho}_{ak}\hat{\rho}_{bk}$$

où $\hat{\rho}_{ak} = \hat{\phi}_{ak}/s_{aa}^{1/2}$ représente la corrélation pour la a-ième variable et la k-ième VGC de l'échantillon.

En utilisant les q premières VGC de l'échantillon pour créer une matrice des variances corrigée au niveau du groupe, c'est-à-dire $\hat{u}_{qi} = \hat{D}_q' y_i$ où $\hat{D}_q = [\hat{d}_1, \ldots, \hat{d}_q]$ servent de variables auxiliaires,

$$\hat{\Sigma}_{yy}(\hat{u}_q) = \bar{S}_{yy} + \bar{B}_{yu_q}'(S_{u_qu_qs_0} - \bar{S}_{u_qu_q})\bar{B}_{yu_q}$$

on élimine les q premiers termes de la décomposition, donc

$$\hat{\Sigma}_{yy}(\hat{u}_q) = S_{yys_1} + \sum_{k=q+1}^{p}(\hat{\theta}_k - 1)\hat{\phi}_k\hat{\phi}_k'$$

et $\text{tr}(S_{yys_1}^{-1}\hat{\Sigma}_{yy}(\hat{u}_q)) = \sum_{k=q+1}^{p}\hat{\theta}_k$. En fait, quand on prend les q premières VGC, on obtient la matrice de rang q qui minimise $\| S_{yys_1} - \hat{\Sigma}_{yy}(\hat{u}_q) \|$. Par conséquent, en étudiant les quantités

$$\sum_{k=q+1}^{p}\hat{\theta}_k \quad \text{et} \quad 1 + \sum_{k=q+1}^{p}(\hat{\theta}_k - 1)\hat{\rho}_{ak}^2$$

pour $q = 0, \ldots, p-1$

il est possible d'établir comment les q premières VGC de l'échantillon expliqueront la partie de l'effet d'agrégation global et l'effet d'agrégation individuel de chaque variable.

Une telle analyse indiquera combien de dimensions sont nécessaires pour expliquer les effets cumulatifs et en supprimer une partie précise. Par ailleurs, en examinant la charge des variables à l'origine des VGC, on devrait être en mesure d'identifier les variables qui «expliqueront» le mieux les effets cumulatifs des autres variables. Les chercheurs devraient s'efforcer d'obtenir les données de niveau unitaire sur ces variables afin de les appliquer à l'estimateur corrigé.

Les résultats qui précèdent ont certaines implications importantes quant à l'usage des données au niveau du groupe complétées par des données de niveau unitaire limitées. En effet, ils autorisent la combinaison des données d'enquête et des données de groupe d'une ou de plusieurs sources, et laissent entrevoir une stratégie d'analyse pour les effets de groupe et les données au niveau du groupe.

4. QUELQUES RÉSULTATS EMPIRIQUES

Nous illustrerons les principes énoncés précédemment par une analyse des données du recensement de la population du Royaume-Uni effectuée en 1991 pour le district local (DL) de Reigate, Banstead et Tandridge. Ce DL compte 188,700 habitants répartis dans 371 DR, ce qui donne une population moyenne de $\bar{n} = 508.6$ par DR. On possède les données de groupe pour les habitants de chaque DR grâce au fichier de données régionales (FDR). Les données de niveau unitaire correspondantes du DL viennent d'un échantillon de 2 pour cent d'enregistrements sur des sujets anonymes. Il est donc impossible d'associer un enregistrement à un DR précis, de telle sorte que $\bar{S}_{yy}$ repose sur les données complètes du fichier se rapportant au DR et que l'estimation de $\bar{S}_{yys_1}$ repose sur l'échantillon de 2 pour cent d'enregistrements anonymes. L'analyse qui suit tient compte de 16 variables de recensement, pour chaque personne.

On s'est servi des données de groupe et des données de niveau unitaire de chaque variable pour calculer l'effet cumulatif $\bar{Q}_a = \bar{s}_{aa}/s_{aas}$. Le paramètre $\delta_{aa} = \Delta_{aa}/\Sigma_{aa}$,

Tableau 1

Effet d'agrégation et corrélation à l'intérieur de la classe pour les variables du recensement dans le DL de Reigate

	Effet d'agrégation	Corrélation
18 à 29 ans	9.20	.016
30 à 44 ans	4.56	.007
45 à 59 ans*	5.97	.010
60 ans et plus*	17.17	.032
Femme	1.08	.000
Personne de couleur*	8.29	.014
Marié	6.24	.010
Maladie débilitante à long terme	7.24	.012
Travailleur à plein temps	8.55	.015
Chômeur	2.27	.003
Autre situation d'emploi	11.19	.020
Soutien de famille né au R.-U.	4.48	.007
Soutien de famille né dans le Nouveau Commonwealth	3.59	.005
Soutien de famille émigré	9.04	.016
≤ 1.5 personne par pièce: densité	27.96	.053
Personnes dans un ménage sans automobile	32.98	.063

* Variable retenue pour la correction.

Source: Reigate et Banstead; données du recensement de 1991 pour le DL de Tandridge.

défini pour les éléments diagonaux appropriés de Δ_{yy} et de Σ_{yy}, correspond à la corrélation de la a-ième variable à l'intérieur du groupe. On peut obtenir l'estimation $\hat{\delta}_{aa}$ de la corrélation à l'intérieur du groupe à partir de (2.18), puisque $\bar{Q}_a = 1 + (\bar{n}^* - 1)\,\hat{\delta}_{aa}$. Les résultats relatifs à ces variables apparaissent au tableau 1. En général, les corrélations à l'intérieur du groupe sont faibles, mais étant donné le nombre d'observations pour chaque DR, les effets cumulatifs peuvent être importants (voir la remarque qui suit l'équation (2.18)).

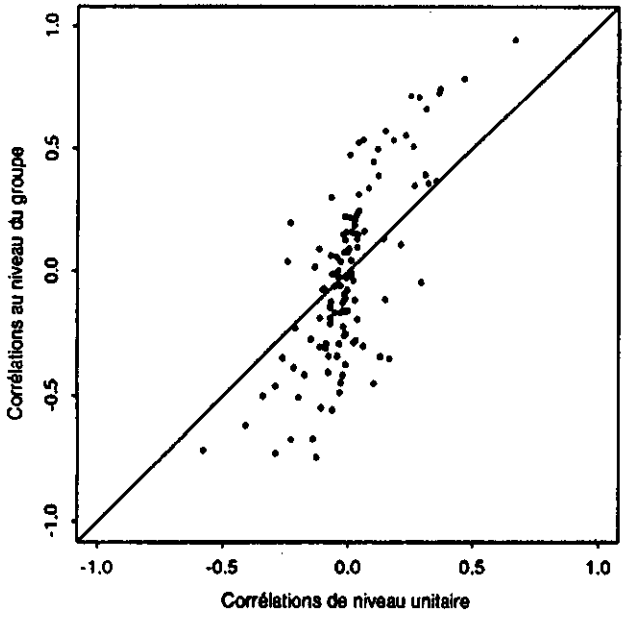

Figure 1a.

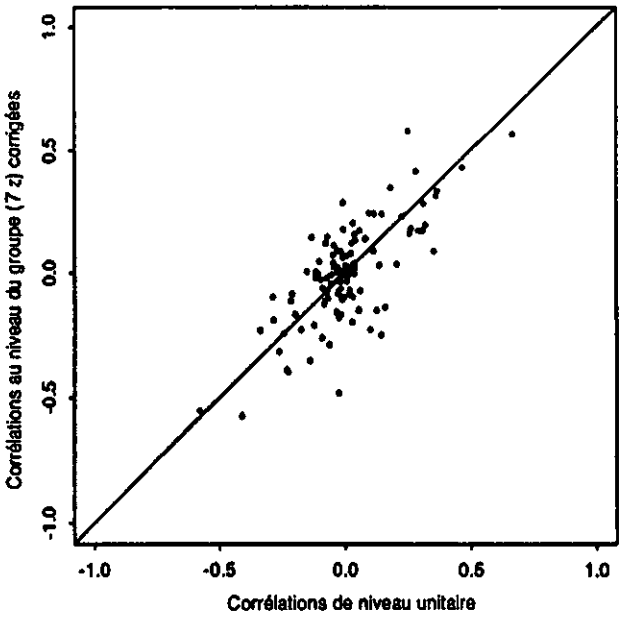

Figure 1b.

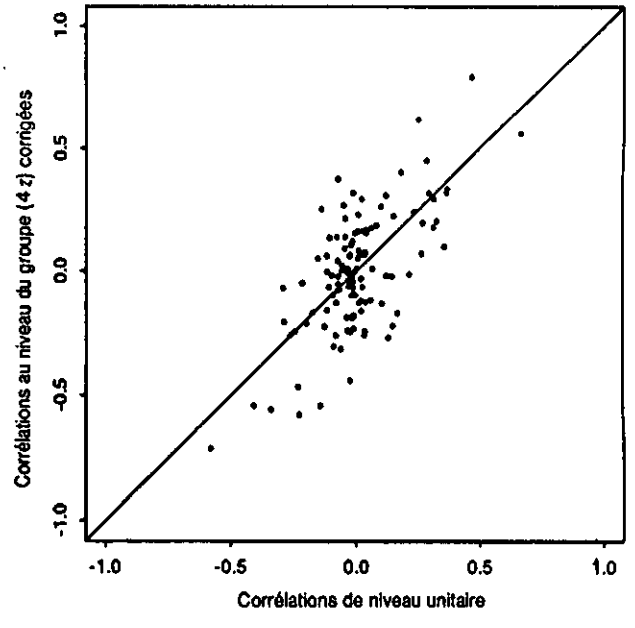

Figure 1c.

La figure 1a montre la courbe de la corrélation au niveau du groupe $\bar{r}_{ab}$ par rapport à la corrélation de niveau individuel r_{ab} pour chaque paire de variables. La forme en S caractéristique de la courbe fait ressortir l'importance des effets dus à l'agrégation. Les petites corrélations de niveau unitaire sont habituellement amplifiées si bien que dans la plupart des cas, $|\,\bar{r}_{ab}\,|$ est beaucoup plus grand que $|\,r_{ab}\,|$.

Tableau 2

Cinq premières VGC des variables du tableau 1

	VGC1	VGC2	VGC3	VGC4	VGC5
18 à 29 ans	0.4	0.3	0.9	1.1	0.1
30 à 44 ans	0.1	0.5	0.36	1.0	0.2
45 à 59 ans*	−0.1	1.2	−0.2	1.0	0.1
60 ans et plus*	0.3	2.2	−0.5	2.6	0.9
Femme	0.1	0.0	0.0	0.3	0.1
Personne de couleur*	0.5	−0.4	1.4	−1.1	5.2
Marié	−0.2	−0.5	−0.4	−0.8	−0.1
Maladie débilitante à long terme	0.3	0.1	−0.2	0.2	0.3
Travailleur à temps plein	0.7	−0.3	0.2	1.2	0.4
Chômeur	0.7	0.0	−0.1	0.0	−0.4
Autre situation d'emploi	0.1	0.1	0.0	−0.2	−0.1
Soutien de famille né au R.-U.	0.5	−0.1	−1.0	0.4	0.2
Soutien de famille né dans le Nouveau Commonwealth	0.0	−0.1	−0.3	0.1	0.6
Soutien de famille émigré	0.2	0.1	1.4	0.6	−1.3
≤ 0.5 personne par pièce	−1.4	0.3	1.2	−0.7	−0.2
Personnes dans un ménage sans automobile	2.2	0.6	0.8	−1.9	−0.7

* Variable retenue pour la correction.
Source: Reigate et Banstead; données du recensement de 1991 pour le DL de Tandridge.

Connaissant $\bar{S}_{yy}$ et S_{yys_1} on peut analyser les variables de groupement canoniques en vue de comprendre la structure particulière du groupe. Le tableau 2 donne la charge des cinq premières variables de groupement canoniques. Ces cinq variables expliquent 89% de l'effet cumulatif des 16 variables utilisées.

La première VGC influe fortement sur la densité de population dans le logement et sur l'accès à une automobile. On pourrait donc la considérer comme un paramètre socio-économique. La deuxième attribue une forte charge aux variables qui identifient les personnes des deux groupes les plus âgés. La contribution des chefs de famille de couleur à la variable suivante est également remarquable. Comme on pouvait s'y attendre, certaines variables comme la proportion de femmes ne présentent pratiquement aucune corrélation à l'intérieur du groupe, donc n'ont aucun effet cumulatif et n'exercent presque aucune influence sur les VGC. Pareilles variables ne changent pas d'une région à l'autre. Elles n'ont donc habituellement aucune valeur explicative.

Dans la pratique, pareille analyse des VGC serait irréalisable, l'existence même de S_{yy} rendant une analyse agrégée inutile. L'analyse des VGC a néanmoins le mérite d'identifier les variables les plus importantes en raison de la charge qu'elles imposent aux premières VGC.

On sait pertinemment qu'au Royaume-Uni, les variables associées au mode d'habitation (qui ne font pas partie des 16 variables intéressantes) partagent des liens très étroits avec maintes variables liées à la situation socio-économique, au comportement et à la santé. Il y a tout lieu de croire qu'en les utilisant comme variables auxiliaires z pour la correction, on tiendrait compte d'une bonne partie de la première dimension socioéconomique. Elles pourraient donc remplacer les variables illustrant la densité de la population dans le logement et l'accès à une automobile, qu'on estime influer fortement sur la première VGC. L'autre raison pour envisager l'utilisation de ces variables est que si l'analyse actuelle doit illustrer ce qui pourrait se produire dans d'autres situations, les variables de base sur le mode d'occupation et sur le logement seront sans doute plus faciles à obtenir que celles sur la densité de la population dans le logement et l'accès à une automobile pour la correction. Compte tenu des résultats de l'analyse des VGC et puisqu'on souhaite des variables d'ajustement faciles à retrouver, peu importe la situation, nous proposons les sept variables d'ajustement potentielles qui suivent, soit les trois variables d'intérêt identifiées au tableau 1 par un astérisque (45-59 ans, 60 ans +, personnes de couleur) et les quatre variables liées au logement du tableau 3, avec leurs effets cumulatifs et leurs corrélations à l'intérieur de la grappe.

Tableau 3

Effet d'agrégation et corrélation à l'intérieur des variables liées au ménage dans le DL de Reigate

Variable		Effet d'agrégation	Corrélation
Mode d'occupation:	Locataire	133.43	0.261
	Propriétaire occupant	90.83	0.177
Type d'habitation:	Unif./semi/terrasse	90.03	0.175
	Commodités adéquates	59.52	0.113

Source: Reigate et Banstead; données du recensement de 1991 pour le DL de Tandridge.

Nous corrigerons ensuite la matrice des covariances au niveau du groupe des 16 variables originales d'après la matrice des covariances de niveau unitaire pour les 7 variables z (trois variables démographiques de base de la série originale et quatre variables relatives au ménage).

Deux mesures globales déterminent l'efficacité de l'ajustement. La première est donnée par

$$1 - \frac{\mathrm{tr}\left(S_{yys_1}^{-1}\,\hat{\Sigma}_{yy}\,(z)\right) - 1}{\mathrm{tr}\left(S_{yys_1}^{-1}\,\bar{S}_{yy}\right) - 1}$$

qui est la réduction de l'effet agrégé multivarié. La seconde est

$$\frac{\| S_{yys_1} - \bar{S}_{yy} \| - \| S_{yys_1} - \hat{\Sigma}_{yy}(z) \|}{\| S_{yys_1} - \bar{S}_{yy} \|}$$

et indique la réduction de la distance généralisée entre les deux matrices de covariances aux niveaux unitaire et de groupe, avant et après la correction.

Le tableau 4 illustre l'effet de diverses combinaisons de variables sur la correction des résultats de l'analyse agrégée. Les deux variables associées à l'âge présentent manifestement de l'importance (elles expliquent 38% de l'effet d'agrégation multivariée et 53% de l'écart généralisé), mais il en va autant des variables sur le mode

Tableau 4

Combinaison des variables Z	Nombre de variables	Réduction de (%)	
		Effet d'agrégation multivarié	Écart général
60 ans et plus	1	16	24
45-59, 60+	2	38	53
Mode d'occupation	2	30	21
Type d'habitation	2	31	19
45-59, 60+, personne de couleur	3	44	54
45-59, 60+, mode d'occupation	4	57	71
45-59, 60+, type d'habitation	4	57	69
45-59, 60+, mode d'occupation, personne de couleur	5	63	72
45-59, 60+, type d'habitation, personne de couleur	5	62	70
45-59, 60+, type d'habitation, mode d'occupation, personne de couleur	7	68	75

d'occupation ou le type d'habitation. Quand on combine ces dernières aux variables liées à l'âge, la réduction obtenue avec chaque mesure, en pourcentage, se rapproche de la somme des effets des variables prises séparément, signe que l'âge et le mode d'occupation ou le type d'occupation entraînent des corrections distinctes. Comme on peut le constater, les meilleurs résultats proviennent de l'application des 7 variables, puisque ces dernières expliquent respectivement 68% et 75% de la correction des deux mesures de l'agrégation.

Les résultats révèlent que la correction supprime environ 70% des effets cumulatifs. Les figures 2a et 2b montrent l'effet de l'ajustement attribuable aux variables. À la figure 2a, l'axe des ordonnées intègre $| s_{ab} - s_{abs_1} |$ le biais absolu de la covariance de chaque paire de variables au niveau du groupe. De son côté, l'axe des abscisses intègre $| \hat{\Sigma}_{ab}(z) - s_{abs_1} |$, le biais absolu de l'estimateur corrigé. La variance des variables est représentée par le symbole vide et la covariance, par le symbole plein. Pratiquement toutes les valeurs du graphique révèlent un biais plus faible (souvent beaucoup plus) que le biais original après correction. L'ajustement entraîne une amélioration appréciable dans presque tous les cas. La figure 2b trace les corrélations plutôt que les covariances (les corrélations y_a, y_a ont été omises pour des raisons évidentes). Encore une fois, on note une nette amélioration de l'analyse au niveau du groupe, le biais résiduel étant beaucoup plus faible que le biais original après la correction. Les résultats sont néanmoins inférieurs à ceux obtenus pour les covariances car, dans certains cas, de légers biais ont empiré lors de l'analyse au niveau du groupe. Ici, les corrections s'appliquent à la covariance et aux deux variances utilisées pour chaque coefficient de corrélation. Il y a plus de risques que la correction relative de chaque élément débouche

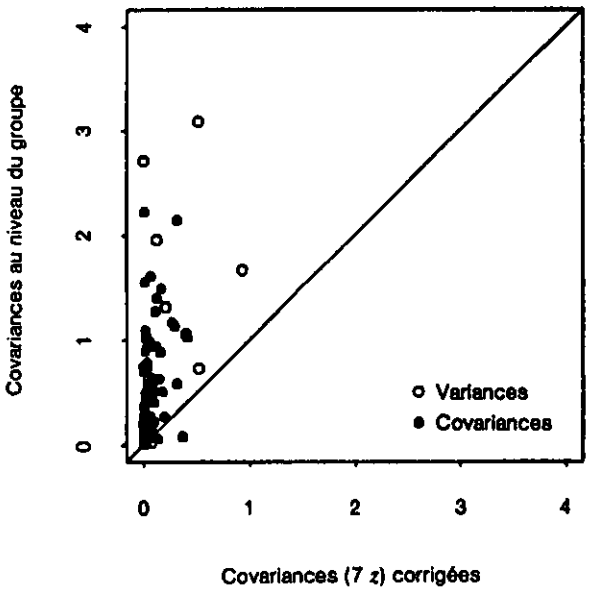

Figure 2a.

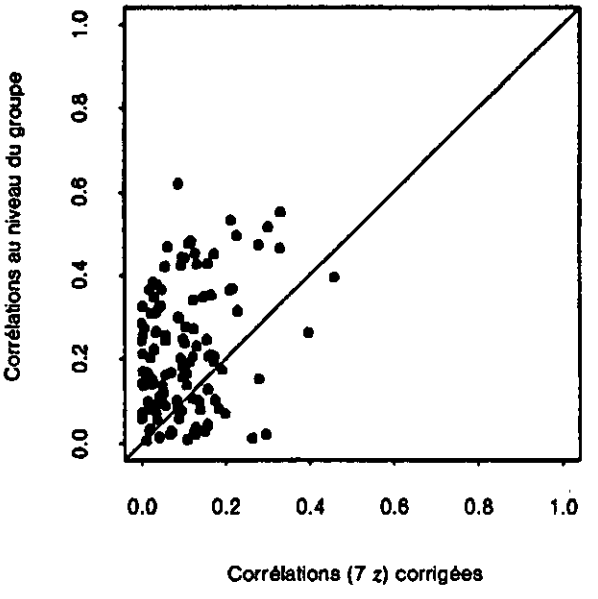

Figure 2b.

sur une moins bonne corrélation que l'originale. La plupart des biais importants relevés au niveau du groupe ont toutefois été atténués.

La figure 1b trace les corrélations corrigées au niveau du groupe $\bar{r}_{ab}(z)$ tirées de $\hat{\Sigma}_{yy(z)}$ par rapport aux corrélations de niveau unitaire. Elle peut être comparée à la figure 1a qui représente les corrélations non corrigées. La courbe en S caractéristique de la figure 1a a été remplacée par une série de points autour de la ligne $\bar{r}_{ab}(z) = r_{ab}$ ce qui est exactement ce qu'on espère obtenir en éliminant le biais agrégé.

Les figures 1b, 2a et 2b révèlent qu'on peut réduire sensiblement l'effet cumulatif en utilisant 4 variables liées à l'habitation et 3 des variables y originales. Les 120 variances et

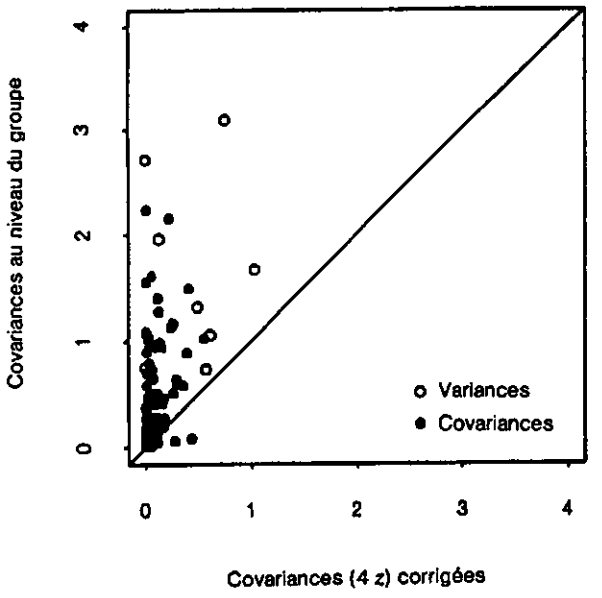

Figure 3a.

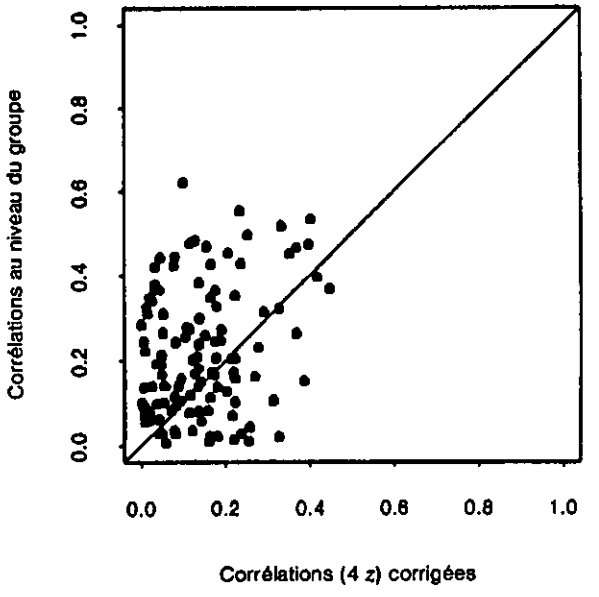

Figure 3b.

covariances de départ de la matrice 16 × 16 doivent donc être corrigées au moyen des 21 variances et covariances des variables z. Pour avoir une idée des résultats qu'on pourrait obtenir avec un minimum d'information, nous avons restreint les variables d'ajustement aux quatre variables liées à l'âge et au mode d'occupation. Le tableau 4 indique que ces variables expliquent respectivement 57% et 71% des deux mesures de l'agrégation. Les figures 3a et 3b donnent le tracé correspondant aux figures 2a et 2b pour ce cas particulier. La figure 1c compare les corrélations corrigées au moyen des quatre variables aux corrélations de niveau individuel. Bien sûr, la correction n'est pas aussi efficace, mais il est encourageant de voir ce qu'on obtient avec si peu de variables. L'écart médian absolu entre $\bar{r}_{ab}$ et r_{ab} (0.186) donne une indication supplémentaire des conséquences de la correction. Après correction au moyen des quatre variables, l'écart n'est plus que de 0.126 (0.090 après correction au moyen de sept variables). Les valeurs médianes correspondantes pour $| \bar{s}_{ab} - s_{ab} |$ sont respectivement de 0.173, de 0.039 et de 0.017.

5. CONCLUSIONS ET DISCUSSION

Les auteurs proposent un modèle qui décompose en deux éléments le biais observé dans l'analyse au niveau du groupe au moyen des matrices des covariances pour des populations groupées. Le premier élément résulte des variables de groupement et le second, des corrélations résiduelles entre les variables y à l'intérieur du groupe, compte tenu des variables de groupement z. Pareille décomposition nous aide à comprendre l'ampleur des effets cumulatifs. Elle indique aussi comment supprimer le biais attribuable aux variables de groupement quand on dispose de renseignements supplémentaires sur la matrice des covariances de niveau unitaire des variables de groupement.

Beaucoup de données de groupe à divers degrés d'agrégation peuvent être extraites du recensement et de nombreuses autres sources dans maints pays. L'expansion des systèmes d'information géographique accroîtra l'accessibilité de telles données. Il est important d'analyser et de décomposer les effets de groupe. La théorie et la méthode exposées ici procurent le cadre nécessaire à un tel exercice. En identifiant les variables qui expliquent la majeure partie des effets de groupe, donc qui devraient permettre la correction des analyses écologiques, on ouvrira la voie à l'utilisation des données agrégées.

REMERCIEMENTS

Le présent projet de recherche n'aurait pu être mené à bien sans la subvention H507 26 5013 de l'Economic and Social Research Council du Royaume-Uni. Les auteurs remercient sincèrement les examinateurs et le rédacteur adjoint pour leurs commentaires utiles.

BIBLIOGRAPHIE

ARBIA, G. (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems.* Dordrecht: Kluman.

BLALOCK, H.M. (1964). *Causal Inference in Nonexperimental Research.* Chapel Hill NC: University of North Carolina Press.

BLALOCK, H.M. (1979). Measurement and conceptualization problems: The major obstacle to integrating theory and research. *American Sociological Review*, 44, 881-894.

BLALOCK, H.M. (1985). Cross level analysis. Dans *The Collection and Analysis of Community Data*, (Éd. J.B. Casterlin), ISI, World Fertility Survey.

CLARK, W.A.V., et AVERY, K.L. (1976). The effect of data aggregation in statistical analysis. *Geographical Analysis*, 8, 428-438.

DUNCAN, D.P., et DAVIS, B. (1953). An alternative to ecological correlation. *American Sociological Review*, 18, 665-666.

FOTHERINGHAM, A.S., et WONG, D.W.S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning*, A, 23, 1025-1044.

GEHLKE, C.E., et BIEHL, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29, Supplement, 169-170.

GOODMAN, L.A. (1959). Some alternatives to ecological correlation. *American Journal of Sociology*, 64, 610-625.

HANNAN, M.T., et BUSTEIN, L. (1974). Estimation from grouped observations. *American Sociological Review*, 39, 374-392.

HOLT, D., SMITH, T.M.F., et WINTER, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society*, A, 143, 474-87.

HOLT, D., et SCOTT, A.J. (1981). Regression analysis using survey data. *The Statistician*, 30, 169-173.

LICHTMAN, A.J. (1974). Correlation, regression, and the ecological fallacy: A critique. *Journal of Interdisciplinary History*, 4, 417-433.

LANGBEIN, L.I., et LICHTMAN, A.J. (1978). *Ecological Inference*. Thousand Oaks, CA: Sage.

OPENSHAW, S. (1984). Ecological fallacies and the analysis of areal census data. *Environment and Planning*, A, 6, 17-31.

OPENSHAW, S., et TAYLOR, P.J. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. Dans *Statistical Applications in the Spatial Sciences*, (Éd. N. Wrigley), 127-144.

PEARSON, K. (1903). On the influence of natural selection on the variability and correlation of organs. *Philological Transactions of the Royal Society*, A, 200, 1-66.

PERLE, E.D. (1977). Scale changes and impacts on factorial ecology structures. *Environment and Planning*, A, 9, 549-558.

RAO, C.R. (1973). *Linear Statistical Inference and its Applications*, (2ième éd.). New York: Wiley.

ROBINSON, W.S. (1950). Ecological correlations and the behaviour of individuals. *American Sociological Review*, 15, 351-357.

SMITH, K.W. (1977). Another look at the clustering perspective on aggregation problems. *Sociological Methods and Research*, 5, 289-316.

SMITH, T.M.F., et HOLMES, D. (1989). Multivariate analysis. Dans *Analysis of Complex Surveys*, (Éds. C.J. Skinner, D. Holt et T.M.F. Smith), 165-187.

STEEL, D. (1985). Statistical Analysis of Populations with Group Structure. Unpublished PhD Thesis, Department of Social Statistics, University of Southampton.

STEEL, D., et HOLT, D. (1994). Analysing and Adjusting Aggregation Effects: The Ecological Fallacy Revisited. Department of Applied Statistics, University of Wollongong, prétirage 1/94.

STEEL, D., et HOLT, D. (1995). Rules for random aggregation. *Environment and Planning* (à paraître).

YULE, U., et KENDALL, M.S. (1950). *An Introduction to the Theory of Statistics*. Glendale, CA: Griffin.

Techniques d'enquête, juin 1996
Vol. 22, n° 1, pp. 17–22
Statistique Canada

Méthodes de linéarisation pour les échantillons à une et deux phases: Une approche de type «recette»

DAVID A. BINDER[1]

RÉSUMÉ

Il existe un certain nombre de méthodes asymptotiquement équivalentes qui permettent de calculer l'approximation, par série de Taylor, des variances pour les statistiques complexes. Binder et Patak (1994) ont présenté la justification théorique pour une classe de méthodes. Toutefois, on peut établir bon nombre de ces méthodes à partir d'exemples pratiques en utilisant des techniques directes qui ne sont pas décrites clairement dans Binder et Patak. Dans cette communication, nous présentons une approche de type «recette», utilisable pour de nombreux exemples et dont les bonnes propriétés d'échantillonnage de taille finie ont été démontrées. La méthode retenue devient normalement claire lorsque l'on utilise des concepts comme méthodes basées sur un modèle ou linéarisation de la méthode du jackknife; toutefois, notre approche permet d'obtenir les résultats recherchés de façon plus directe. En outre, nous présentons de nouveaux résultats pour l'application de ces techniques à des échantillons à deux phases.

MOTS CLÉS: Enquêtes complexes; estimation de la variance; estimateur par quotient; estimateur par régression; test de la somme des rangs de Wilcoxon; équations d'estimation.

1. LA MÉTHODE

Le calcul de la variance asymptotique pour une grande classe d'estimateurs provenant d'échantillons d'enquête complexes est maintenant bien décrit dans la littérature, du moins pour les approximations du premier ordre. Toutefois, il existe un certain nombre d'autres estimateurs de la variance, tous étant asymptotiquement équivalents. Dans la présente communication, nous présentons une méthode simple pour calculer, de façon générale, l'un de ces estimateurs les plus utilisés. Cette méthode simple de calcul peut s'avérer utile pour les statisticiens qui peuvent hésiter devant les choix disponibles et recherchent une solution rapide à leur problème.

Nous débutons par un exemple simple de la méthode utilisant l'estimateur par quotient pour la valeur totale d'un ensemble. Dans ce cas-ci, l'estimateur est:

$$\hat{Y}_R = \hat{R}X, \tag{1}$$

pour

$$\hat{R} = \hat{Y}/\hat{X}, \quad \text{et} \quad \hat{Y} = \sum_{k \in s} w_k y_k,$$

où s est l'ensemble des indices correspondant aux unités échantillonnées et w_k est le poids d'échantillonnage, normalisé de telle sorte que $\sum w_k$ soit un estimateur de la valeur totale de l'ensemble; p. ex., $w_k = 1/\pi_k$, où π_k est la probabilité d'inclusion du premier ordre. La définition de $\hat{X}$ est analogue à celle de $\hat{Y}$. En prenant la différentielle totale des deux membres de l'expression (1), nous obtenons:

$$(d\hat{Y}_R) = (d\hat{R})X, \tag{2a}$$

où

$$(d\hat{R}) = \frac{(d\hat{Y})}{\hat{X}} - \frac{\hat{Y}}{\hat{X}^2}(d\hat{X}) \tag{2b}$$

$$= \frac{1}{\hat{X}}[(d\hat{Y}) - \hat{R}(d\hat{X})].$$

Nous constatons que la différentielle totale pour $\hat{T} = g(\hat{Y}_1, \ldots, \hat{Y}_m)$ est donnée, en général, par:

$$(d\hat{T}) = \sum \left[\frac{\partial g(\hat{Y})}{\partial \hat{Y}_i}\right](d\hat{Y}_i).$$

Nous aurions pu toutefois éviter l'utilisation de $\hat{R}$ dans l'expression (1) en définissant tout simplement

$$Y_{\hat{R}} = \frac{\hat{Y}}{\hat{X}}X,$$

ce qui élimine la nécessité de définir explicitement $(d\hat{R})$ dans (2b), mais nous l'avons fait afin de rendre plus clairs les exemples complexes qui seront présentés à la section 1.2. Signalons également que l'expression (2a) ne comprend pas la différentielle totale de X, qui est la valeur totale de la variable x pour l'ensemble, car X est supposé être fixe et connu.

L'étape suivante consiste à remplacer toutes les différentielles totales des quantités estimées par les écarts par rapport à leurs valeurs probables respectives. Dans le membre de droite, nous remplaçons $(d\hat{Y})$ par l'expression $(\sum w_k y_k - Y)$, *etc.* Pour la quantité qui nous intéresse, soit $\hat{Y}_R$, nous remplaçons $d\hat{Y}_R$ par $\hat{Y}_R - Y$. À partir de l'expression (2), et faisant ces substitutions, nous obtenons:

$$\hat{Y}_R - Y \doteq \frac{X}{\hat{X}}\left[\left(\sum w_k y_k - Y\right) - \hat{R}\left(\sum w_k x_k - X\right)\right]. \tag{3}$$

[1] David A. Binder, Directeur, Division des méthodes d'enquêtes-entreprises, Statistique Canada, R.H. Coats, 11 A, Ottawa (Ontario), Canada, K1A 0T6.

Nous constatons que cette expression contient un certain nombre d'estimateurs pondérés, ceux qui ont une dépendance explicite par rapport aux valeurs w_k, ($\sum w_k y_k$ et $\sum w_k x_k$) et ceux pour lesquels les valeurs w_k sont implicites dans l'expression ($\hat{X}$ et $\hat{R}$).

La dernière étape consiste à isoler z_k, ce que l'on fait en réécrivant l'expression (3) sous la forme:

$$\hat{Y}_R - Y \doteq \sum w_k z_k + \begin{array}{l}\text{autres termes ne dépendant}\\ \text{pas } \textit{explicitement} \text{ de } w_k.\end{array}$$

Nous obtenons ainsi:

$$z_k = \frac{X}{\hat{X}}(y_k - \hat{R}x_k). \tag{4}$$

Nous justifierons à la section 4 la raison pour laquelle nous ignorons les termes qui ne dépendent pas explicitement de w_k. On remarquera que $\sum w_k z_k$ a la forme d'une estimation de la valeur totale de la variable z pour l'ensemble.

Pour obtenir la variance de $\hat{Y}_R$, nous insérons la nouvelle variable z_k dans l'enregistrement du k-ième échantillon, et nous utilisons une méthode standard pour calculer la variance d'une valeur totale, appliquée à cette variable. Nous supposons qu'il existe un estimateur de variance possédant de bonnes propriétés pour le plan d'échantillonnage à l'étude.

Nous pouvons résumer comme suit la méthode:

1. Comme estimateur de T, nous utilisons $\hat{T}$ et nous prenons sa différentielle totale. Nous supposons que $\hat{T}$ est asymptotiquement convergent avec le plan.
2. Nous remplaçons la différentielle totale de $\hat{T}$, $d\hat{T}$, par $\hat{T} - T$. Nous remplaçons toutes les autres différentielles totales des quantités estimées par l'écart par rapport à leurs valeurs probables respectives, comme nous avons remplacé $(d\hat{Y})$ par l'expression $(\sum w_k y_k - Y)$, *etc.*
3. La dernière étape consiste à isoler z_k, lorsque nous réécrivons le résultat de l'étape (2) sous la forme suivante:

$$\hat{T} - T \doteq \sum w_k z_k + \begin{array}{l}\text{autres termes ne dépendant}\\ \text{pas } \textit{explicitement} \text{ de } w_k.\end{array}$$

4. Enfin, pour obtenir la variance estimée de $\hat{T}$, nous insérons la nouvelle variable z_k dans chaque enregistrement échantillonné, et nous utilisons la méthode standard (dont on sait qu'elle a de bonnes propriétés) afin de calculer la variance d'un total, appliqué à cette variable.

1.1 Cas général le plus simple

Pour les échantillons à une phase, examinons un cas général simple où l'estimateur peut s'exprimer sous la forme d'une fonction différentiable des totaux estimés pour certaines variables de l'enquête, dont certaines peuvent être calculées au niveau de l'unité d'échantillonnage finale. Dans ce cas-ci, notre méthode donne les résultats suivants:

$$\hat{T} = g(\hat{Y}_1, \ldots, \hat{Y}_m)$$

$$(d\hat{T}) = \sum \left[\frac{\partial g(\hat{Y})}{\partial \hat{Y}_i}\right](d\hat{Y}_i)$$

$$\hat{T} - T \doteq \sum_i \left[\frac{\partial g(\hat{Y})}{\partial \hat{Y}_i}\right]\left(\sum_k w_k y_{ik} - Y_i\right) \tag{5}$$

$$= \sum w_k z_k + \ldots,$$

où

$$z_k = \sum_i \left[\frac{\partial g(\hat{Y})}{\partial \hat{Y}_i}\right] y_{ik} = \left[\frac{\partial g(\hat{Y})}{\partial \hat{Y}}\right]' y_k. \tag{6}$$

En quoi cette formulation diffère-t-elle des méthodes standard qui utilisent l'expression de Taylor? La principale différence réside dans le traitement de l'expression (5). Dans les méthodes standard, on calcule les dérivées partielles à leurs valeurs probables avant de calculer z_k. Ici, pour les composantes de z_k qui sont inconnues, on substitue un estimateur. Pour l'estimateur par quotient, (1), cela ferait disparaître $X/\hat{X}$ de z_k dans l'expression (4), car $\hat{X}$ est remplacé par sa valeur probable. La valeur $X/\hat{X}$ devient égale à l'unité. Le $\hat{R}$ demeure dans l'expression, car on l'utilise pour calculer R, dont la valeur est requise pour le calcul habituel de z_k.

Kott (1990) soutient que l'estimateur de variance pour le quotient que nous avons calculé possède de bonnes propriétés conditionnelles, par rapport à l'estimateur qui omet le facteur $X/\hat{X}$. Plusieurs autres auteurs en sont arrivés à des conclusions similaires. Rao (1995) a démontré que la méthode donne les mêmes résultats que la méthode linéarisée du jackknife. Notre conjecture est la suivante: comme les dérivées partielles dans l'expression (5) sont évaluées à la valeur $\hat{Y}$ plutôt qu'à la valeur Y, la linéarisation «épouse» mieux la statistique originale $\hat{T}$, de sorte que les variances qui en découlent ont de meilleures propriétés. Cette conjecture n'est évidemment pas une démonstration technique, mais plutôt une justification intuitive de la méthode.

On remarquera que, dans l'expression (6) pour z_k, tous les termes sont directement obtenus de l'échantillon, de sorte qu'il n'y a pas lieu de substituer des estimateurs pour des quantités inconnues.

1.2 Cas avec paramètres additionnels

Dans de nombreux exemples, on définit plus facilement l'estimateur en termes qui comprennent l'utilisation de paramètres servant uniquement à simplifier la définition du paramètre qui nous intéresse. Pour l'estimateur par quotient, $\hat{R}$ est un exemple d'un tel *paramètre additionnel*. Dans ce cas précis, on dispose d'une équation explicite pour l'estimateur du paramètre additionnel. En présence de paramètres additionnels, on peut écrire comme suit la méthode générale:

$$\hat{T} = g_1(\hat{Y}_1, \ldots, \hat{Y}_m, \hat{\lambda}), \quad \text{où} \quad \hat{\lambda} = g_2(\hat{Y}_1, \ldots, \hat{Y}_m),$$

$$(d\hat{T}) = \sum \left[\frac{\partial g_1(\hat{Y}, \hat{\lambda})}{\partial \hat{Y}_i}\right](d\hat{Y}_i) + \sum \left[\frac{\partial g_1(\hat{Y}, \hat{\lambda})}{\partial \hat{\lambda}_j}\right](d\hat{\lambda}_j),$$

où

$$(d\hat{\lambda}_j) = \sum_i \left[\frac{\partial g_{2j}(\hat{Y})}{\partial \hat{Y}_i}\right](d\hat{Y}_i),$$

$$\hat{T} - T \doteq \sum \frac{\partial g_1(\hat{Y}, \hat{\lambda})}{\partial \hat{Y}_i}\left(\sum_k w_k y_{ik} - Y_i\right)$$

$$+ \sum \frac{\partial g_1(\hat{Y}, \hat{\lambda})}{\partial \hat{\lambda}_j} \sum_i \frac{\partial g_{2j}(\hat{Y})}{\partial \hat{Y}_i}\left(\sum_k w_k y_{ik} - Y_i\right)$$

$$= \sum w_k z_k + \ldots,$$

où

$$z_k = \left[\frac{\partial g_1(\hat{Y}, \hat{\lambda})}{\partial \hat{Y}}\right]' y_k + \left[\frac{\partial g_1(\hat{Y}, \hat{\lambda})}{\partial \hat{\lambda}}\right]' \left[\frac{\partial g_2(\hat{Y})}{\partial \hat{Y}}\right] y_k. \quad (7)$$

Lorsque les paramètres additionnels sont définis seulement de manière implicite par les équations d'estimation, nous obtenons la généralisation suivante:

$$\hat{T} = g(\hat{Y}_1, \ldots, \hat{Y}_m, \hat{\lambda}),$$

où

$$\hat{U}(\hat{Y}_1, \ldots, \hat{Y}_m, \hat{\lambda}) = 0. \quad (8)$$

$$(d\hat{T}) = \sum \left[\frac{\partial g(\hat{Y}, \hat{\lambda})}{\partial \hat{Y}_i}\right](d\hat{Y}_i) + \left[\frac{\partial g(\hat{Y}, \hat{\lambda})}{\partial \hat{\lambda}}\right]'(d\hat{\lambda}),$$

en prenant la différentielle totale de (8) et en isolant $(d\hat{\lambda})$, nous obtenons:

$$(d\hat{\lambda}) = -\left[\frac{\partial \hat{U}(\hat{Y}, \hat{\lambda})}{\partial \hat{\lambda}}\right]^{-1} \sum \left[\frac{\partial \hat{U}(\hat{Y}, \hat{\lambda})}{\partial \hat{Y}_i}\right](d\hat{Y}_i). \quad (9)$$

$$\hat{T} - T \doteq \sum_i \left(\frac{\partial g}{\partial \hat{Y}_i}\right)\left(\sum_k w_k y_{ik} - Y_i\right)$$

$$- \left(\frac{\partial g}{\partial \hat{\lambda}}\right)'\left[\frac{\partial \hat{U}}{\partial \hat{\lambda}}\right]^{-1} \sum_i \left(\frac{\partial \hat{U}}{\partial \hat{Y}_i}\right)\left(\sum_k w_k y_{ik} - Y_i\right)$$

$$= \sum w_k z_k + \ldots,$$

où

$$z_k = \left[\frac{\partial g}{\partial \hat{Y}}\right]' y_k - \left[\frac{\partial g}{\partial \hat{\lambda}}\right]' \left[\frac{\partial \hat{U}}{\partial \hat{\lambda}}\right]^{-1} \left[\frac{\partial \hat{U}}{\partial \hat{Y}}\right]' y_k. \quad (10)$$

Nous constatons évidemment que l'expression (10) est une généralisation des formes précédentes pour z_k données en (6) et (7).

2. AUTRES EXEMPLES

Les expressions (6), (7) et (10) ci-dessus sont présentées uniquement afin de donner les formules précises pour les divers cas. Toutefois, dans la pratique, nous recommandons d'utiliser les étapes de base et de partir des principes premiers. Pour démontrer ceci, voici deux exemples. L'un est l'estimateur de régression généralisé (GREG) bien connu; l'autre aboutit à quelques nouveaux résultats pour le test de la somme des rangs de Wilcoxon pour les données des enquêtes complexes.

2.1 Estimateur de régression généralisé

On peut écrire comme suit l'estimateur de régression généralisé (GREG) qui est donné, par exemple, dans Särndal, Swensson et Wretman (1989):

$$\hat{Y}_{GREG} = \hat{Y} + \hat{\beta}'(X - \hat{X}), \quad (11)$$

où le paramètre additionnel $\hat{\beta}$ est défini comme étant la solution à:

$$\sum_k w_k x_k (y_k - x_k' \hat{\beta})/c_k = 0,$$

et où c_k est le facteur qui rend la variance hétéroscédastique dans le modèle de régression. Ceci est équivalent à:

$$\hat{S}_{xx}\hat{\beta} - \hat{S}_{xy} = 0, \quad (12)$$

les définitions de $\hat{S}_{xx}$ et $\hat{S}_{xy}$ étant évidentes. Si l'on prend les différentielles totales dans (12), nous obtenons:

$$(d\hat{S}_{xx})\hat{\beta} + \hat{S}_{xx}(d\hat{\beta}) - (d\hat{S}_{xy}) = 0,$$

de telle sorte que

$$(d\hat{\beta}) = \hat{S}_{xx}^{-1}[(d\hat{S}_{xy}) - (d\hat{S}_{xx})\hat{\beta}].$$

Nous avons donc:

$$\hat{\beta} - \beta \doteq \sum w_k \hat{S}_{xx}^{-1}[x_k(y_k - x_k'\hat{\beta})]/c_k + \ldots.$$

Maintenant, en prenant les différentielles totales de (11), nous obtenons:

$$(d\hat{Y}_{GREG}) = (d\hat{Y}) - \hat{\beta}'(d\hat{X}) + (d\hat{\beta})'(X - \hat{X})$$
$$= (d\hat{Y}) - \hat{\beta}'(d\hat{X}) + [(d\hat{S}'_{xy}) - \hat{\beta}'(d\hat{S}_{xx})]S_{xx}^{-1}(X - \hat{X}).$$

Après quelques manipulations algébriques, nous obtenons:

$$\hat{Y}_{GREG} - Y = \sum w_k e_k[1 + x'_k \hat{S}_{xx}^{-1}(X - \hat{X})/c_k] + \ldots,$$

où $e_k = y_k - x'_k\hat{\beta}$. Nous définissons donc:

$$z_k = e_k[1 + x'_k \hat{S}_{xx}^{-1}(X - \hat{X})/c_k].$$

La variance de ce total estimé pour cette variable z est identique à la variance proposée par Särndal, Swensson et Wretman (1989). Ces auteurs soutiennent que, en se basant sur la validité du modèle de régression, cette variance est préférable aux autres estimateurs par développement de Taylor pour le calcul de la variance. Nous constatons que le calcul de cette variable z se fait naturellement dans notre méthode.

2.2 Statistique du test de la somme des rangs de Wilcoxon

Nous démontrons maintenant comment notre méthode fonctionne dans un cas non standard plus difficile. Nous supposons que nos unités échantillonnées appartiennent à l'une des deux sous-populations, que nous nommons respectivement population 1 et population 2. Nous définissons:

$$I\{x \le y\} = \begin{cases} 1 \text{ si } x \le y, \\ 0 \text{ autrement,} \end{cases} \quad \text{et} \quad \delta_k = \begin{cases} 1 \text{ si } k \in \text{Pop. 1} \\ 0 \text{ autrement.} \end{cases}$$

Nous posons:

$$\hat{N}_1(t) = \sum_{k \in s} w_k \delta_k I\{x_k \le t\},$$

ce qui correspond au nombre estimé d'unités de la population 1 qui ont des valeurs inférieures ou égales à t. Nous définissons $\hat{N}_2(t)$ de façon analogue. Notons que $\hat{N}_j = \hat{N}_j(\infty)$, le nombre estimé d'unités dans la population j. Une version pondérée de la statistique du test de Wilcoxon est:

$$\hat{T}_W = \int_0^\infty [\hat{N}_1(t) + \hat{N}_2(t)]d\hat{N}_1(t). \qquad (13)$$

Cette expression correspond à la somme pondérée des rangs de la population 1 parmi les rangs pondérés de l'échantillon combiné. Afin de calculer la valeur probable asymptotique de $\hat{T}_W$ dans (13), posons $N_i(t) = E[\hat{N}_i(t)]$ pour $i = 1, 2$, et substituons $N_i(t)$ pour $\hat{N}_i(t)$ dans (13). Nous définissons alors $F_i(t) = N_i(t)/N_i$, où $N_i = E(\hat{N}_i)$ et nous appliquons l'hypothèse nulle lorsque, disons, $F_1(t) = F_2(t) = F(t)$. Nous obtenons ainsi la valeur probable asymptotique suivante:

$$\int_0^1 (N_1 + N_2)F(t)N_1 dF(t) = N_1(N_1 + N_2)/2.$$

On notera que dans le cas des échantillons indépendants de taille N_1 et N_2 provenant des populations 1 et 2, respectivement, et en supposant que la fonction de distribution dans chaque population est continue et que les échantillons sont prélevés par échantillonnage aléatoire simple, la valeur probable exacte de $\hat{T}_W$ dans (13) est $N_1(N_1 + N_2 + 1)/2$.

Considérons maintenant la statistique

$$\hat{T}^*_W = \int_0^\infty [\hat{N}_1(t) + \hat{N}_2(t)]d\hat{N}_1(t) - \frac{\hat{N}_1(\hat{N}_1 + \hat{N}_2)}{2}.$$

Nous utilisons Δ plutôt que d pour dénoter la différentielle totale, car d est utilisé sous l'intégrale. Nous avons donc

$$(\Delta\hat{T}^*_W) = \int_0^\infty [\Delta\hat{N}_1(t) + \Delta\hat{N}_2(t)]d\hat{N}_1(t)$$
$$+ \int_0^\infty [\hat{N}_1(t) + \hat{N}_2(t)]d\Delta\hat{N}_1(t)$$
$$- \frac{(\Delta\hat{N}_1)(\hat{N}_1 + \hat{N}_2) + \hat{N}_1(\Delta\hat{N}_1 + \Delta\hat{N}_2)}{2}.$$

En poursuivant selon notre méthode habituelle, nous obtenons:

$$\hat{T}^*_W - T^*_W \doteq \int_0^\infty \left(\sum w_k I\{x_k \le t\}\right) d\hat{N}_1(t)$$
$$+ \sum w_k \delta_k [\hat{N}_1(x_k) + \hat{N}_2(x_k)]$$
$$- \frac{\sum w_k \delta_k (\hat{N}_1 + \hat{N}_2) + \hat{N}_1 \sum w_k}{2} + \ldots,$$

de sorte que

$$z_k = \sum_j w_j \delta_j I\{x_k \le x_j\} + \delta_k[\hat{N}_1(x_k) + \hat{N}_2(x_k)]$$
$$- \frac{\delta_k(\hat{N}_1 + \hat{N}_2) + \hat{N}_1}{2}. \qquad (14)$$

Nous ne savons pas si ce résultat a déjà été publié. On peut démontrer que lorsque l'hypothèse nulle est vraie et que nous choisissons indépendamment des données dans deux populations par échantillonnage aléatoire simple, et lorsque les populations présentent des fonctions de distribution continues, la variance obtenue à partir des variables z dans (14) est asymptotiquement équivalente à celle obtenue avec la formule classique habituelle.

3. ÉCHANTILLONS À DEUX PHASES

La méthode décrite ci-dessus s'étend assez facilement au cas des échantillons à deux phases. Par exemple, examinons l'estimateur par quotient à deux phases d'une population totale, donné par:

$$\hat{Y}_{R(2)} = \frac{\hat{Y}}{\hat{X}} \hat{X}^{(1)} = \hat{R}\hat{X}^{(1)}, \qquad (15)$$

où $\hat{X}^{(1)} = \sum w_k x_k$ est l'estimation de la première phase de X basée sur les poids de la première phase $\{w_k\}$, et $\hat{Y}$ et $\hat{X}$ sont les estimations de Y et X, respectivement, basées sur les unités d'échantillonnage de la deuxième phase avec les poids $\{w_k w_{2k}\}$, où w_{2k} est le poids assigné à l'unité de la deuxième phase choisie et est soumise à la condition de se trouver dans l'échantillon de la première phase. En particulier, posons:

$$a_k = \begin{cases} 1 & \text{si la } k\text{-ième unité est dans l'échantillon de la seconde phase} \\ 0 & \text{autrement,} \end{cases}$$

nous obtenons,

$$\hat{Y} = \sum_{k \in s} w_k w_{2k} a_k y_k,$$

où s est l'ensemble des indices correspondant aux unités de l'échantillon de la première phase.

En prenant les différentielles totales de (15), nous obtenons:

$$(d\hat{Y}_{R(2)}) = \left(\frac{\hat{X}^{(1)}}{\hat{X}}\right)[(d\hat{Y}) - \hat{R}(d\hat{X})] + \hat{R}(d\hat{X}^{(1)}).$$

Nous remplaçons maintenant les différentielles totales par les sommes pondérées sur les unités de la première phase:

$$\hat{Y}_{R(2)} - \hat{Y} \doteq$$

$$\sum_{k \in s} w_k \left[a_k w_{2k} \left(\frac{\hat{X}^{(1)}}{\hat{X}}\right)(y_k - \hat{R}x_k) + \hat{R}x_k \right] + \ldots,$$

de sorte que

$$z_k = a_k w_{2k} \left(\frac{\hat{X}^{(1)}}{\hat{X}}\right)(y_k - \hat{R}x_k) + \hat{R}x_k. \qquad (16)$$

Nous voyons que les étapes que nous avons suivies sont essentiellement les mêmes que dans le cas d'un échantillon à une phase. Toutefois, il importe de souligner que z_k contient maintenant la variable aléatoire a_k, que l'on utilise pour indiquer si l'unité échantillonnée se trouve ou non dans l'échantillon de la deuxième phase. Cette variable est nécessaire pour calculer l'estimateur de la variance à deux phases.

Les variances obtenues pour la variable z dans (16) sont identiques à celles données par Rao et Sitter (1995), qui ont obtenu leur résultat par linéarisation de la méthode du jackknife.

L'extension de cette méthode aux autres problèmes d'estimation dans les échantillons à deux phases ne présente pas de problème. Supposons, par exemple, que $(\hat{Y}_1, \ldots, \hat{Y}_m)$ soient des estimations de $(Y_1, \ldots, Y_m)$ pour les échantillons de la deuxième phase, et que $(\hat{X}_1^{(1)}, \ldots, \hat{X}_p^{(1)})$ soient des estimations des variables disponibles seulement pour les unités de l'échantillon de la première phase. Nous supposons qu'un ensemble de paramètres additionnels, λ, soit défini seulement en termes des unités dans la deuxième phase, et que la variable qui nous intéresse soit définie en termes de ces paramètres additionnels et des valeurs $\hat{X}_j^{(1)}$. Nous obtenons donc l'expression formelle suivante:

$$U(\hat{\lambda}, \hat{Y}) = 0,$$

et

$$\hat{T} = g(\hat{X}^{(1)}, \hat{\lambda}).$$

En prenant les différentielles totales, nous obtenons, comme dans l'expression (9):

$$(d\hat{\lambda}) = -\left[\frac{\partial \hat{U}}{\partial \hat{\lambda}}\right]^{-1}\left[\frac{\partial \hat{U}}{\partial \hat{Y}}\right](d\hat{Y}),$$

de sorte que:

$$\hat{T} - T \doteq \left[\frac{\partial g}{\partial \hat{X}^{(1)}}\right]' \left(\sum_k w_k x_k - X\right)$$

$$- \left[\frac{\partial g}{\partial \hat{\lambda}}\right]' \left[\frac{\partial \hat{U}}{\partial \hat{\lambda}}\right]^{-1} \left[\frac{\partial \hat{U}}{\partial \hat{Y}}\right] \left(\sum_k a_k w_k w_{2k} y_k - Y\right).$$

Ainsi, l'expression générale pour z_k est:

$$z_k = \left[\frac{\partial g}{\partial \hat{X}^{(1)}}\right]' x_k - \left[\frac{\partial g}{\partial \hat{\lambda}}\right]' \left[\frac{\partial \hat{U}}{\partial \hat{\lambda}}\right]^{-1} \left[\frac{\partial \hat{U}}{\partial \hat{Y}}\right] a_k w_{2k} y_k.$$

Il est alors nécessaire de placer la variable z dans l'algorithme qui calcule la variance de l'estimateur d'un total provenant d'un échantillon à deux phases.

4. JUSTIFICATION

La technique que nous venons de décrire peut être considérée comme un résultat direct de la formulation donnée par Binder et Patak (1994). Nous résumerons ici l'un des principaux résultats de cette communication. Supposons que nous soyons intéressés par le paramètre θ, défini comme étant la solution à:

$$\hat{U}_1(\theta,\hat{\lambda}_\theta) = \sum_{k\in s} w_k u_1(y_k,\theta,\hat{\lambda}_\theta) = 0,$$

où $\hat{\lambda}_\theta$ est l'estimation d'un paramètre additionnel, défini comme étant la solution à:

$$\hat{U}_2(\theta,\hat{\lambda}_\theta) = \sum_{k\in s} w_k u_2(y_k,\theta,\hat{\lambda}_\theta) = 0,$$

pour une valeur θ donnée. Par un argument basé sur l'élimination des paramètres additionnels pour les problèmes de test d'hypothèse sur θ, Binder et Patak recommandent de baser les inférences au sujet de θ sur la variable

$$u^* = u_1(y,\theta,\hat{\lambda}_\theta) - \left[\frac{\partial \hat{U}_1}{\partial \hat{\lambda}_\theta}\right]\left[\frac{\partial \hat{U}_2}{\partial \hat{\lambda}_\theta}\right]^{-1} u_2(y,\theta,\hat{\lambda}_\theta). \quad (17)$$

En particulier, les intervalles de confiance bilatéraux pour θ devraient être basés sur:

$$\left\{\theta \;\middle|\; \frac{\hat{U}_1^2(\theta,\hat{\lambda}_\theta)}{\hat{W}} \leq \chi^2_{1-\alpha}(1)\right\},$$

où $\hat{W}$ est la variance estimée de l'estimateur d'un total lorsque la variable estimée est u^*.

Posons $u_1 = g(\lambda_1,\lambda_2) - \theta$. Le noyau des équations d'estimation pour les totaux sur y sera donné par $u_{21} = y - \lambda_1$ et le noyau des équations d'estimation pour λ_2 est donné par $u_{22}(\lambda_1,\lambda_2)$. Posons

$$\hat{U}_2 = \sum w_k \begin{bmatrix} u_{21} \\ u_{22} \end{bmatrix} = \begin{bmatrix} \hat{Y} - \hat{N}\hat{\lambda}_1 \\ \hat{N}u_{22} \end{bmatrix}, \quad \text{où} \quad \hat{N} = \sum w_k.$$

Après quelques manipulations algébriques, nous obtenons, de (17), la variance qui nous intéresse, soit la variance du total estimé basé sur la variable u^*, donnée par:

$$\left[\frac{\partial g(\hat{\lambda}_1,\hat{\lambda}_2)}{\partial \hat{\lambda}_1}\right]' y$$

$$- \left[\frac{\partial g(\hat{\lambda}_1,\hat{\lambda}_2)}{\partial \hat{\lambda}_2}\right]' \left[\frac{\partial u_{22}(\hat{\lambda}_1,\hat{\lambda}_2)}{\partial \hat{\lambda}_2}\right]^{-1} \left[\frac{\partial u_{22}(\hat{\lambda}_1,\hat{\lambda}_2)}{\partial \hat{\lambda}_1}\right] y$$

+ des termes constant.

Ceci est équivalent à l'expression (10), ce qui démontre que les méthodes démontrées ici sont cohérentes avec celles de Binder et Patak (1994).

REMERCIEMENTS

Je désire remercier Georgia Roberts et Alain Théberge, dont les suggestions nombreuses et utiles ont permis d'améliorer la lisibilité d'une ébauche antérieure. Je remercie également J.N.K. Rao pour les discussions nombreuses et utiles sur ce sujet, ainsi qu'un arbitre anonyme pour ses remarques constructives.

BIBLIOGRAPHIE

BINDER, D.A., et PATAK, Z. (1994). Use of estimating functions for interval estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1043.

KOTT, P.S. (1990). Estimating the conditional variance of a design consistent regression estimator. *Journal of Statistical Planning and Inference*, 24, 287-296.

RAO, J.N.K. (1995). Communication privée.

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

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.

Techniques d'enquête, juin 1996
Vol. 22, n° 1, pp. 23–31
Statistique Canada

Linéarisation des estimateurs de variance jackknife dans un échantillonnage stratifié à degrés multiples

W. YUNG et J.N.K. RAO[1]

RÉSUMÉ

Les auteurs examinent l'estimation de la variance d'une totalisation issue d'un échantillonnage stratifié à degrés multiples pour l'estimateur de stratification a posteriori et l'estimateur de régression généralisée. En linéarisant l'estimateur de variance jackknife, on obtient un nouvel estimateur, différent de celui obtenu par la méthode de linéarisation ordinaire. En matière de calcul, cet estimateur est plus simple à utiliser que l'estimateur de variance jackknife. Pourtant, il donne des valeurs qui s'approchent de celles de la méthode du jackknife. Les auteurs étudient les propriétés de l'estimateur de variance jackknife linéarisé, de l'estimateur de variance linéarisé ordinaire et de l'estimateur de variance jackknife dans le cadre d'une simulation. D'après l'écart entre le total estimatif des variables auxiliaires et les totaux connus de la population, les trois estimateurs donnent de bons résultats, conditionnellement ou non. Un estimateur de variance jackknife reposant sur une nouvelle pondération incorrecte a donné de piètres résultats, signe qu'il est important de procéder de façon adéquate à une nouvelle pondération quand on recourt à la méthode du jackknife.

MOTS CLÉS: Estimateur de régression généralisée; estimateur de variance jackknife; estimateur de variance linéarisé; estimateur de stratification a posteriori.

1. INTRODUCTION

Les enquêtes-échantillons de grande envergure recourent souvent à des plans d'échantillonnage stratifiés à plusieurs degrés et un nombre important de strates, L, ainsi que relativement peu d'unités primaires d'échantillonnage (grappes), $n_h (\geq 2)$, dans chaque strate. On prélève certains éléments (unités finales d'échantillonnage) dans chaque grappe au moyen d'une méthode d'échantillonnage. Le nombre de degrés ou les méthodes d'échantillonnage utilisés après le premier degré de l'échantillonnage ne sont pas spécifiés, mais on suppose que le sous-échantillonnage à l'intérieur des grappes débouche sur une estimation non biaisée des valeurs totales des grappes, $Y_{hi}, i = 1, \ldots, n_h$; $h = 1, \ldots, L$.

Les spécifications du plan d'échantillonnage de l'enquête servent à établir les poids de base $w_{hik} (> 0)$ associés au (hik)-ième élément. Les poids de base w_{hik} doivent souvent être ajustés après stratification pour être cohérents avec les totaux connus des variables de stratification a posteriori. En présence d'un seul paramètre de stratification a posteriori, on ajuste les poids par ratio à la population connue (à savoir, chiffres selon l'âge et le sexe). Avec deux paramètres de stratification a posteriori ou plus et des chiffres de population marginaux connus, on peut étalonner les poids w_{hik} par régression généralisée (voir partie 4), comme on le fait dans l'Enquête sur la population active du Canada (EPA).

L'EPA recourt à la méthode du jackknife pour estimer la variance de l'estimateur de régression généralisée. Cette méthode exige des calculs importants à l'ordinateur, mais elle s'applique aisément aux statistiques généralement lisses, contrairement à la linéarisation. Par ailleurs, elle présente de bonnes propriétés conditionnelles. Par exemple, avec l'échantillonnage aléatoire simple et l'estimateur de ratio, Royall et Cumberland (1981) ont montré que l'estimateur de variance jackknife suit la variance conditionnelle, étant donné la moyenne de la variable auxiliaire x pour l'échantillon.

La présente communication a principalement pour but d'examiner l'estimation de la variance pour l'estimateur de stratification a posteriori ajusté par quotient et pour l'estimateur de régression généralisée, avec un échantillonnage stratifié. En linéarisant l'estimateur de variance jackknife, on obtient un nouvel estimateur, différent de l'estimateur de variance linéarisé ordinaire. Dans le cas de l'estimateur de stratification a posteriori, ce nouvel estimateur est identique à l'estimateur de variance de Rao (1985). Le nouvel estimateur de variance proposé est plus simple à calculer que l'estimateur de variance jackknife, mais donne des résultats similaires.

La partie 2 présente l'estimateur de variance jackknife pour l'estimateur de base d'un total, Y. Les estimateurs de variance jackknife et jackknife linéarisé pour l'estimateur de stratification a posteriori sont présentés à la partie 3. Les résultats sont étendus à l'estimateur de régression généralisée pour les variables de stratification multiple a posteriori à la partie 4. La partie 5 traite de l'estimation de la variance pour un ratio de deux totaux, obtenu chacun au moyen de l'estimateur de régression généralisée. Les résultats d'une simulation démontrant l'efficacité relative de l'estimateur de variance linéarisé ordinaire, de la méthode du jackknife et de l'estimateur de variance jackknife linéarisé apparaissent à la partie 6.

[1] W. Yung, Statistique Canada, Division des méthodes d'enquêtes-ménage, immeuble R.H. Coats, parc Tunney, Ottawa, Ontario, K1A 0T6; et J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6.

2. ESTIMATEUR DE BASE

Partant des poids de base w_{hik}, l'estimateur non biaisé de la population totale Y s'écrit

$$\hat{Y} = \sum_{(hik)\in s} w_{hik} y_{hik}, \tag{2.1}$$

où s représente l'échantillon d'éléments et y_{hik} correspond à la valeur de la caractéristique à laquelle on s'intéresse, associée à l'élément $(hik)\in s$ de l'échantillon. Pour plus de simplicité, on suppose qu'il n'y a que des répondants.

La pratique courante consiste à prélever des grappes sans remise. Néanmoins, lorsqu'on doit estimer la variance, on simplifie considérablement les calculs en traitant l'échantillon comme si les grappes avaient été choisies avec remise. Pareille approximation entraîne habituellement une surestimation de la variance de $\hat{Y}$, mais le biais relatif sera probablement faible si les taux d'échantillonnage au premier degré le sont également.

L'estimateur de la variance de $\hat{Y}$ s'exprime sous la forme suivante:

$$v(\hat{Y}) = \sum_{h=1}^{L} \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 = v(y_{hi}), \tag{2.2}$$

où $y_{hi} = \sum_k (n_h w_{hik}) y_{hik}$ et $\bar{y}_h = (1/n_h)\sum_i y_{hi}$. La notation $v(y_{hi})$ de l'opérateur indique que $v(\hat{Y})$ ne dépend que de y_{hi}.

Pour introduire la méthode du jackknife, il faut utiliser l'estimateur $\hat{Y}_{(gj)}$ pour chaque (gj) de l'échantillon après avoir omis les données de la j-ième grappe prélevée dans la g-ième strate ($j = 1, \ldots, n_g$; $g = 1, \ldots, L$). Partant de (2.1), on y parvient simplement en supposant que $w_{gjk} = 0$, en remplaçant $w_{gik} (i \neq j)$ par $n_g w_{gik}/(n_g - 1)$ et en gardant les poids originaux w_{hik} pour $h \neq g$, c'est-à-dire:

$$w_{hik(gj)} = \begin{cases} 0 & \text{si } (hi) = (gj) \\ \dfrac{n_g}{(n_g-1)} w_{gik} & \text{si } h = g \text{ et } i \neq j \\ w_{hik} & \text{si } h \neq g. \end{cases}$$

Ces poids jackknife $w_{hik(gj)}$ sont établis pour chaque grappe (gj). L'estimateur résultant de Y est

$$\hat{Y}_{(gj)} = \sum_{(hik)\in s} w_{hik(gj)} y_{hik}.$$

L'estimateur de variance jackknife prend la forme suivante:

$$v_J(\hat{Y}) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{(gj)} - \hat{Y})^2. \tag{2.3}$$

On applique l'estimateur de variance (2.3) aux statistiques lisses générales, par exemple $\hat{\theta} = g(\hat{Y})$, en remplaçant simplement $\hat{Y}_{(gj)}$ et $\hat{Y}$ par $\hat{\theta}_{(gj)} = g(\hat{Y}_{(gj)})$ et $\hat{\theta}$, respectivement. Dans le cas linéaire $\hat{\theta} = \hat{Y}$, l'estimateur de variance jackknife est identique à l'estimateur de variance ordinaire (2.2).

3. ESTIMATEUR DE STRATIFICATION A POSTERIORI

Supposons que la population soit répartie a posteriori en C strates dont la population connue est ${}_cM$, $c = 1, \ldots, C$. Nous nous servirons de l'indice antérieur c pour désigner les strates a posteriori. Un estimateur de ${}_cM$ est donné par:

$${}_c\hat{M} = \sum_{(hik)\in {}_cs} w_{hik}, \tag{3.1}$$

où ${}_cs$ représente l'échantillon d'éléments appartenant à la c-ième strate a posteriori. De même, l'estimateur du total de la strate a posteriori ${}_cY$ est

$${}_c\hat{Y} = \sum_{(hik)\in {}_cs} w_{hik} y_{hik}.$$

En utilisant les estimateurs ${}_c\hat{Y}$ et ${}_c\hat{M}$, on obtient l'estimateur de stratification a posteriori du total Y:

$$\hat{Y}_{ps} = \sum_c \frac{{}_cM}{{}_c\hat{M}} {}_c\hat{Y}. \tag{3.2}$$

L'équation qui précède peut être récrite de la façon suivante:

$$\hat{Y}_{ps} = \sum_c \sum_{(hik)\in {}_cs} {}_cw_{hik} y_{hik}$$

où ${}_cw_{hik} = w_{hik}({}_cM/{}_c\hat{M})$ correspond au poids de $(hik)\in {}_cs$ ajusté par ratio. Si y_{hik} représente la variable indicatrice d'une strate a posteriori, par exemple c, alors $\hat{Y}_{ps} = {}_cM$, ce qui garantit la cohérence avec les totaux connus, ${}_cM$.

L'équation (2.2) donne l'estimateur de variance linéarisé ordinaire quand on remplace y_{hi} par

$$\tilde{e}_{hi} = \sum_c \sum_{k\in {}_cs} (n_h w_{hik})\, {}_ce_{hik},$$

où ${}_ce_{hik} = y_{hik} - {}_c\hat{Y}/{}_c\hat{M}$ pour le k-ième élément de la (hi)-ième grappe de ${}_cs$, c'est-à-dire:

$$v_L(\hat{Y}_{ps}) = v(\tilde{e}_{hi}). \tag{3.3}$$

Rao (1985) propose un autre estimateur de variance linéarisé reposant sur les poids ${}_cw_{hik}$ ajustés par ratio:

$$v_R(\hat{Y}_{ps}) = v(e^*_{hi}) \tag{3.4}$$

où

$$e_{hi}^{*} = \sum_{c} \sum_{k \in_{c} s} ({}_{h}n\, {}_{c}w_{hik})\, {}_{c}e_{hik}.$$

Avec la méthode du jackknife, il est nécessaire de recalculer les poids de stratification a posteriori ${}_{c}w_{hik}$ chaque fois qu'on supprime une grappe (gj). On se sert donc des poids jackknife $w_{hik(gj)}$ de (3.1) pour obtenir ${}_{c}\hat{M}_{(gj)}$, puis on utilise ${}_{c}w_{hik(gj)} = ({}_{c}M/{}_{c}\hat{M}_{(gj)})\, w_{hik(gj)}$ afin d'établir

$$\hat{Y}_{ps(gj)} = \sum_{c} \sum_{(hik) \in_{c} s} {}_{c}w_{hik(gj)}\, y_{hik}.$$

L'estimateur de variance jackknife devient donc:

$$v_J(\hat{Y}_{ps}) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{ps(gj)} - \hat{Y}_{ps})^2. \qquad (3.5)$$

En linéarisant (3.5), on parvient à un estimateur de variance jackknife linéarisé, $v_{JL}(\hat{Y}_{ps})$, identique à l'estimateur de variance de Rao (3.4); lire aussi Valliant (1993). Dans le cas particulier, non moins important, où $n_h = 2$ grappes par strate, les équations (3.4) et (3.5) sont asymptotiquement égales en tenant compte de termes d'ordre supérieur, lorsque le nombre de strates L augmente (Yung 1996).

Rao (1985) justifie (3.4) par des arguments heuristiques en soulignant que l'équation se réduit à un estimateur de variance valable pour un échantillonnage aléatoire simple, étant donné les tailles d'échantillon dans les strates a posteriori, à l'inverse de l'estimateur de variance linéarisé ordinaire (3.3). Särndal, Swensson et Wretman (1994) sont parvenus à un estimateur de variance similaire à (3.4) avec un échantillonnage à un degré dont le cadre était articulé sur un modèle. Puisque $v_{JL}(\hat{Y}_{ps})$ et $v_J(\hat{Y}_{ps})$ sont presque égaux, les résultats qui précèdent suggèrent deux estimateurs de variance «robustes», c'est-à-dire présentant de bonnes propriétés conditionnelles selon la taille estimée des strates a posteriori. Valliant (1993) a effectué une simulation pour éprouver la «robustesse» de $v_J(\hat{Y}_{ps})$ et $v_{JL}(\hat{Y}_{ps})$.

4. ESTIMATEUR DE RÉGRESSION GÉNÉRALISÉE

Dans la pratique, on crée couramment des strates a posteriori en fonction de deux variables auxiliaires ou davantage. Sachant le chiffre de population résultant pour chaque cellule, on peut se servir de l'estimateur de stratification a posteriori ajusté par ratio pour rendre les estimations plus précises. En réalité cependant, il se pourrait que le chiffre de population de la cellule soit inconnu. On pourrait, par exemple savoir le chiffre marginal de certains groupes d'âge et groupes ethniques, mais pas le chiffre des cellules combinant âge et race. Dans un tableau bidimensionnel, il s'ensuivrait donc qu'on connaîtrait les chiffres marginaux, mais pas les chiffres par cellule. Lorsque les paramètres de stratification a posteriori sont plus nombreux et qu'on connaît les chiffres de population marginaux, on peut se servir d'un estimateur de régression généralisée de Y en utilisant des variables auxiliaires indicatrices pour désigner les catégories des paramètres de stratification a posteriori (Huang et Fuller 1978; Deville et Särndal 1992).

Soit x_{hik}, un vecteur des variables auxiliaires dont la population totale connue est X. L'estimateur de régression généralisée de Y s'exprime comme suit

$$\hat{Y}_r = \hat{Y} + (X - \hat{X})^T \hat{B}, \qquad (4.1)$$

où

$$\hat{X} = \sum_{(hik) \in s} w_{hik} x_{hik},$$

et où $\hat{B}$ représente le vecteur des coefficients de régression estimatifs

$$\hat{B} = \hat{A}^{-1} \hat{b},$$

où

$$\hat{A} = \sum_{(hik) \in s} w_{hik} x_{hik} x_{hik}^T,$$

et

$$\hat{b} = \sum_{(hik) \in s} w_{hik} x_{hik} y_{hik}.$$

L'estimateur de stratification a posteriori $\hat{Y}_{ps}$ constitue un cas particulier de (4.1) où x_{hik} indique le vecteur des variables indicatrices des strates a posteriori. Dans ce cas, $\hat{X} = ({}_{1}\hat{M}, \ldots, {}_{C}\hat{M})^T$, $X = ({}_{1}M, \ldots, {}_{C}M)^T$ et $\hat{B} = ({}_{1}\hat{R}, \ldots, {}_{C}\hat{R})^T$, où ${}_{c}\hat{R} = {}_{c}\hat{Y}/{}_{c}\hat{M}$. Par conséquent,

$$\hat{Y}_r = \hat{Y} + \sum_{c} {}_{c}\hat{R}({}_{c}M - {}_{c}\hat{M}) = \hat{Y}_{ps}.$$

Avec deux paramètres de stratification a posteriori ou plus, X correspond à la valeur vectorielle des données de population marginales.

On peut récrire l'estimateur de régression généralisée de la façon suivante:

$$\hat{Y}_r = \sum_{(hik) \in s} w_{hik}^{*} y_{hik},$$

où

$$w_{hik}^{*} = w_{hik} a_{hik} \qquad (4.2)$$

correspond au poids «final» ou au poids «d'étalonnage» et

$$a_{hik} = 1 + x_{hik}^T \hat{A}^{-1} (X - \hat{X}).$$

Dans le cas particulier $\hat{Y}_{ps}$, $a_{hik} = {}_{c}M/{}_{c}\hat{M}$ pour $(hik) \in_{c} s$. En remplaçant $\hat{Y}_r$ dans l'opérateur par $\hat{Y}_r(y_{hik})$, on se rend facilement compte que l'estimateur de régression généralisée $\hat{X}_r = \hat{Y}_r(x_{hik}) = X$, ce qui garantit la cohérence avec les totaux connus X.

Quand on passe à l'estimation de la variance, l'équation (2.2) donne une fois de plus l'estimateur de variance linéarisé ordinaire où y_{hi} devient

$$\tilde{e}_{hi} = \sum_k (n_h w_{hik}) e_{hik},$$

où

$$e_{hik} = y_{hik} - \boldsymbol{x}_{hik}^T \hat{\boldsymbol{B}} \tag{4.3}$$

correspond aux résidus estimatifs, c'est-à-dire

$$v_L(\hat{Y}_r) = v(\tilde{e}_{hi}). \tag{4.4}$$

Pour la méthode du jackknife, il est nécessaire de calculer les poids d'étalonnage w^*_{hik} chaque fois qu'on supprime une grappe (gj). Ces poids sont donnés par

$$w^*_{hik(gj)} = w_{hik(gj)} a_{hik(gj)},$$

où

$$a_{hik(gj)} = 1 + \boldsymbol{x}_{hik}^T \hat{\boldsymbol{A}}^{-1}_{(gj)} (\boldsymbol{X} - \hat{\boldsymbol{X}}_{(gj)}),$$

$$\hat{\boldsymbol{A}}_{(gj)} = \sum_{(hik)\in s} w_{hik(gj)} \boldsymbol{x}_{hik} \boldsymbol{x}_{hik}^T,$$

et

$$\hat{\boldsymbol{X}}_{(gj)} = \sum_{(hik)\in s} w_{hik(gj)} \boldsymbol{x}_{hik}.$$

L'estimateur de régression généralisée résultant se présente de la façon suivante:

$$\hat{Y}_{r(gj)} = \sum_{(hik)\in s} w^*_{hik(gj)} y_{hik}$$

$$= \hat{Y}_{(gj)} + (\boldsymbol{X} - \hat{\boldsymbol{X}}_{(gj)})^T \hat{\boldsymbol{B}}_{(gj)}$$

où $\hat{\boldsymbol{B}}_{(gj)}$ est le vecteur des coefficients de régression estimatifs lorsque la (gj)-ième grappe est supprimée:

$$\hat{\boldsymbol{B}}_{(gj)} = \hat{\boldsymbol{A}}^{-1}_{(gj)} \hat{\boldsymbol{b}}_{(gj)}$$

avec

$$\hat{\boldsymbol{b}}_{(gj)} = \sum_{(hik)\in s} w_{hik(gj)} \boldsymbol{x}_{hik} y_{hik}.$$

L'estimateur de variance jackknife de $\hat{Y}_r$ est donc représenté par

$$v_J(\hat{Y}_r) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{r(gj)} - \hat{Y}_r)^2. \tag{4.5}$$

On verra à l'annexe qu'en linéarisant l'estimateur de variance jackknife (4.5), on obtient

$$v_{JL}(\hat{Y}_r) = v(e^*_{hi}) \tag{4.6}$$

pour laquelle

$$e^*_{hi} = \sum_k (n_h w^*_{hik}) e_{hik}$$

où w^*_{hik} est défini en (4.2) et e_{hik}, en (4.3). Fait intéressant, l'estimateur de variance jackknife linéarisé (4.6) ressemble à l'estimateur de variance fondé sur un modèle que proposent Särndal, Swensson et Wretman (1989) dans le cadre d'un échantillonnage à un degré. Yung (1996) a établi l'équivalence asymptotique de $v_J(\hat{Y}_r)$ et $v_{JL}(\hat{Y}_r)$ aux ordres supérieurs dans le cas particulier mais important où on retrouve $n_h = 2$ grappes par strate. Soulignons que les résultats qui précèdent s'appliquent aussi aux variables auxiliaires générales $\boldsymbol{x}_{hik}$.

Binder (1996) propose une nouvelle méthode de linéarisation qui aboutit elle aussi à $v_{JL}(\hat{Y}_r)$. En vertu de cette méthode, on évalue les dérivées partielles par rapport aux valeurs estimatives $\hat{Y}$, $\hat{\boldsymbol{X}}$ et $\hat{\boldsymbol{B}}$, et non par rapport aux chiffres de population Y, $\boldsymbol{X}$ et $\boldsymbol{B}$ comme avec la méthode de linéarisation classique. Puisque v_J et v_{JL} sont convergents selon le plan d'échantillonnage (Yung 1996) et présentent de bonnes propriétés conditionnelles, nos résultats apportent une justification théorique à la méthode de type «recette» que propose Binder.

Le calcul de l'estimateur de variance jackknife suppose l'inversion de la matrice $\hat{\boldsymbol{A}}_{(gj)}$ pour chaque (gj). Toutefois, on peut obtenir la valeur approximative de cet estimateur en gardant l'inverse pour l'échantillon complet, $\hat{\boldsymbol{A}}^{-1}$, puis en utilisant les poids modifiés

$$\tilde{w}_{hik(gj)} = w_{hik(gj)} \tilde{a}_{hik(gj)}$$

où

$$\tilde{a}_{hik(gj)} = 1 + (w_{hik}/w_{hik(gj)}) \boldsymbol{x}'_{hik} \hat{\boldsymbol{A}}^{-1} (\boldsymbol{X} - \hat{\boldsymbol{X}}_{(gj)}).$$

L'estimateur résultant de Y, à la suppression de la (gj)-ième grappe, est donné par

$$\tilde{Y}_{r(gj)} = \sum_{(hik)\in s} \tilde{w}_{hik(gj)} y_{hik}$$

et l'estimateur de variance jackknife correspondant est

$$v_{J1}(\hat{Y}_r) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\tilde{Y}_{r(gj)} - \hat{Y}_r)^2. \tag{4.7}$$

Il est facile de voir que (4.7) est identique à l'estimateur de variance linéarisé ordinaire (4.4).

5. ESTIMATION D'UN RATIO

On a souvent besoin d'établir le ratio de deux totaux estimatifs. Dans le cadre d'une enquête sur les dépenses des familles, par exemple, on s'intéresse à la proportion du revenu consacré à l'habillement. Soit:

$$\hat{Y}_r = \hat{Y} + (X - \hat{X})^T \hat{B}_1$$

un estimateur de régression généralisée de la somme totale allouée à l'achat de vêtements, Y. De même, soit

$$\hat{Z}_r = \hat{Z} + (X - \hat{X})^T \hat{B}_2$$

un estimateur de régression généralisée du revenu total, Z. La proportion à laquelle on s'intéresse est exprimée par $\theta = Y/Z$, qu'on peut estimer par

$$\hat{\theta} = \hat{Y}_r/\hat{Z}_r.$$

L'estimateur de variance jackknife est représenté par

$$v_J(\hat{\theta}) = \sum_g \frac{n_g - 1}{n_g} \sum_j (\hat{\theta}_{(gj)} - \hat{\theta})^2 \qquad (5.1)$$

où

$$\hat{\theta}_{(gj)} = \hat{Y}_{r(gj)}/\hat{Z}_{r(gj)}.$$

Lorsqu'on linéarise l'estimateur (5.1), on en obtient un nouveau

$$v_{JL}(\hat{\theta}) = v(r_{hi}^{**}) \qquad (5.2)$$

où

$$r_{hi}^{**} = \frac{1}{\hat{Z}_r} \sum_k (n_h w_{hik}^*) e_{hik}^*$$

pour laquelle

$$e_{hik}^* = e_{hik} - \frac{\hat{Y}_r}{\hat{Z}_r} \tilde{e}_{hik},$$

et

$$e_{hik} = y_{hik} - x_{hik}^T \hat{B}_1, \quad \tilde{e}_{hik} = z_{hik} - x_{hik}^T \hat{B}_2.$$

La preuve de (5.2) est omise pour plus de simplicité.

6. SIMULATION

On a procédé à une simulation afin de vérifier les propriétés conditionnelles et inconditionnelles des estimateurs de variance pour un échantillon de population finie avec une et deux variables de stratification a posteriori. On a recouru pour cela à une population fixe finie, déjà envisagée par Valliant (1993), soit 10,841 personnes de l'Enquête sur la population courante (EPC) américaine de septembre 1988. La variable qui nous intéresse, y, correspond à la rémunération hebdomadaire de chaque sujet. La stratification a posteriori à une dimension reposait sur l'âge, la race et le sexe, alors que celle à deux dimensions portait sur cinq niveaux d'âge et deux niveaux de race (voir tableaux 1 et 2 pour plus de détails).

Tableau 1

Attribution des catégories âge/race/sexe aux strates a posteriori: stratification a posteriori unidimensionnelle

Âge	Autre race		Race noire	
	Homme	Femme	Homme	Femme
19 ans et moins	1	1	1	1
20-24	2	3	3	3
25-34	5	6	4	4
35-64	7	8	4	4
65 ans et plus	2	3	3	1

Note: Le numéro des cellules (1-8) correspond au numéro d'identification des strates.

Tableau 2

Attribution des catégories âge/race aux strates a posteriori: stratification a posteriori bidimensionnelle

Âge	Autre race	Race noire	
19 ans et moins	(1,1)	(1,2)	PS1(1)
20-24	(2,1)	(2,2)	PS1(2)
25-34	(3,1)	(3,2)	PS1(3)
35-64	(4,1)	(4,2)	PS1(4)
65 ans et plus	(5,1)	(5,2)	PS1(5)
	PS2(1)	PS2(2)	

Note: Le numéro en marge correspond au numéro d'identification des strates. Les cellules (i,j) indiquent la strate ($i = 1, \ldots, 5; j = 1, 2$).

La population utilisée pour la simulation comprenait 2,826 segments géographiques composés chacun d'environ quatre ménages voisins. On a créé 100 strates ($L = 100$) comptant à peu près le même nombre de ménages. Nous nous sommes servis d'un plan d'échantillonnage stratifié à deux degrés en prenant les segments comme grappes et les personnes comme unités de deuxième degré. On a sélectionné dans chaque strate $n_h = 2$ segments selon une probabilité proportionnelle au nombre de personnes dans chaque segment, et $m_{hi} = 4$ personnes ont été choisies par échantillonnage aléatoire simple sans remise quand le segment comprenait plus de quatre personnes. Dans les segments de quatre personnes ou moins, tous les sujets ont été retenus. Grâce à ce plan d'échantillonnage, on a obtenu deux ensembles de 10,000 échantillons indépendants, un pour la stratification a posteriori unidimensionnelle, l'autre pour la stratification a posteriori bimensionnelle.

Nous avons calculé l'estimateur de base, l'estimateur de stratification a posteriori pertinent, $\hat{Y}_{ps}$ ou $\hat{Y}_r$, et quatre estimateurs de variance pour chaque échantillon: l'estimateur de variance linéarisé ordinaire v_L, l'estimateur de variance jackknife linéarisé v_{JL}, l'estimateur de variance jackknife v_J et un estimateur de variance jackknife incorrect v_J^*. Quand on applique la méthode du jackknife, une question se pose: faut-il ou non calculer à nouveau les poids «finals» ou «étalonnés» chaque fois qu'une grappe est supprimée? Le bon estimateur de variance jackknife donne effectivement un nouveau poids «final» chaque

fois qu'une grappe est supprimée, contrairement à l'estimateur de variance jackknife incorrect. Dans la stratification a posteriori unidimensionnelle, $v_J^*(\hat{Y}_{ps})$ applique l'ajustement intégral ${}_cM/{}_c\hat{M}$ au lieu de ${}_cM/{}_c\hat{M}_{(gj)}$, lorsque la (gj)-ième grappe est supprimée. Bref, $\hat{Y}_{ps(gj)}$ utilise $({}_cM/{}_c\hat{M})w_{hik(gj)}$ comme poids plutôt que $({}_cM/{}_c\hat{M}_{(gj)})$ $w_{hik(gj)}$. Pareillement, dans la stratification a posteriori bidimensionnelle, $v_J^*(\hat{Y}_r)$ fait appel à l'ajustement complet a_{hik} plutôt qu'à $a_{hik(gj)}$, quand la (gj)-ième grappe disparaît, c'est-à-dire que $\hat{Y}_r$ utilise les poids $w_{hik(gj)}a_{hik}$ plutôt que $w_{hik(gj)}a_{hik(gj)}$. La version linéarisée de v_J^* est identique à l'estimateur de variance v_R (équation 3.4), dans laquelle ${}_ce_{hik}$ est remplacé par y_{hik} pour $\hat{Y}_{ps}$, et à v_{JL} (équation 4.6) dans laquelle e_{hik} est remplacé par y_{hik} pour l'estimateur de régression généralisée $\hat{Y}_r$. En d'autres termes,

$$v_J^*(\hat{Y}_{ps}) = v(y_{hi}^*)$$

où

$$y_{hi}^* = \sum_c \sum_{k \epsilon_c s} (n_h \, {}_cw_{hik}) y_{hik}$$

et

$$v_J^*(\hat{Y}_r) = v(y_{hi}^*)$$

où

$$y_{hi}^* = \sum_{k \epsilon s} (n_h w_{hik}^*) y_{hik}.$$

Puisque v_J^* recourt à y plutôt qu'aux résidus e, v_J^* devrait manifestement surestimer la variance véritable de l'estimateur, même si le calcul s'avère plus simple que pour v_J.

(i) Résultats inconditionnels

Pour comparer l'efficacité inconditionnelle des estimateurs de variance, nous avons calculé le biais relatif empirique (RB) de chaque estimateur de variance. Le RB d'un estimateur de variance v est donné par:

$$\text{RB} = \frac{1}{\text{MSE}} \left[\frac{1}{10{,}000} \sum_i v_i \right] - 1$$

où v_i représente la valeur de v pour le i-ième échantillon simulé ($i = 1, \ldots, 10{,}000$) et MSE correspond à l'erreur quadratique moyenne (EQM) empirique de l'estimateur, par exemple $\tilde{Y}$:

$$\text{MSE} = \frac{1}{10{,}000} \sum_i (\tilde{Y}_i - Y)^2$$

où $\tilde{Y}_i$ est la valeur de $\tilde{Y}$ pour le i-ième échantillon simulé.

On a aussi calculé le taux d'erreur des intervalles de confiance théoriques normaux de Y, dans son ensemble, pour chaque estimateur de variance, au moyen d'un taux d'erreur nominal de 5%:

taux d'erreur =

$$1 - \frac{1}{10{,}000} \text{ (nombre d'échantillons où } L_i \leq Y \leq U_i),$$

où $L_i \leq Y \leq U_i$ représente l'intervalle de confiance de Y pour le i-ième échantillon simulé. Les taux d'erreur minimum et maximum sont:

taux d'erreur minimum =

$$\frac{1}{10{,}000} \text{ (nombre d'échantillons où } Y < L_i)$$

taux d'erreur maximum =

$$\frac{1}{10{,}000} \text{ (nombre d'échantillons où } Y > U_i).$$

La longueur moyenne de l'intervalle de confiance correspond à:

$$\text{longueur moyenne} = \frac{1}{10{,}000} \sum_i (U_i - L_i).$$

Le tableau 3 reproduit des résultats inconditionnels pour l'estimateur de stratification a posteriori $\hat{Y}_{ps}$ obtenus au moyen des mesures de rendement qui précèdent. En ce qui concerne le biais relatif, v_{JL} et v_J donnent de bons résultats quand RB < 1%, mais l'estimateur de variance jackknife incorrect v_{JL}^* surestime de façon appréciable l'EQM (RB = 37%). Notons que v_L donne aussi une bonne estimation de l'EQM de $\hat{Y}_{ps}$ de façon inconditionnelle (RB < 1%), contrairement à ce que prétend Valliant (1993). Ce dernier avait signalé un biais relatif de 35% pour v_L avec les mêmes données. Face à la convergence de v_L selon le plan d'échantillonnage, qu'appuient les résultats de la simulation pour v_L, on peut supposer que les calculs de Valliant pour v_L sont erronés.

Tableau 3

Résultats inconditionnels pour l'estimateur de stratification a posteriori

Mesure de rendement	$v_L(\hat{Y}_{ps})$	$v_{JL}(\hat{Y}_{ps})$	$v_J(\hat{Y}_{ps})$	$v_J^*(\hat{Y}_{ps})$
Biais relatif (%)	−0.44	0.12	0.26	37.16
Taux d'erreur (%)	5.20	5.09	5.06	2.41
Taux d'erreur minimum (%)	2.41	2.35	2.33	0.99
Taux d'erreur maximum (%)	2.79	2.74	2.73	1.42
Longueur moyenne	3.81	3.82	3.83	4.48

En ce qui concerne le rendement relatif à l'intervalle de confiance, le tableau 3 montre que le taux d'erreur associé à v_J, à v_{JL} et à v_L s'approche du taux nominal de 5%, tandis que le taux d'erreur de v_J^* est considérablement plus faible (environ 2.5%). On remarque une performance

analogue pour ce qui est des taux d'erreur minimum et maximum. Les estimateurs de variance v_J, v_{JL} et v_L donnent des résultats semblables pour la longueur moyenne de l'intervalle de confiance, mais la longueur moyenne de l'intervalle associé à v_J^* est significativement plus importante à cause du biais résultant de la surestimation. Enfin, signalons que les valeurs de rendement pour v_J et v_{JL} sont très proches, ce qui confirme l'équivalence asymptotique de v_J et v_{JL}.

Le tableau 4 présente les résultats inconditionnels pour l'estimateur de régression généralisée $\hat{Y}_r$. Ainsi qu'on a pu le remarquer avec $\hat{Y}_{ps}$, les estimateurs de variance v_J, v_{JL} et v_L donnent de bons résultats tant en ce qui concerne le biais relatif que le taux d'erreur de l'intervalle de confiance. L'estimateur jackknife incorrect v_J^*, par contre, débouche sur une sérieuse surestimation, qui se reflète dans un taux d'erreur inférieur au taux nominal et dans un intervalle de confiance de longueur moyenne supérieure.

Tableau 4

Résultats inconditionnels pour l'estimateur de régression généralisée

Mesure de rendement	$v_L(\hat{Y}_r)$	$v_{JL}(\hat{Y}_r)$	$v_J(\hat{Y}_r)$	$v_J^*(\hat{Y}_r)$
Biais relatif (%)	−0.96	0.76	0.57	25.87
Taux d'erreur (%)	5.30	5.27	5.23	3.07
Taux d'erreur minimum (%)	2.24	2.21	2.19	1.08
Taux d'erreur maximum (%)	3.06	3.06	3.04	1.99
Longueur moyenne	3.94	3.95	3.95	4.44

(ii) Résultats conditionnels

Nous nous sommes également attardés aux propriétés conditionnelles des estimateurs de variance, à l'instar de Valliant (1993). Pour l'estimateur de stratification a posteriori, les 10,000 échantillons simulés ont été répartis en 10 groupes comptant chacun 1,000 échantillons avec la mesure suivante (Valliant 1993):

$$D_{ps} = \sum_c \left(\frac{{}_c\hat{M}}{{}_cM} - 1 \right).$$

On a calculé D_{ps} pour chaque échantillon et les 10,000 échantillons ont été classés par ordre croissant selon cette valeur, puis divisés en groupes. On peut considérer D_{ps} comme une indication de l'«équilibre» de l'échantillon par rapport à la distribution de la population dans les strates a posteriori.

Pour l'estimateur de régression généralisée, on a retenu l'extension naturelle que voici de D_{ps}:

$$D_r = \sum_a \left(\frac{{}_a\hat{M}}{{}_aM} - 1 \right) + \sum_b \left(\frac{{}_b\hat{M}}{{}_bM} - 1 \right),$$

où a et b désignent le niveau des deux variables de stratification a posteriori, et $({}_a\hat{M}, {}_aM)$ et $({}_b\hat{M}, {}_bM)$ représentent les chiffres de population marginaux correspondants. La valeur D_r donne une idée de la façon dont l'échantillon est «équilibré» par rapport à la répartition de la population marginale dans les strates a posteriori.

Tableau 5

Biais relatif conditionnel (%) de l'estimateur de stratification a posteriori

Groupe	$v_L(\hat{Y}_{ps})$	$v_{JL}(\hat{Y}_{ps})$	$v_J(\hat{Y}_{ps})$	$v_J^*(\hat{Y}_{ps})$
1	−5.00	−8.05	−7.88	17.83
2	0.55	−1.18	−1.01	28.06
3	8.33	7.03	7.19	41.29
4	−1.10	−1.56	−1.42	31.82
5	−0.76	−0.69	−0.55	34.77
6	2.50	3.39	3.53	41.69
7	6.10	7.51	7.66	48.86
8	6.60	8.82	8.96	53.54
9	−4.46	−1.43	−1.31	41.11
10	−13.56	−9.17	−9.07	36.63

Tableau 6

Taux d'erreur conditionnel (%) de l'estimateur de stratification a posteriori

Groupe	$v_L(\hat{Y}_{ps})$	$v_{JL}(\hat{Y}_{ps})$	$v_J(\hat{Y}_{ps})$	$v_J^*(\hat{Y}_{ps})$
1	5.5	5.9	5.9	3.4
2	4.6	4.8	4.8	2.9
3	3.7	3.8	3.8	1.9
4	5.7	5.8	5.8	2.9
5	4.9	4.8	4.7	2.6
6	5.1	5.0	4.8	2.2
7	5.2	4.8	4.8	2.1
8	4.5	4.3	4.3	1.3
9	5.8	5.4	5.4	2.4
10	7.0	6.3	6.3	2.4

Les tableaux 5 et 6 présentent les résultats pour l'estimateur de stratification a posteriori: soit le biais relatif conditionnel au tableau 5 et le taux d'erreur conditionnel (taux nominal de 5%) au tableau 6. Ces mesures ont été obtenues de la même manière que pour l'essai inconditionnel, mais séparément pour chaque groupe. Les deux tableaux montrent clairement que v_J, v_{JL} et v_L donnent de bons résultats, bien que ceux de v_L laissent un peu plus à désirer pour les groupes 1 et 10 aux extrémités de la fourchette, v_J^* produit d'aussi piètres résultats que dans le cas précédent. Il est quelque peu surprenant de voir v_L donner de si bons résultats conditionnellement. Une explication plausible est qu'avec ce plan d'échantillonnage particulier, où $\hat{M} = \sum_{(hik)\in s} w_{hik} = M$,

$$\sum_c {}_c\hat{M} = \hat{M} = M.$$

Pour cette raison, il est impossible d'obtenir des échantillons mal équilibrés, puisque si ${}_c\hat{M}$ débouche sur une surestimation grossière avec certaines strates, les autres chiffres effectuent la correction à cause de la contrainte décrite plus haut. Par conséquent, on obtient des échantillons passablement bien équilibrés, auquel cas v_L devrait déboucher sur de bons résultats.

Tableau 7

Biais relatif conditionnel (%) de l'estimateur de régression généralisée

Groupe	$v_L(\hat{Y}_r)$	$v_{JL}(\hat{Y}_r)$	$v_J(\hat{Y}_r)$	$v_J^*(\hat{Y}_r)$
1	9.25	4.95	5.13	26.51
2	3.99	1.50	1.67	24.96
3	−3.24	−4.76	−4.59	17.53
4	−2.66	−3.43	−3.26	20.53
5	7.90	7.61	7.80	35.46
6	−3.60	−3.12	−2.94	23.38
7	−9.24	−8.27	−8.08	17.41
8	3.34	5.30	5.50	35.84
9	−3.75	−0.85	−0.62	30.84
10	−8.68	−4.15	−3.92	28.50

Tableau 8

Taux d'erreur conditionnel (%) de l'estimateur de régression généralisée

Groupe	$v_L(\hat{Y}_r)$	$v_{JL}(\hat{Y}_r)$	$v_J(\hat{Y}_r)$	$v_J^*(\hat{Y}_r)$
1	4.3	4.5	4.4	3.0
2	4.9	5.0	5.0	3.3
3	5.0	5.1	5.1	3.8
4	5.7	5.9	5.9	3.3
5	3.9	4.0	4.0	2.3
6	5.7	5.8	5.7	3.0
7	5.9	5.8	5.8	2.9
8	5.8	5.7	5.7	2.8
9	5.5	5.1	4.9	3.0
10	6.3	5.8	5.8	3.3

Les tableaux 7 et 8 reproduisent les résultats pour l'estimateur de régression généralisée: soit le biais relatif conditionnel au tableau 7 et le taux d'erreur conditionnel (taux nominal de 5%) au tableau 8. Ces résultats ressemblent beaucoup à ceux obtenus de la stratification unidimensionnelle. Dans les deux cas, on remarque une fois de plus la proximité des valeurs de rendement de v_J et v_{JL}, ce qui confirme l'équivalence asymptotique de v_J et v_{JL}.

En résumé, les trois estimateurs de variance v_J, v_{JL} et v_L produisent à peu près les mêmes résultats. L'estimateur jackknife incorrect v_J^*, par contre, laisse à désirer, signe qu'il faut effectuer une nouvelle pondération chaque fois qu'on supprime une grappe.

7. CONCLUSIONS

Beebakhee (1995) a appliqué les trois estimateurs de variance v_J, v_{JL} et v_L à diverses enquêtes-ménages entreprises par Statistique Canada. Ses résultats empiriques montrent que l'estimateur de variance jackknife linéarisé, v_{JL}, exige constamment moins de temps et d'argent que l'estimateur de variance jackknife v_J pour les enquêtes, tout en approximant très bien v_J. Il s'agit de résultats importants sur le plan pratique, car les utilisateurs souhaitent un estimateur plus simple dont les résultats s'approcheront au maximum des résultats de l'estimateur v_J actuel. L'estimateur de variance linéarisé ordinaire v_L fonctionne aussi bien que v_{JL} pour ce qui est du temps et de l'argent investis, mais comparativement à v_{JL}, les valeurs obtenues ne s'approchent pas autant de celles de v_J.

Si on s'intéresse surtout à l'estimation des totaux ou des ratios, on pourrait être tenté par l'estimateur de variance jackknife linéarisé v_{JL}, car il est plus facile à manipuler que l'estimateur de variance jackknife ordinaire v_J et produit des valeurs proches. Pour les statistiques lisses générales, v_{JL} a les mêmes inconvénients que l'estimateur de variance linéarisé ordinaire v_L, car ils exigent tous deux l'établissement d'une formule distincte pour chaque statistique, à l'inverse de v_J. Pour les propriétés statistiques, la simulation suggère que les trois estimateurs de variance, v_J, v_{JL} et v_L, se soldent par des résultats analogues. Par contre, l'estimateur jackknife incorrect v_J^*, qui utilise le même ajustement chaque fois qu'on supprime une grappe, donne de piètres résultats, signe qu'il faut procéder à une nouvelle pondération à chaque suppression.

REMERCIEMENTS

Ce travail n'aurait pu être mené à bien sans une subvention du Conseil de recherches en sciences naturelles et en génie du Canada.

ANNEXE

Preuve que $v_J(\hat{Y}_r) \approx v_{JL}(\hat{Y}_r)$

Pour obtenir le résultat désiré, il faut d'abord approximer la différence $\hat{A}_{(gj)}^{-1} - \hat{A}^{-1}$. Partant de l'identité matricielle

$$(I + PQ)^{-1} = I - P(I + QP)^{-1}Q$$

on obtient

$$\begin{aligned}\hat{A}_{(gj)}^{-1} - \hat{A}^{-1} &= \hat{A}^{-1}[I + (\hat{A}_{(gj)} - \hat{A})\hat{A}^{-1}]^{-1} - \hat{A}^{-1} \\ &= \hat{A}^{-1}[I - (\hat{A}_{(gj)} - \hat{A}) \\ &\quad (I + \hat{A}^{-1}(\hat{A}_{(gj)} - \hat{A}))^{-1}\hat{A}^{-1}] - \hat{A}^{-1} \\ &\approx -\hat{A}^{-1}(\hat{A}_{(gj)} - \hat{A})\hat{A}^{-1}. \qquad \text{(A.1)}\end{aligned}$$

On parvient à l'approximation (A.1) en notant (i) que $\hat{A}_{(gj)} - \hat{A}$ est d'un ordre inférieur à $\hat{A}$, selon l'hypothèse que la contribution d'aucune grappe n'est disproportionnée, à mesure que le nombre de strates L augmente (lire Yung (1996) pour plus de précisions sur les conditions de régularité) et (ii) que $[I + \hat{A}^{-1}(\hat{A}_{(gj)} - \hat{A})]^{-1} \approx I - \hat{A}^{-1}(\hat{A}_{(gj)} - \hat{A})$.

De (A.1), on obtient

$$\begin{aligned}\hat{B}_{(gj)} - \hat{B} &= (\hat{A}_{(gj)}^{-1} - \hat{A}^{-1} + \hat{A}^{-1})(\hat{b}_{(gj)} - \hat{b} + \hat{b}) - \hat{A}^{-1}\hat{b} \\ &\approx (\hat{A}_{(gj)}^{-1} - \hat{A}^{-1})\hat{b} + \hat{A}^{-1}(\hat{b}_{(gj)} - \hat{b}) \\ &\approx -\hat{A}^{-1}(\hat{A}_{(gj)} - \hat{A})\hat{B} + \hat{A}^{-1}(\hat{b}_{(gj)} - \hat{b}). \end{aligned} \quad \text{(A.2)}$$

Étant donné (A.2), il s'ensuit que

$$\begin{aligned}\hat{Y}_{r(gj)} - \hat{Y}_r &\approx (\hat{Y}_{(gj)} - \hat{Y}) - (\hat{X}_{(gj)} - \hat{X})^T\hat{B} - (\hat{X} - X)^T(\hat{B}_{(gj)} - \hat{B}) \\ &\approx \frac{1}{n_g - 1}(\bar{e}_g^* - e_{gj}^*), \end{aligned} \quad \text{(A.3)}$$

où $e_{gj}^* = \sum_k (n_g w_{gjk}^*) e_{gjk}$ et $\bar{e}_g^* = (1/n_g)\sum_j e_{gj}^*$. On s'est servi des résultats ci-dessous pour parvenir à (A.3):

$$(\hat{Y}_{(gj)} - \hat{Y}) - (\hat{X}_{(gj)} - \hat{X})^T\hat{B} = \frac{1}{n_g - 1}(\bar{e}_g - e_{gj})$$

et

$$(\hat{X} - X)^T(\hat{B}_{(gj)} - \hat{B}) \approx (X - \hat{X})^T\hat{A}^{-1}\left[\frac{1}{n_g - 1}(\bar{u}_g - u_{gj})\right],$$

où $e_{gj} = \sum_k (n_g w_{gjk}) e_{gjk}$ et $u_{gj} = \sum_k (n_g w_{gjk}) x_{gjk} e_{gjk}$.

Par conséquent, de (A.3), on déduit que:

$$\begin{aligned} v_J(\hat{Y}_r) &\approx \sum_{h=1}^{L} \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} (e_{hi}^* - \bar{e}_h^*)^2 \\ &= v(e_{hi}^*) = v_{JL}(\hat{Y}_r). \end{aligned}$$

BIBLIOGRAPHIE

BEEBAKHEE, R. (1995). Une comparaison des deux techniques d'estimation des variances: la méthode du jackknife et la méthode du jackknife linéarisée. Direction de la méthodologie, document de travail, DMEM-95-005F. Statistique Canada.

BINDER, D.A. (1996). Méthodes de linéarisation pour les échantillons à une et deux phases: une approche de type «recette». *Techniques d'enquête*, 22, 17-22.

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

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

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

RAO, J.N.K. (1985). Inférence conditionnelle dans les enquêtes par sondage. *Techniques d'enquête*, 11, 17-35.

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

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

STATISTIQUE CANADA (1990). *Méthodologie de l'enquête sur la population active du Canada*. N° 71-526 au catalogue.

VALLIANT, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.

YUNG, W. (1996). Contributions to poststratification in stratified multi-stage samples. Thèse de doctorat inédite, Carleton University, Ottawa, Canada.

Techniques d'enquête, juin 1996
Vol. 22, n° 1, pp. 33–41
Statistique Canada

Estimation des caractéristiques des petites régions au moyen d'un modèle gaussien inverse

Y.P. CHAUBEY, F. NEBEBE et P.S. CHEN[1]

RÉSUMÉ

Les auteurs examinent la méthodologie de l'analyse de la variance dans le cas d'une distribution gaussienne inverse et l'adaptent à l'estimation des paramètres régionaux de populations finies. Grâce à une étude de Monte Carlo, ils démontrent que le choix de ces estimateurs est défendable pour les données d'enquête étalées vers la droite, telles que celles sur le revenu ou sur le rendement d'un secteur particulier.

MOTS CLÉS: Interactions; gaussien inverse; Monte Carlo; estimations de régression; estimations synthétiques; estimateur de Särndal-Hidiroglou; modèle asymétrique.

1. INTRODUCTION

Un grand nombre de méthodes visant à résoudre la question des estimations pour des petites régions ont été publiées récemment, dont celles de Prasad et Rao (1990), Särndal et Hidiroglou (1989), Choudhry et Rao (1988), et Särndal (1984), ainsi que les études auxquelles se réfèrent ces auteurs, en particulier Särndal et Råbäck (1983), Fay et Herriot (1979), Shaible (1979), Holt, Smith et Tomberlin (1979), et Gonzalez et Hoza (1978), pour n'en nommer que quelques-unes. La nécessité d'obtenir des estimations régionales de plusieurs caractéristiques d'une population donnée a mené à l'élaboration de diverses méthodes précieuses qui permettent de produire des estimations réalistes et suffisamment exactes pour des secteurs restreints et d'autres sous-groupes particuliers. Plusieurs techniques proposées par les auteurs susmentionnés sont, implicitement et (ou) explicitement, fondées sur un modèle et sur l'application de la distribution théorique normale standard. D'autres chercheurs se sont attaqués au calcul d'estimateurs pour les secteurs restreints selon une approche bayesienne théorique ou empirique, en recherchant un compromis entre la moyenne d'échantillon d'une région (dont la population est présumée normale) et un estimateur calculé par régression en fonction d'une ou de plusieurs covariables (voir, p. ex., Stroud 1987, MacGibbon et Tomberlin 1989). Pour une revue détaillée des progrès récents en ce qui concerne les estimations régionales, le lecteur devrait consulter Ghosh et Rao (1994).

L'application de la distribution théorique normale standard à l'analyse des plans d'expérience factoriels est parfois inappropriée quand les données sont produites à partir d'une population dont la distribution présente une asymétrie positive nette. Bien que la plupart des procédures d'inférence soient retraçables sur le plan analytique, l'exactitude et la fiabilité des résultats sont douteuses dans le cas de nombreuses applications pratiques. Par conséquent, dans de telles situations, une analyse fondée sur des distributions à asymétrie positive se justifie.

La présente étude vise à examiner les méthodes d'inférence applicables à des plans d'expérience bifactoriels, tant asymétriques que symétriques, au moyen d'un modèle gaussien inverse permettant de produire des estimations pour les petites régions. Hidiroglou et Särndal (1985) ont décrit une étude de Monte Carlo effectuée en choisissant un estimateur de régression corrigé comme compromis entre l'estimateur synthétique et l'estimateur de régression généralisé. Ces auteurs (Särndal et Hidiroglou 1989) ont aussi présenté d'autres comparaisons d'estimateurs fondées sur l'inférence conditionnelle. Fondamentalement, l'estimateur de régression généralisé est calculé d'après le modèle de régression d'une superpopulation, sans hypothèse quant à la distribution. Chaubey (1991), quant à lui, a examiné les modèles de superpopulation proposés par Durbin (1959) dans le cas d'une distribution gamma ou gaussienne inverse auxiliaire, auquel cas l'estimateur de régression généralisé s'avère le meilleur prédicteur linéaire non biaisé (voir Prassad et Rao 1990). En fait, le meilleur estimateur linéaire non biaisé pour l'ensemble de la population étant indépendant de la forme de la distribution de la variable caractéristique, cette méthode est préférable, compte tenu de la difficulté éventuelle à calculer les estimations du maximum de vraisemblance (EMV). Ayant observé que les distributions des superpopulations (transposées aux populations), peuvent être fort semblables aux distributions gaussiennes inverses de diverses populations, nous aimerions tirer parti de cette caractéristique.

Le recours à une distribution gaussienne inverse n'est pas un exercice futile. Cette distribution, qui a été appliquée avec succès dans de nombreuses situations (voir Folks et Chhikara 1978), est fort similaire à la distribution gamma, à la distribution lognormal et à la distribution de Weibull utilisées couramment pour la modélisation de variables aléatoires non négatives dont la distribution présente une asymétrie positive. Dans le présent article, nous étudions l'application du modèle gaussien inverse aux estimations régionales. L'approche de Fries et de Bhattacharyya (1983), lesquels discutent de l'analyse de plans d'expérience

[1] Y.P. Chaubey, Professeur, Department of Mathematics and Statistics; F. Nebebe, Professeur agrégé, Department of Decision Sciences & M.I.S.; et P.S. Chen, assistant à la recherche, Department of Finance, Concordia University, Montréal, Canada.

bifactoriels au moyen d'un modèle gaussien inverse, est d'une très grande importance. Ces auteurs présentent le calcul d'estimations au moyen d'un modèle symétrique, sans interaction. Nous avons étendu cette approche au cas d'un modèle asymétrique, démarche essentielle à l'estimation des totaux ou des moyennes par domaine. À cet égard, on pourrait adapter la méthode générale de régression multiple de Bhattacharyya et Fries (1986), et Whitmore (1983), mais nous avons préféré la méthode directe. À la Section 2, nous précisons le modèle et présentons les estimateurs que nous proposons en vertu du modèle gaussien inverse. À la Section 3, nous effectuons une étude numérique en vue d'évaluer la performance de l'estimateur proposé grâce à une simulation de Monte Carlo. Enfin, à la Section 4, nous présentons un résumé et les conclusions.

2. LE MODÈLE DE RÉGRESSION GAUSSIEN INVERSE APPLIQUÉ AU CALCUL DES ESTIMATIONS RÉGIONALES

Supposons qu'on divise une population finie $\mathfrak{U}$ en D domaines qui ne se chevauchent pas $U_{d.}$, $d = 1(1)D$, où $N_{d.}$ représente la taille de $U_{d.}$. La population est en outre divisée suivant une deuxième dimension, en G groupes qui ne se chevauchent pas $U_{.g}$, $g = 1(1)G$, où la taille de $U_{.g}$ est représentée par $N_{.g}$. Le recoupement des domaines et des groupes produit DG cellules de population U_{dg}, $d = 1(1)D$, $g = 1(1)G$, où N_{dg} représente la taille de U_{dg}. Nous pouvons exprimer la taille de la population N par la relation $N = \sum_d N_{d.} = \sum_g N_{.g} = \sum_{dg} N_{dg}$. Nous désirons estimer les totaux par domaine $t_d = \sum_{U_{d.}} y_k$, où y représente la variable caractéristique et y_k, l'observation pour la k-ième unité. Un échantillon s de taille n est tiré de $\mathfrak{U}$ par simple échantillonnage aléatoire. Représentons par $s_{d.}$, $s_{.g}$ et s_{dg} les parties de s qui tombent dans $U_{d.}$, $U_{.g}$ et U_{dg}. Les tailles d'échantillon correspondantes sont représentées par $n_{d.}$, $n_{.g}$ et n_{dg}, et respectivement.

2.1 Méthode de régression pour les données gaussiennes inverses

Nous conseillons au lecteur de consulter les revues détaillées des travaux récents sur la distribution gaussienne inverse publiées par Chhikara et Folks (1989) et par Iyengar et Patwardhan (1988). La densité de probabilité d'une variable gaussienne inverse dont les paramètres sont (θ, σ), $IG(\theta, \sigma)$ s'écrit

$$f(y;\theta,\sigma) = (2\pi\sigma)^{-1/2}y^{-3/2}\exp[-(2\sigma y)^{-1}(y\theta^{-1} - 1)^2]; \tag{2.1}$$

où $y > 0, \theta > 0, \sigma > 0$. La moyenne et la variance de cette distribution sont θ et $\theta^3\sigma$, respectivement. Bhattacharyya et Fries (1982) ont proposé un modèle linéaire réciproque pour θ. Plus précisément, ils supposent que le modèle a la forme $\theta_k^{-1} = x_k'\eta$. Dans ces conditions, un estimateur de η similaire à l'estimateur du paramètre de régression du modèle linéaire habituel (voir Särndal 1984) s'exprime par

$$\hat{\eta} = \left(\sum_{k\in S_{d.}} \frac{x_k x_k' y_k}{\pi_k}\right)^{-1} \sum_{k\in S_{d.}} \frac{x_k}{\pi_k}. \tag{2.2}$$

Cet estimateur est appelé pseudo estimateur du maximum de vraisemblance, car, puisqu'on le calcule par maximisation inconditionnelle de la fonction de vraisemblance, l'inégalité $x_k'\hat{\eta} > 0$ n'est pas nécessairement satisfaite pour toutes les valeurs de k. Par conséquent, à l'exemple de l'estimateur de régression corrigé de Särndal (1984), nous pouvons représenter l'estimateur du total t_d du d-ième domaine par la relation

$$\hat{t}_{dIG} = \sum_{k\in U_{d.}} \hat{y}_k + \sum_{k\in S_{d.}} \frac{e_k}{\pi_k} \tag{2.3}$$

où $\hat{y}_k = x_k'\hat{\eta}$ et $e_k = y_k - \hat{y}_k$. Dans l'exposé qui suit, nous représentons la moyenne de la cellule (d,g) par θ_{dg}, et nous envisageons un échantillon aléatoire simple, auquel cas π_k est constant. En premier lieu, nous examinons la prédiction des observations au moyen de l'équation (2.3), en nous fondant sur un modèle à effets additifs représenté par

$$\theta_{dg}^{-1} = \mu + \alpha_d + \beta_g, \quad \sum \alpha_d = \sum \beta_g = 0, \tag{2.4}$$

où μ, les facteurs α_d et β_g représentent respectivement l'effet global, les effets par domaine ou par rangée, et les effets par groupe ou par colonne. Dans le cas d'une distribution gaussienne inverse, il faut aussi que $\theta_{dg} > 0$ pour toutes les cellules (d,g) et que $\sigma > 0$. Donc, les paramètres μ, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_D)$, $\beta = (\beta_1, \beta_2, \ldots, \beta_G)$, ainsi que σ sont compris dans l'ensemble $\Omega = \{(\mu, \alpha, \beta, \sigma): \sum_d \alpha_d = 0, \sum_g \beta_g = 0; \mu + \alpha_d + \beta_g > 0, \forall(d,g); \sigma > 0\}$. Dans ces conditions, on peut estimer les paramètres de prédiction par maximisation inconditionnelle de la fonction de vraisemblance. Compte tenu de la taille de la population et de celle des échantillons n_{dg}, et en se référant à (2.1) et à (2.3), la fonction logarithmique de vraisemblance des paramètres s'exprime par

$$\ell = -\frac{1}{2}\log\sigma \sum_d \sum_g n_{dg}$$

$$-(2\sigma)^{-1}\sum_d \sum_g \sum_k y_{dgk}^{-1}[y_{dgk}(\mu + \alpha_d + \beta_g) - 1]^2. \tag{2.5}$$

En premier lieu, observons que les paramètres sont effectivement donnés par $(\mu, \alpha_d, \beta_g, d = 1, 2, \ldots, D-1; g = 1, 2, \ldots, G-1)$. Donc, par différentiation de l'équation susmentionnée par rapport à $(\mu, \alpha_d, \beta_g, d = 1, 2, \ldots, D-1; g = 1, 2, \ldots, G-1)$ et en posant les dérivées partielles obtenues comme égales à zéro, on obtient les équations qui suivent pour les estimateurs $(\hat{\mu}, \hat{\alpha}_d, \hat{\beta}_g, d = 1, 2, \ldots, D-1; g = 1, 2, \ldots, G-1)$,

$$\hat{\mu} y_{..} + \sum_{d=1}^{D-1} \hat{\alpha}_d (y_{d.} - y_{D.}) + \sum_{g=1}^{G-1} \hat{\beta}_g (y_{.g} - y_G) = n_{..},$$

$$\hat{\mu}(y_{d.} - y_{D.}) + \hat{\alpha}_d y_{d.} + \sum_{j=1}^{D-1} \alpha_j y_{D.}$$

$$+ \sum_{g=1}^{G-1} \hat{\beta}_g \{ (y_{dg} - y_{Dg}) - (y_{dG} - y_{DG}) \} = n_{d.} - n_{D.},$$

$$\hat{\mu}(y_{.g} - y_{.G}) + \sum_{d=1}^{D-1} \hat{\alpha}_d \{ (y_{dg} - y_{dG}) - (y_{Dg} - y_{DG}) \}$$

$$+ \hat{\beta}_g y_{.g} + \sum_{j=1}^{G-1} \hat{\beta}_j y_{.G} = n_{.g} - n_{.G}, \quad (2.6)$$

où les totaux et les moyennes sont représentés par les expressions

$$y_{dg} = \sum_k y_{dgk}, \; y_{d.} = \sum_g y_{dg}, \; y_{.g} = \sum_d y_{dg}, \quad (2.7a)$$

$$n_{d.} = \sum_g n_{dg}, \; n_{.g} = \sum_d n_{dg}, \; n_{..} = \sum_d \sum_g n_{dg}. \quad (2.7b)$$

Les solutions $(\hat{\mu}, \hat{\alpha}_d, \hat{\beta}_g)$, $d = 1(1)D$, $g = 1(1)G$, qui fournissent le pseudo estimateur du maximum de vraisemblance, n'aboutissent pas nécessairement à des estimations de réponse non négatives, mais correspondent à l'estimateur du maximum de vraisemblance (EMV) correct quand $n_{dg} \to \infty$ (voir Fries et Bhattacharyya 1983), la probabilité étant égale à un. Nous pouvons donc tronquer les valeurs des estimateurs de réponse à zéro pour annuler les valeurs négatives.

Dans le cas du modèle $IG(\theta, \sigma)$ avec interactions, le paramétrage habituel des effets d'interaction donne à penser que le modèle prend la forme

$$\theta_{dg}^{-1} = \mu + \alpha_d + \beta_g + \gamma_{dg},$$

$$\sum \alpha_d = \sum \beta_g = \sum_d \gamma_{dg} = \sum_g \gamma_{dg} = 0, \quad (2.8)$$

où γ_{dg} représente maintenant l'effet d'interaction quand on considère le d-ième domaine et le g-ième groupe. Dans ces conditions, nous pouvons obtenir les estimateurs des paramètres en appliquant la méthode décrite ci-dessus. Toutefois, puisque l'estimateur du maximum de vraisemblance (EMV) de θ_{dg} est $\bar{y}_{dg}$, et qu'il existe une relation de parité entre les paramètres du modèle reparamétré en ce qui concerne $(\mu, \alpha_d, \beta_g, \gamma_{dg})$, d'une part, et les paramètres originaux, θ_{dg}, d'autre part, il n'est pas nécessaire d'établir des formules explicites pour l'EMV de divers paramètres. Donc, selon l'équation (2.3), pour un modèle bifactoriel avec interactions, notre estimateur est

$$\hat{t}_{dWI} = \sum_g N_{dg} \bar{y}_{dg}, \quad (2.9)$$

expression qui correspond à l'estimateur stratifié a posteriori et ne présente donc aucun intérêt supplémentaire en ce qui concerne l'estimation des caractéristiques des petites régions. Dans le cas du modèle avec interactions, l'estimateur s'exprime par

$$\hat{t}_{dWOI} = \sum_g N_{dg} \hat{\theta}_{dg} + \sum_g \hat{N}_{dg} (\bar{y}_{dg} - \hat{\theta}_{dg}), \quad (2.10)$$

où $\hat{\theta}_{dg}^{-1} = \hat{\mu} + \hat{\alpha}_d + \hat{\beta}_g$, les estimateurs étant calculés à partir de (2.6) et de $\hat{N}_{dg} = n_{dg} N / n_{..}$.

Afin d'évaluer l'efficacité de cet estimateur, nous avons effectué une étude numérique dont les résultats sont exposés à la section suivante.

3. ÉTUDE NUMÉRIQUE DE L'ESTIMATEUR DE RÉGRESSION GAUSSIEN INVERSE

Nous présentons ici les résultats d'une étude en simulation effectuée pour évaluer la performance des estimateurs développés à la section précédente. À titre de référence, nous utilisons l'estimateur de régression corrigé proposé par Särndal et Hidiroglou (1989) dont voici l'expression:

$$\hat{t}_{dS-H} = \sum_g N_{dg} \bar{y}_{.g} + \sum_g F_d \hat{N}_{dg} (\bar{y}_{dg} - \bar{y}_{.g}), \quad (3.1)$$

où $F_d = N_{d.} / \hat{N}_{d.}$, si $\hat{N}_{d.} \geq N_{d.}$, sinon, $F_d = \hat{N}_{d.} / N_{d.}$. Ici, $\hat{N}_{d.} = n_{d.} N / n_{..}$. On peut obtenir une autre expression de cet estimateur, qui tient compte à la fois des effets de groupe et de domaine, en remplaçant $\bar{y}_{.g}$ par $\bar{y}_{.g} + \bar{y}_{d.} - \bar{y}_{..}$ mais cette approche n'a pas été poursuivie ici. Il convient de noter qu'on ne peut calculer les estimateurs susmentionnés quand n_{dg} est égal à zéro. Le cas échéant, on choisit simplement comme estimateurs les moyennes d'échantillons des domaines respectifs. Aux fins de comparaison, nous incluons également la version modifiée de $\hat{t}_{dWOI}$ que voici:

$$\hat{t}_{dWOIM} = \sum_g N_{dg} \hat{\theta}_{dg} + \sum_g F_d \hat{N}_{dg} (\bar{y}_{dg} - \hat{\theta}_{dg}). \quad (3.2)$$

3.1 Conception de l'étude en simulation

Nous nous sommes servis des données sur le revenu des ménages au Canada en 1986, tirées du fichier de microdonnées sur le revenu des ménages et sur les équipements ménagers de Statistique Canada (1987), pour calculer la valeur des paramètres utilisés pour la simulation. Après avoir réparti les données sur le revenu des ménages en dix domaines, selon la province, et en six groupes, selon le niveau d'éducation, nous avons commencé par ajuster un modèle gaussien inverse représenté par l'équation (2.4). Puis, à partir des estimations des paramètres, nous avons calculé les paramètres réels de la superpopulation gaussienne inverse qui sont résumés à l'annexe A. Les valeurs de D,

G, N_{dg} sont tirées de cette population (voir l'annexe B), D représentant le nombre de provinces (c.-à-d. $D = 10$) et G représentant le nombre de groupes de niveau d'éducation (c.-à-d. $G = 6$). Nous avons obtenu d'autres ensembles de valeurs de θ_{dg} et de σ en considérant diverses combinaisons de (c_1, c_2); $c_1 = 0(1)4$ et de $c_2 = 1, .25, .1, .01$, où c_1 est utilisé pour transformer θ_{dg} en $10^{-c_1}\theta_{dg}$ et c_2, pour transformer σ en $c_2\sigma$. Il convient de noter que quand $c_1 = 0$ et $c_2 = 1$, on obtient les valeurs des paramètres de la population originale. En outre, les valeurs plus élevées de c_1 indiquent des valeurs plus faibles des moyennes et celles de c_2, une valeur plus élevée du paramètre de dispersion.

Pour effectuer l'étude en simulation, nous avons commencé par produire, pour un ensemble donné de valeurs de θ_{dg} et de σ, un échantillon aléatoire gaussien inverse au moyen de l'algorithme de Michael et coll. (1976), le nombre d'observations correspondant aux valeurs données à l'annexe B. Puis, nous nous sommes servis de cet échantillon aléatoire comme d'une population finie dont nous avons tiré 1,000 échantillons aléatoires pour des fractions d'échantillonnage correspondant à 1% et à 5% avec remplacement. Nous avons, en fait, sélectionné plusieurs échantillons aléatoires et obtenu des résultats comparables à ceux exposés ici. Pour chaque échantillon, nous avons calculé les estimateurs des totaux pour les 10 domaines au moyen des estimateurs $\hat{t}_{dS-H}$, $\hat{t}_{dWOI}$ et $\hat{t}_{dWOIM}$. Nous avons choisi comme critères d'évaluation de la performance des estimateurs l'erreur relative absolue moyenne (ERAM) et le biais relatif absolu (BRA) définis comme suit:

$$\text{ERAM}(\hat{t}_d) = \frac{1}{1000}\sum_{i=1}^{1000} |\hat{t}_{di} - t_d|/t_d \qquad (3.3)$$

$$\text{BRA}(\hat{t}_d) = \left|\frac{1}{1000}\sum_{i=1}^{1000} \hat{t}_{di} - t_d\right|/t_d. \qquad (3.4)$$

Tableau 1

Erreur relative absolue moyenne (%) de divers estimateurs

Domaine	Échantillon de 1%			Échantillon de 5%			Échantillon de 1%			Échantillon de 5%		
	SH	WOI	WOIM	SH	WOI	WOIM	SH	WOI	WOIM	SH	WOI	WOIM
	$c_1 = 0, c_2 = 1$						$c_1 = 0, c_2 = .01$					
1	13.27	13.05	13.19	6.60	6.48	6.47	3.72	2.46	2.45	1.80	0.89	0.89
2	14.57	13.61	14.20	7.53	7.61	7.69	3.79	3.56	3.48	2.10	0.59	0.60
3	25.27	27.86	26.88	19.07	20.74	20.80	2.52	1.51	1.52	1.19	0.77	0.77
4	11.83	11.70	11.74	5.29	5.61	5.59	1.83	1.08	1.09	0.93	0.58	0.58
5	10.57	11.72	11.68	6.80	7.10	7.11	0.92	0.90	0.91	0.42	0.40	0.40
6	7.12	7.45	7.52	3.85	3.95	3.97	1.94	1.22	1.22	0.93	0.64	0.64
7	11.78	13.91	14.23	7.39	8.01	8.05	1.22	1.13	1.14	0.86	0.64	0.64
8	11.48	12.56	12.46	6.70	7.15	7.14	1.29	0.93	0.94	0.76	0.67	0.68
9	7.43	7.92	7.99	3.61	3.74	3.75	3.47	2.99	2.96	3.13	2.97	2.96
10	15.32	17.43	17.16	11.20	11.81	11.80	0.93	0.94	0.95	0.52	0.52	0.53
	$c_1 = 2, c_2 = 1$						$c_1 = 2, c_2 = .01$					
1	3.34	2.18	2.15	1.66	0.79	0.78	2.99	1.48	1.44	1.47	0.08	0.08
2	4.14	3.94	3.82	2.14	1.07	1.06	0.54	3.37	3.27	1.86	0.14	0.13
3	2.44	1.67	1.65	1.17	0.71	0.70	1.81	0.45	0.44	0.87	0.07	0.07
4	2.05	1.70	1.69	0.98	0.70	0.70	1.32	0.36	0.35	0.66	0.07	0.07
5	1.08	1.17	1.16	0.50	0.51	0.51	0.27	0.13	0.13	0.11	0.05	0.05
6	1.74	1.14	1.14	0.78	0.52	0.52	1.29	0.13	0.13	0.55	0.05	0.05
7	1.90	1.57	1.56	0.91	0.72	0.72	1.22	0.31	0.31	0.56	0.07	0.07
8	1.48	1.38	1.38	0.70	0.60	0.60	0.81	0.18	0.18	0.38	0.06	0.06
9	1.41	1.30	1.29	0.67	0.59	0.58	0.69	0.14	0.14	0.30	0.06	0.06
10	1.22	1.38	1.38	0.56	0.59	0.59	0.26	0.15	0.15	0.10	0.06	0.06
	$c_1 = 4, c_2 = 1$						$c_1 = 4, c_2 = .01$					
1	2.99	1.48	1.44	1.47	0.08	0.08	2.99	1.45	1.41	1.47	0.01	0.01
2	3.54	3.37	3.27	1.86	0.14	0.13	3.54	3.36	3.25	1.87	0.05	0.05
3	1.81	0.45	0.44	0.87	0.07	0.07	1.80	0.38	0.37	0.86	0.01	0.01
4	1.32	0.36	0.35	0.66	0.07	0.07	1.31	0.28	0.27	0.66	0.01	0.01
5	0.27	0.13	0.13	0.11	0.05	0.05	0.24	0.06	0.06	0.10	0.01	0.01
6	1.29	0.13	0.13	0.55	0.05	0.05	1.29	0.06	0.06	0.54	0.01	0.01
7	1.22	0.31	0.31	0.56	0.07	0.07	1.20	0.24	0.24	0.55	0.01	0.01
8	0.81	0.18	0.18	0.38	0.06	0.06	0.79	0.09	0.09	0.37	0.01	0.01
9	0.69	0.14	0.14	0.30	0.06	0.06	0.68	0.06	0.06	0.29	0.01	0.01
10	0.26	0.15	0.15	0.10	0.06	0.06	0.23	0.07	0.07	0.09	0.01	0.01

Tableau 2

Biais relatif absolu (en %) de divers estimateurs

Domaine	Échantillon de 1%			Échantillon de 5%			Échantillon de 1%			Échantillon de 5%		
	SH	WOI	WOIM	SH	WOI	WOIM	SH	WOI	WOIM	SH	WOI	WOIM
	$c_1 = 0, c_2 = 1$						$c_1 = 0, c_2 = .01$					
1	4.34	2.40	2.51	1.87	0.26	0.27	2.66	1.58	1.54	1.22	0.03	0.03
2	8.88	3.46	4.39	2.18	0.30	0.23	3.15	3.40	3.31	1.38	0.04	0.04
3	3.13	3.47	2.74	0.51	1.12	1.15	1.44	0.31	0.32	0.68	0.01	0.01
4	1.57	0.51	0.53	0.50	0.21	0.22	1.11	0.29	0.30	0.53	0.03	0.03
5	0.13	0.33	0.35	0.20	0.16	0.18	0.10	0.03	0.02	0.05	0.01	0.01
6	1.09	0.14	0.04	0.02	0.39	0.42	1.09	0.03	0.03	0.43	0.02	0.01
7	1.20	1.09	1.59	0.54	0.28	0.30	0.99	0.22	0.23	0.43	0.01	0.01
8	0.40	0.04	0.12	0.20	0.53	0.54	0.55	0.00	0.01	0.28	0.03	0.03
9	1.03	0.47	0.36	0.24	0.04	0.01	1.01	0.35	0.37	0.45	0.14	0.14
10	1.05	2.27	2.03	0.04	0.30	0.29	0.08	0.02	0.01	0.06	0.01	0.01
	$c_1 = 2, c_2 = 1$						$c_1 = 2, c_2 = .01$					
1	2.40	1.37	1.33	1.13	0.01	0.01	2.47	1.43	1.39	1.15	0.01	0.01
2	3.00	3.28	3.16	1.33	0.02	0.01	3.06	3.34	3.24	1.36	0.03	0.03
3	1.53	0.39	0.38	0.70	0.04	0.04	1.46	0.35	0.34	0.65	0.01	0.01
4	1.00	0.25	0.25	0.53	0.04	0.04	1.01	0.23	0.23	0.49	0.00	0.00
5	0.10	0.02	0.03	0.04	0.00	0.01	0.10	0.01	0.02	0.04	0.00	0.00
6	1.16	0.01	0.01	0.47	0.02	0.02	1.15	0.01	0.00	0.46	0.00	0.00
7	1.00	0.27	0.27	0.42	0.00	0.00	0.95	0.21	0.21	0.41	0.00	0.00
8	0.48	0.04	0.04	0.25	0.01	0.01	0.57	0.04	0.04	0.26	0.00	0.00
9	0.64	0.06	0.05	0.27	0.02	0.02	0.61	0.01	0.00	0.26	0.00	0.00
10	0.01	0.02	0.02	0.02	0.00	0.00	0.06	0.01	0.01	0.03	0.00	0.00
	$c_1 = 4, c_2 = 1$						$c_1 = 4, c_2 = .01$					
1	2.47	1.43	1.39	1.15	0.01	0.01	2.48	1.43	1.39	1.15	0.00	0.00
2	3.06	3.34	3.24	1.36	0.03	0.03	3.07	3.35	3.24	1.36	0.04	0.04
3	1.46	0.35	0.34	0.65	0.01	0.01	1.45	0.34	0.34	0.64	0.00	0.00
4	1.01	0.23	0.23	0.49	0.00	0.00	1.01	0.24	0.24	0.49	0.00	0.00
5	0.10	0.01	0.02	0.04	0.00	0.00	0.11	0.01	0.02	0.04	0.00	0.00
6	1.15	0.01	0.00	0.46	0.00	0.00	1.15	0.01	0.00	0.46	0.00	0.00
7	0.95	0.21	0.21	0.41	0.00	0.00	0.94	0.20	0.20	0.41	0.00	0.00
8	0.57	0.04	0.04	0.26	0.00	0.00	0.58	0.04	0.05	0.26	0.00	0.00
9	0.61	0.01	0.00	0.26	0.00	0.00	0.60	0.00	0.00	0.25	0.00	0.00
10	0.06	0.01	0.01	0.03	0.00	0.00	0.06	0.01	0.01	0.03	0.00	0.00

Ici, $\hat{t}_d$ désigne un estimateur typique de t_d et $\hat{t}_{di}$ représente la valeur du i-ième échantillon de la simulation de Monte Carlo ($i = 1, \ldots, 1000$).

3.2 Analyse des résultats

Les valeurs de l'ERAM calculées conformément à (3.3) et celles du BRA calculées d'après (3.4) pour ces trois estimateurs et pour différentes tailles d'échantillon sont résumées aux tableaux 1 et 2, respectivement, pour certaines paires (c_1, c_2). Les valeurs choisies pour c_1 représentent des moyennes fortes (comme dans la population originale, $c_1 = 0$), des moyennes modérées ($c_1 = 2$), et des moyennes faibles ($c_1 = 4$), tandis que celles choisies pour c_2 représentent le paramètre de dispersion original ($c_2 = 1$) et une autre valeur, plus faible ($c_2 = .01$). Il est intéressant de noter que, si on augmente c_1 d'une unité tandis que c_2 demeure constant, le coefficient diminue d'un facteur 10.

Certaines valeurs de l'ERAM et du BRA des tableaux 1 et 2 sont représentées graphiquement aux figures 1 et 2, respectivement, pour les échantillons de 1%, aux fins de visualisation.

La comparaison des valeurs de l'ERAM et du BRA obtenues pour les trois estimateurs révèle une diminution du biais ainsi que de l'erreur relative dans de nombreux cas, tant pour les échantillons de 1% que de 5%. On constate que les valeurs de l'ERAM et du BRA diminuent parallèlement à la valeur de la moyenne et à celle du paramètre de dispersion σ. Les diminutions sont appréciables, en particulier dans le cas de l'échantillon de 5% et (ou) quand la moyenne est nulle. On remarquera aussi que la diminution du biais est généralement plus importante que celle de l'erreur. Soulignons également que, selon Johnson et Kotz (1970, pp. 141), pour une valeur constante de la moyenne, la distribution gaussienne inverse standardisée tend à devenir normale à mesure que le coefficient de variation tend vers zéro. Puisque l'augmentation de la valeur de l'ERAM et de celle du BRA est plus importante quand la valeur du coefficient de variation

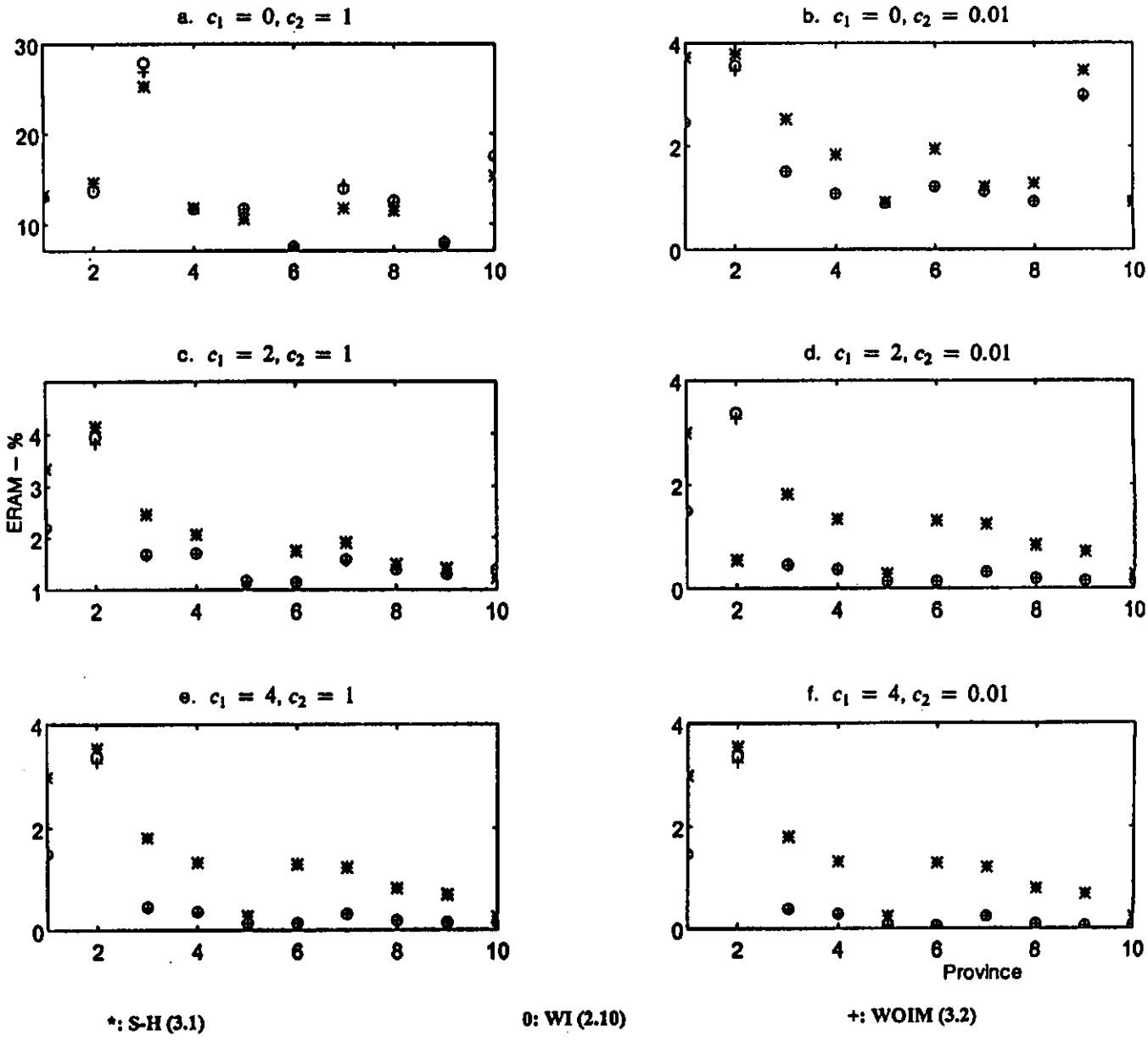

Figure 1. Erreur relative absolue moyenne de divers estimateurs pour un échantillon de 1%.

est faible, nous concluons que, pour le calcul d'estimations par modélisation, il est important de bien choisir le modèle de la moyenne quand le coefficient de variation est petit.

Nous observons en outre que l'ERAM et le BRA des estimateurs $\hat{t}_{dWOI}$ et $\hat{t}_{dWOIM}$ sont presque identiques, donc que la modification de l'estimateur représentée par (2.10) est inutile. Soulignons, en revanche, que Hidiroglou et Särndal (1985) ont démontré que l'estimateur $\hat{t}_{dS-H}$ est nettement supérieur à l'estimateur non corrigé correspondant proposé par Särndal (1984).

Certains ayant critiqué le fait que $\hat{t}_{dWOI}$ et $\hat{t}_{dWOIM}$ dépendent du modèle, nous avancerons les arguments qui suivent pour défendre le choix de ces estimateurs. La distribution gaussienne inverse peut prendre diverses formes et permet d'approcher la distribution lognormale, la distribution gamma, la distribution de Weibull et d'autres distributions à asymétrie positive. Donc, si l'on pense que la distribution de la caractéristique principale présente une asymétrie positive, la méthodologie que nous venons d'exposer est valable et utile.

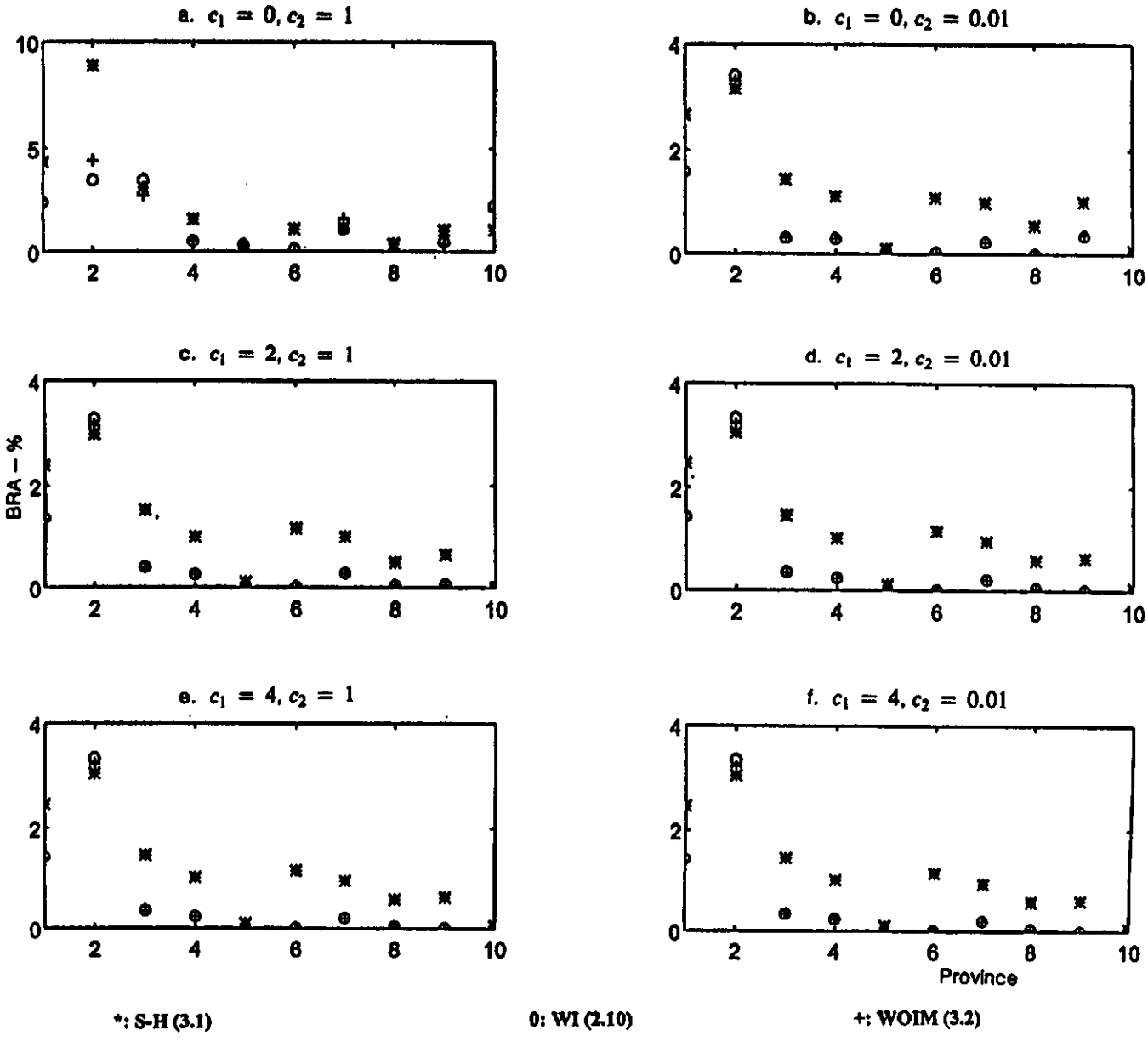

Figure 2. Biais relatif absolu de divers estimateurs pour un échantillon de 1%.

4. RÉSUMÉ ET CONCLUSIONS

Nous avons envisagé la généralisation de la méthodologie de l'analyse de la variance à une population gaussienne inverse à distribution asymétrique. Puis, nous avons étudié et appliqué les modèles sans interaction factorielle à l'estimation des paramètres régionaux d'une population finie. À partir de données d'enquêtes canadiennes, nous avons produit des populations synthétiques grâce à une simulation de Monte Carlo. Nous avons montré ainsi que les estimateurs proposés donnent de bons résultats dans diverses conditions, si la population peut être considérée comme un échantillon aléatoire tiré d'une population à distribution gaussienne inverse. Cette méthode représente un choix valable pour l'estimation des paramètres de données d'enquête dont la distribution présente une asymétrie positive.

REMERCIEMENTS

Les professeurs Y.P. Chaubey et F. Nebebe remercient vivement le Conseil de recherches en sciences naturelles et en génie du Canada pour l'appui financier qu'il leur a accordé. Les auteurs remercient le rédacteur, les professeurs J.N.K. Rao et A.B. Sim, ainsi que les lecteurs extérieurs, pour leurs commentaires précieux.

ANNEXE A

Valeurs des paramètres servant à définir la population IG

$$\mu = 3.13241147 \times 10^{-5}, \quad \sigma = 2.5447984 \times 10^{-5}$$

d	1	2	3	4	5
$10^6 \times \alpha_d$	3.1902855	2.8235779	1.5676078	.8056079	−.95350458

d	6	7	8	9	10
$10^6 \times \alpha_d$	−4.0661125	.49944356	.0061694263	−2.7414128	−1.1316622

g	1	2	3	4	5	6
$10^5 \times \beta_g$	1.0938451	.36781639	−.012707035	−.11561414	−.30936835	−1.023972

Valeurs de θ_{dg}:

d/g	1	2	3	4	5	6
1	22,000.82	26,183.11	29,080.48	29,977.59	31,826.13	41,195.19
2	22,179.76	26,436.94	29,393.94	30,310.79	32,201.96	41,827.05
3	22,815.33	27,344.90	30,520.70	31,510.37	33,559.25	44,146.20
4	23,219.00	27,926.81	31,247.41	32,285.58	34,439.96	45,682.95
5	24,207.76	29,369.63	33,064.91	34,229.61	36,661.02	49,674.90
6	26,180.44	32,324.63	36,858.30	38,311.45	41,383.33	58,760.34
7	23,385.24	28,167.65	31,549.24	32,607.90	34,806.97	46,330.96
8	23,658.15	28,564.53	32,047.98	33,140.96	35,415.03	47,414.57
9	25,302.90	30,997.31	35,142.43	36,461.01	39,232.58	54,516.76
10	24,312.62	29,524.12	33,260.85	34,439.64	36,902.04	50,118.45

ANNEXE B

Valeurs de la taille des cellules N_{dg}

d/g	1	2	3	4	5	6	Total
1	627	360	277	84	215	110	1,673
2	285	212	198	72	68	83	918
3	597	483	616	148	204	231	2,279
4	729	397	568	151	239	219	2,303
5	1,372	761	1,216	202	473	511	4,535
6	1,177	888	1,795	517	707	800	5,884
7	639	432	673	165	236	222	2,367
8	850	512	888	264	349	297	3,160
9	700	699	1,350	385	696	572	4,401
10	456	540	1,083	342	393	407	3,221

BIBLIOGRAPHIE

BHATTACHARYYA, G.K., et FRIES, A. (1986). On the inverse Gaussian multiple regression and model checking procedures. Dans *Reliability and Quality Control*, (A.P. Basu, Éd.). New York: North Holland, 86-100.

CHAUBEY, Y.P. (1991). A study of ratio and product estimators under super population. *Communications in Statistics*, A, 20 (5 et 6), 1731-1746.

CHOUDHRY, G.H., et RAO, J.N.K. (1988). Evaluation of small area estimators: an empirical study. Communication présentée à l'International Symposium on Small Area Statistics, New Orleans.

CHHIKARA, R.S., et FOLKS, J.L. (1989). *The Inverse Gaussian Distribution*. New York: Marcel Dekker, Inc.

DURBIN, J. (1959). A Note on the application of Quenouille's method of bias reduction to the estimation of ratio. *Biometrika*, 46, 477-480.

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.

FOLKS, J.L., et CHHIKARA, R.S. (1978). The inverse Gaussian distribution and its statistical applications-A review. *Journal of the Royal Statistical Society*, Séries B, 40, 263-275.

FRIES, A., et BHATTACHARYYA, G.K. (1983). Analysis of two-factor experiments under an inverse Gaussian model. *Journal of the American Statistical Association*, 78, 820-826.

GONZALES, 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.

GHOSH, M., et RAO, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-76.

HIDIROGLOU, M.A., et SÄRNDAL, C.-E. (1985). Étude empirique de quelques estimateurs de régression pour petits domaines. *Techniques d'enquête*, 6, 73-85.

HOLT, D., SMITH, T.M.F., et TOMBERLIN, T.J. (1979). A model-based approach to estimation for small subgroups of a population. *Journal of the American Statistical Association*, 74, 405-410.

IYENGAR, S., et PATWARDHAN, G. (1988). Recent developments in the inverse Gaussian distribution. Dans *Handbook of Statistics*, New York: Elsevier Science 479-490.

JOHNSON, N. L., et KOTZ, S. (1970). *Continuous Univariate Distributions-1, Distributions in Statistics*. New York: Wiley.

MACGIBBON, B., et TOMBERLIN, T.J. (1989). Estimation de proportions pour petites régions par des méthodes empiriques de Bayes. *Techniques d'enquêtes*, 15, 247-262.

MICHAEL, J.R., SCHUCANY, W.R., et HASS, R.W. (1976). Generating random variables using transformations with multiple roots. *American Statistician*, 30(2), 88-90.

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

PURCELL, N.J., et KISH, L. (1980). Postcensal estimates for local areas (or domains). *Bulletin de l'Institut International de Statistique*, 48, 3-18.

SÄRNDAL, C.-E. (1984). Design consistent versus model dependent estimation 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.

SÄRNDAL, C.-E., et RÅBÄCK, G. (1983). Variance reduction and unbiasedness for small domain estimators. *Statistical Review*, 21, (Dissertations en l'honneur de T.E. Dalenius), 33-40.

SCHAIBLE, W.L. (1979). A composite estimator for small area statistics. Dans *Synthetic Estimates for Small Areas*, NIDA Research Monograph 24, (J. Steinberg, Éd.). Rockville, MD: National Institute on Drug Abuse, 36-53.

STATISTIQUE CANADA (1987). Fichier de microdonnées, Revenu des ménages et équipement des ménages (1987), Statistique Canada, Division des enquêtes-ménages.

STROUD, T.W.F. (1987). Bayes and empirical Bayes approaches to small area estimation. Dans *Small Area Statistics*, (R.Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh, Éds.). New York: Wiley, 124-137.

WHITMORE, G.A. (1983). A regression method for censored inverse Gaussian data. *La Revue Canadienne de Statistique*, 11, 305-315.

Techniques d'enquête, juin 1996
Vol. 22, n° 1, pp. 43–53
Statistique Canada

Comparaison de quelques méthodes de correction de la non-réponse d'un panel

LOU RIZZO, GRAHAM KALTON et J. MICHAEL BRICK[1]

RÉSUMÉ

Dans certaines enquêtes, on peut utiliser, pour la correction de la non-réponse, de nombreuses variables auxiliaires relatives aux répondants et aux non-répondants. On peut s'interroger sur le choix des variables auxiliaires à retenir aux fins de cette correction et sur la façon des les utiliser. Dans la présente recherche, nous examinons diverses méthodes de correction de la non-réponse fondées sur des modèles de régression logistique, sur des algorithmes de recherche par catégories et sur la méthode du quotient généralisée. Ces méthodes servent à la pondération de la non-réponse pour l'enquête Survey of Income and Program Participation (SIPP). Les estimations issues des diverses méthodes de correction sont comparées entre elles et également à des estimations de référence provenant d'autres sources.

MOTS CLÉS: Biais de non-réponse; enquêtes par panel; méthode du quotient généralisée; estimations de référence.

1. INTRODUCTION

On utilise généralement la pondération, dans l'analyse des données d'enquêtes, afin de compenser les probabilités inégales de sélection des échantillons et pour la non-réponse de certaines unités, et de rendre les distributions de certaines variables de l'échantillon pondéré conformes à des distributions de population connues de ces variables (en compensant ainsi les carences dans la couverture et en améliorant la précision des estimations de l'enquête) (Kish 1992). Compte tenu de ces trois objectifs, la pondération est habituellement réalisée en trois étapes. Premièrement, un poids de base est calculé pour chaque élément échantillonné et correspond à l'inverse de la probabilité de sélection de chacun de ces éléments. Deuxièmement, les poids de base correspondant aux éléments répondants sont multipliés par un poids de non-réponse afin de corriger en fonction des non-répondants. Troisièmement, le poids ajusté est modifié pour faire en sorte que les distributions d'échantillons pondérés pour certaines variables soient conformes aux informations externes portant sur ces distributions.

Le présent article traite des méthodes de pondération qui cherchent à compenser la non-réponse des unités. Une des méthodes généralement utilisées pour le calcul des poids consiste à diviser l'échantillon total en une série de catégories de pondération fondées sur les informations recueillies tant sur les répondants que sur les non-répondants, puis d'augmenter les poids de base des répondants appartenant à une catégorie de pondération donnée afin de représenter les non-répondants de cette catégorie (Oh et Scheuren 1983; Kalton 1983). Dans beaucoup d'enquêtes, les informations disponibles sur les non-répondants, mises à part les unités primaires d'échantillonnage et la strate dont elles proviennent, sont assez limitées. Dans de tels cas, le choix des catégories de pondération possibles est restreint et la méthode est donc relativement simple à appliquer.

Dans certaines enquêtes, toutefois, on possède une masse considérable d'informations sur les non-répondants. Ces informations peuvent provenir de la base de sondage (p. ex., échantillonnage d'employés à partir des dossiers du personnel) ou résulter d'un appariement des éléments échantillonnés et des données administratives. En outre, avec les enquêtes par panel et d'autres types d'enquêtes comportant plus d'une étape de collecte des données, les réponses recueillies lors des premières étapes fournissent une foule de renseignements sur les non-répondants que l'on peut utiliser aux étapes ultérieures.

La présente étude porte principalement sur les méthodes de correction de la non-réponse à utiliser lorsqu'on dispose de renseignements détaillés sur les caractéristiques des non-répondants. Dans une telle situation, le choix de la méthode à retenir dépendra des variables auxiliaires utilisées et de la façon dont on les utilisera.

Pour les fins de notre exposé, nous appliquerons diverses méthodes de pondération aux données d'une enquête particulière intitulée "Survey of Income and Program Participation" (SIPP). L'enquête SIPP est une enquête par panel réalisée en continu auprès des ménages par le U.S. Bureau of the Census. Les non-répondants d'un panel de l'enquête SIPP peuvent être répartis en deux groupes: ceux qui ne répondent pas à la première vague de collecte de données (non-répondants de la vague initiale), et ceux qui répondent à la vague initiale, mais qui ne répondent pas à la suivante ou aux vagues subséquentes du panel auxquels ils sont admissibles (non-répondants du panel). Dans ce dernier cas, la correction de la non-réponse du panel peut s'appuyer sur une masse importante de renseignements tirés de la vague initiale de collecte de données. Les méthodes de correction étudiées ici portent uniquement sur les non-répondants du panel. Ces méthodes modifient les poids des répondants du panel (c'est-à-dire, ceux qui fournissent des renseignements dans toutes les vagues où ils sont admissibles) et permettent de corriger en fonction des non-répondants.

[1] Lou Rizzo, Graham Kalton et J. Michael Brick, Westat Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.

Aux fins de l'enquête SIPP, on procède chaque année à des entrevues auprès d'un échantillon national de ménages, et toutes les personnes âgées de quinze ans et plus vivant dans ces ménages à l'époque de la vague initiale deviennent les membres du panel qui fera l'objet du suivi subséquent. Jusqu'à maintenant, les panels de l'enquête SIPP ont eu une durée de vie de 32 mois, mais cette durée sera portée à quatre ans avec le panel de 1996. Les membres des panels sont interviewés tous les quatre mois. Les données recueillies portent sur le revenu, la participation aux programmes de soutien du revenu et d'autres facteurs qui peuvent influer sur le revenu et la situation économique. On recueille également des données sur les enfants. Pour de plus amples informations sur le plan de sondage de l'enquête SIPP, voir Nelson, McMillen et Kasprzyk (1985) et Jabine, King et Petroni (1990).

L'étude qui fait l'objet du présent document a porté sur le panel de l'enquête SIPP de 1987. Les renseignements utilisés provenaient des bases de données à grande diffusion. Ce panel a débuté avec un échantillon d'environ 12,300 ménages, et le suivi comprenait en sept vagues successives de collecte de données. Le taux de non-réponse des ménages à la vague initiale était de 6.7% (Jabine et coll. 1990). Les ménages qui ont répondu à la vague initiale renfermaient 30,841 personnes, enfants compris. De ce total, 20.8% ont omis de fournir des données pour l'ensemble des vagues auxquelles elles étaient admissibles: c'étaient les non-répondants du panel.

Outre le choix des variables auxiliaires et l'étude des diverses méthodes d'utilisation de ces variables aux fins de la correction de la non-réponse du panel, nous avons procédé à une évaluation comparative des méthodes. Les estimations obtenues à l'aide de diverses méthodes de rechange ont d'abord été comparées entre elles, puis comparées chacune à des estimations de référence. La partie finale du présent article résume les résultats obtenus et tire certaines conclusions sur l'efficacité des diverses méthodes de correction examinées. Rizzon, Kalton et Brick (1994) étudient ces questions d'une façon plus détaillée.

2. PRÉDICTEURS DE LA PROPENSION À RÉPONDRE

La première étape de la correction de la non-réponse du panel consiste à décider quels sont, parmi les nombreux items de la première vague de collecte de données, ceux qui devraient servir à la pondération. La présente section porte sur ce choix. Il s'agit de déterminer les items dont les réponses permettront de distinguer les participants en fonction de leur propension à répondre au cours des vagues subséquentes de l'enquête. Little (1986) appelle cette méthode une stratification de la propension à répondre; il montre que l'important biais des estimations peut être réduit en multipliant le poids de base par l'inverse de la probabilité de réponse d'un élément.

Dans l'enquête par panel SIPP de 1987, 58 des items de la vague initiale de collecte des données (vague 1) pouvaient servir de variables explicatives possibles de la non-réponse du panel. Tous les items utilisés effectivement par le Bureau of the Census aux fins de la correction de la non-réponse du panel de l'enquête SIPP faisaient partie de ce groupe de 58, à l'exception des renseignements sur la Metropolitan Statistical Area (MSA), exclus de la base de données à grande diffusion pour des raisons de confidentialité.

Nous avons jugé que l'analyse de régression logistique était une méthode toute désignée pour le choix d'un modèle de la non-réponse du panel, le rapport des répondants sur les non-répondants du panel servant de variable dépendante. Toutefois, avant d'entreprendre cette modélisation, nous avons procédé à une présélection des variables afin de faciliter les opérations subséquentes. D'une manière générale, les items retenus aux fins de l'analyse de régression logistique ont été ceux appartenant à des catégories laissant constater entre elles une différence statistiquement significative et atteignant au moins 4 points. Cependant, pour toutes sortes de raisons, certains items ont également été retenus même s'ils ne répondaient à ces exigences. Par exemple, la différence entre les taux de réponse des hommes et des femmes du panel était inférieure à 2%, ce qui ne nous a pas empêché d'utiliser le sexe dans certaines des analyses subséquentes.

La présélection nous a permis de réduire de 58 à 31 le nombre d'items de l'analyse de régression logistique. Voici quels ont été les items retenus: statut d'occupation du logement, logement social, type de ménage, région de recensement, scolarité du ménage, taille du ménage, revenu du ménage, possession d'instruments financiers (obligations), sexe, race, origine hispanique, lien de parenté avec la personne de référence (PPR), âge, état civil, type de famille, scolarité, état d'étudiant, état d'emploi, revenu personnel, emplois multiples, catégorie d'emploi, bénéficiaire des programmes Medicare, Medicaid, Women, Infants, and Children (WIC), Aid to Families with Dependent Children (AFDC) ou de coupons alimentaires, bénéficiaire d'un programme d'aide général, de la Sécurité sociale ou d'un autre programme de bien-être, état d'ancien combattant et nombre d'items imputés à la vague 1.

Le dernier item (nombre d'items imputés) a été retenu pour servir d'indice de collaboration à la vague 1. D'autres études ont en effet démontré que les personnes qui coopèrent moins à la vague initiale d'une enquête par panel ont un plus grand risque de devenir des non-répondants lors des vagues ultérieures (voir, par exemple, Kalton, Lepkowski, Montanari et Maligalig 1990). Comme nous le verrons plus loin, cet indice s'est avéré étroitement lié à la non-réponse du panel.

2.1 Analyse de régression logistique

Puisque les 31 items retenus à l'issue de la présélection étaient au moins en partie liés à la non-réponse du panel, ils constituaient tous des variables utilisables aux fins de la pondération visant à réduire le biais de non-réponse du panel dans les estimations de l'enquête. Toutefois, la présélection a été imparfaite puisqu'elle n'a pas tenu compte des rapports existant entre les items, retenant ainsi plus de variables qu'il n'en fallait pour la correction de la non-réponse du panel. Par exemple, deux items qui dépendent

Tableau 1

Estimations des paramètres du modèle de régression logistique

Prédicteurs	Estimations des paramètres
Coordonnée à l'origine	−0.465
Âge ($\chi^2 = 184.9, p < .0001$).	
< 16	−0.179
16-24	0.446
25-50	0.187
51-71	−0.056
> 71	0.0
Race ($\chi^2 = 214.0, p < .0001$).	
Blanche	−0.351
Noire	0.255
Autre	0.0
PPR ($\chi^2 = 69.0, p < .0001$).	
Membre de la famille	−0.251
Non-membre de la famille	0.0
Région de recensement ($\chi^2 = 327.3, p < .0001$).	
Nouvelle-Angleterre	0.009
Côte Atlantique moyenne	0.167
Côte Atlantique sud	0.027
Centre sud-est	−0.231
Centre nord	−0.396
Rocheuses/Centre sud-ouest	0.425
Pacifique	0.0
Statut d'occupation du logement ($\chi^2 = 207.2, p < .0001$).	
Propriétaire	−0.154
Locataire	0.331
Autre	0.0
Items imputés ($\chi^2 = 434.2, p < .0001$).	
0	−0.626
1	−0.244
2 à 3	0.296
> 3	0.0
Instruments financiers (obligations) ($\chi^2 = 97.1, p < .0001$).	
Pas d'obligations	0.168
Un certain nombre d'obligations	0.0
État d'emploi ($\chi^2 = 33.4, p < .0001$).	
Non-chômeur	−0.179
Chômeur	0.0
Coupons alimentaires ($\chi^2 = 39.3, p < .0001$).	
Non-bénéficiaire	−0.191
Bénéficiaire	0.0
Catégorie d'emploi ($\chi^2 = 31.4, p < .0001$).	
Affaires	0.100
Autre	0.103
Fonctionnaire	0.0
Scolarité ($\chi^2 = 12.8, p = .0003$).	
Dixième ou onzième année	−0.075
Autre	0.0
Revenu du ménage ($\chi^2 = 14.9, p = .0006$).	
Moins de $1,200/mois	0.117
$1,200 à $8,000/mois	−0.088
Plus de $8,000/mois	0.0
Sexe ($\chi^2 = 10.3, p = .0013$).	
Homme	0.047
Femme	0.0
Rapport PPR/ < 16 ans ($\chi^2 = 10.1, p = .0015$).	
Membre de la famille, enfant	0.096
Autre	0.0

très étroitement de la propension à répondre risquent également d'être étroitement liés l'un à l'autre de telle sorte qu'un seul des deux pourrait suffire à la pondération. En conséquence, l'étape suivante de la sélection des prédicteurs de la non-réponse du panel a consisté à examiner quelles étaient les combinaisons d'items qui permettaient le mieux de prévoir la propension à répondre du panel.

Nous avons utilisé une méthode de régression logistique pour examiner les rapports conjoints de plusieurs items avec la propension à répondre du panel. Les modèles de régression ont été ajustés en utilisant les poids de la première vague qui tenaient compte des probabilités de sélection inégales et de la non-réponse à la première vague. Après avoir examiné un certain nombre de modèles possibles, nous avons retenu un modèle comportant 13 variables d'effet majeur et un terme d'interaction, et offrant une représentation raisonnablement adéquate des données.

Nous présentons au tableau 1 les estimations des paramètres pour chaque niveau de chaque prédicteur de ce modèle, chacune accompagnée de la valeur correspondante de la statistique de Wald (χ^2). La valeur du paramètre du dernier niveau de chaque prédicteur (niveau repère) est fixée à zéro. Les estimations des paramètres des autres niveaux de chaque prédicteur représentent les différences dans la propension à répondre à partir du niveau repère. Comme on le constate à la lecture des statistiques de Wald, les prédicteurs jouent tous un rôle hautement significatif dans le modèle.

Ce modèle présente la particularité de ne contenir qu'un seul terme d'interaction, soit le rapport «lien de parenté avec la personne de référence/personne de moins de 16 ans». Toutes les autres interactions examinées présentaient des valeurs de χ^2 plus petites que cette dernière. Même le rapport «PPR/ < 16 ans» avait une efficacité prédictive relativement faible. En fait, cette interaction ainsi que les trois derniers prédicteurs du tableau 1 (scolarité, revenu du ménage et sexe) ont été exclus de la plupart des opérations de pondération abordées plus loin à cause de leur puissance prédictive limitée de la propension à répondre du panel. Les méthodes de correction sont principalement fondées sur un modèle d'effets majeurs réduits comprenant les 10 premières variables prédictives énumérées au tableau 1.

3. AUTRES MÉTHODES DE CORRECTION DES POIDS

La méthode utilisée dans le cadre de l'enquête SIPP pour corriger la pondération de la non-réponse du panel est décrite par Chapman, Bailey et Kasprzyk (1986). Il s'agit essentiellement de former des cellules de correction de la non-réponse, puis de corriger les poids en utilisant l'inverse des taux de réponse de chacune des cellules. Les cellules sont formées par classification croisée des réponses tirées d'une série de variables de la vague 1 que l'on juge liées à la réponse du panel. Les petites cellules sont combinées de manière que la taille de l'échantillon obtenu dans chaque groupe de cellules soit de 30 ou plus. La réciproque

du taux de réponse observé (pondéré) dans chaque groupe de cellules correspond à la correction de la non-réponse du panel pour ce groupe. La correction de la non-réponse du panel est ensuite multipliée par le poids de la vague 1 afin de créer le poids corrigé de la non-réponse. Le poids de la vague 1 comprend une correction pour la non-réponse de la vague 1, mais il ne comprend pas celle de stratification a posteriori pour cette vague.

Dans la présente section, nous examinons diverses méthodes de correction des poids de non-réponse du panel. Ces méthodes peuvent être réparties en trois groupes:

- Méthodes de régression logistique;
- Méthodes CHAID;
- Méthodes du quotient généralisées.

Nous abordons ci-après chacune de ces méthodes. Nous décrivons en détail les procédures d'élaboration de correction des poids et les propriétés statistiques importantes des corrections.

3.1 Corrections fondées sur les modèles logistiques

La première série de corrections de la pondération que nous examinons s'inspire directement du modèle de régression logistique décrit dans la section précédente. Cette correction de la pondération de la non-réponse du panel, appelé «correction logistique des taux de réponse prévus», a été calculée en utilisant l'inverse des taux de réponses prévus du modèle de régression logistique des effets majeurs réduits pour chacune des cellules de la classification croisée des 10 variables prédictives de ce modèle.

Comme les paramètres utilisés pour le calcul des taux de réponse prévus sont estimés avec un modèle des effets majeurs à partir des réponses marginales pour les variables, la petite taille des cellules de la classification croisée de toutes les variables ne pose pas de difficultés. Toutefois, cet avantage est obtenu au prix d'une dépendance complète vis-à-vis de la validité du modèle des effets majeurs; autrement dit, nous devons présumer qu'il n'existe pas, entre les variables, d'interaction dont nous devrions tenir compte.

Un des moyens de réduire notre dépendance vis-à-vis du modèle des effets majeurs consiste à fonder les corrections sur les taux de réponse observés dans les cellules dont les échantillons sont suffisamment grands pour assurer la stabilité des taux de réponse, et de fonder les corrections sur les taux de réponse prévus dans les autres cellules. La deuxième méthode de correction fondée sur la régression logistique utilise cette stratégie mixte. Dans les cellules qui contiennent 25 personnes échantillonnées ou plus, la correction de la non-réponse correspond à l'inverse du taux de réponse observé dans la cellule. Dans les cellules qui contiennent moins de 25 personnes échantillonnées, la correction de la non-réponse correspond à l'inverse du taux de réponse prévu pour chaque cellule. On donne à cette méthode le nom de «correction logistique des taux de réponse mixtes».

Une troisième méthode de correction du poids de non-réponse fondée sur les modèles logistiques rappelle les opérations utilisées dans le cadre de l'enquête SIPP. Les cellules initiales ont été définies par classification croisée des 10 variables indépendantes utilisées dans la régression logistique. Les cellules ont ensuite été groupées jusqu'à ce que la taille de l'échantillon de chaque cellule dépasse 30, puis l'inverse du taux de réponse observé à l'intérieur de chaque groupe de cellules est devenu la correction de la non-réponse. On a pris soin de grouper les cellules qui présentaient des taux de réponses prévus comparables. On a donné à cette méthode de correction du poids de non-réponse le nom de «correction logistique en cellules groupées». Même si cette méthode ressemble à celle utilisée pour la correction de la non-réponse du panel dans le cadre de l'enquête SIPP, il existe certaines différences dans les variables utilisées pour définir les cellules et dans les méthodes utilisées pour combiner les petites cellules.

Pour ces trois méthodes de correction du poids fondées sur le modèle de régression logistique, les taux de réponse observés et prévus ont été calculés à partir de dénombrements pondérés du nombre de cas, (au lieu d'utiliser des nombres non pondérés), en utilisant les poids corrigés de la non-réponse de la vague 1. En pratique toutefois, les corrections pondérés et non pondérés étaient presque identiques.

3.1.1 Corrections fondées sur les modèles CHAID

La deuxième catégorie de méthodes de correction du poids de non-réponse du panel est fondée sur l'utilisation de l'algorithme de recherche par catégories CHAID pour la répartition de l'ensemble de données en cellules de correction. Il s'agit d'une manière générale d'assimiler des cellules de correction à des combinaisons de réponses aux variables prédictives qui présentent le plus grand pouvoir discriminatif en ce qui a trait aux taux de réponse du panel, à condition cependant que chaque cellule contienne un échantillon minimal de 25 personnes. La correction de la non-réponse du panel correspond à l'inverse du taux de réponse observé dans chaque cellule.

L'algorithme CHAID produit des cellules en séparant progressivement les séries de données en une structure arborescente. La séparation pour chaque nouvelle branche est effectuée en choisissant la variable qui maximise un critère χ^2. Lorsque la séparation porte sur une variable polychotomique, elle risque d'intéresser plusieurs branches. Les tests de χ^2 sont modifiés en utilisant des corrections de type Bonferroni afin d'éviter que les variables ne soient choisies simplement parce qu'elles englobent un plus grand nombre de catégories. Le modèle CHAID est une version parmi d'autres du détecteur automatique d'interactions (DAI) mis au point pour les variables catégoriques. Kass (1980) expose la théorie qui sous-tend la technique CHAID. Lepkowski, Kalton et Kasprzyk (1989) et Kalton, Lepkowski et Lin (1985) ont utilisé une autre version de cette méthode pour la modélisation de la non-réponse dans les enquêtes SIPP.

En ce qui concerne l'analyse présente, nous avons examiné deux modèles CHAID incluant des séries différentes de variables prédictives. Le premier modèle comprenait les sept prédicteurs les plus importants du modèle de régression logistique (âge, lien de parenté avec la personne de référence, race du chef du ménage, statut d'occupation du logement, région du recensement, items imputés et

possession d'instruments financiers) en plus du sexe. Ce modèle a donné 99 cellules de correction de la non-réponse. La correction de la non-réponse fondé sur ce modèle a reçu le nom de CHAID 1. Le deuxième modèle CHAID comprenait les 13 variables de prévision du modèle de régression logistique énumérées au tableau 1. Ce modèle a donné 142 cellules de correction de la non-réponse. La correction de la non-réponse pour ce modèle a reçu le nom de CHAID 2.

3.1.2 Corrections fondées sur la méthode du quotient généralisée

La troisième catégorie de méthodes de correction du poids de non-réponse du panel est fondée sur la méthode du quotient généralisée. À la différence des autres catégories de méthodes, l'élaboration des cellules de correction de la non-réponse n'a pas été fondée sur la classification croisée des variables prédictives. Nous avons plutôt utilisé la méthode du quotient pour forcer l'égalité entre les distributions marginales des répondants du panel pour chacune des variables prédictives (calculées à l'aide des poids corrigés) et les distributions correspondantes des répondants et des non-répondants combinées (calculées à l'aide des poids originaux de la vague 1). Kalton et Kasprzyk (1986) donnent à cette méthode le nom de «méthodes du quotient fondée sur l'échantillon». Nous avons utilisé les 10 variables prédictives du modèle de régression logistique réduit pour définir les distributions marginales. Ainsi, le problème de correction comportait 10 dimensions: une pour chaque variable de prévision.

La méthode du quotient exige une modification des poids originaux pour satisfaire à certaines contraintes marginales tout en minimisant la distance qui sépare les poids originaux des poids ajustés. Deville et Särndal (1992) décrivent certaines fonctions de distance auxquelles on peut recourir et obtiennent les méthodes de correction correspondantes. L'algorithme de la méthode du quotient de Deming et Stephan (1942), qui utilise implicitement une fonction de distance conduisant à une solution multiplicative, constitue une des formes de correction fondée sur la méthode du quotient généralisée.

Nous avons utilisé le logiciel CALMAR décrit par Deville, Särndal et Sautory (1993) pour calculer les corrections. Trois fonctions de distance différentes ont été examinées: la méthode multiplicative, la méthode linéaire et la méthode multiplicative tronquée. Les corrections correspondant à chacune des trois fonctions de distance se sont avérées presque identiques. Ce résultat empirique est conforme aux résultats obtenus par Deville et Särndal (1992) qui montrent que les estimateurs qui utilisent des poids générés avec des fonctions de distance différentes sont asymptotiquement équivalents si les fonctions de distance répondent à certaines conditions de lissage. Les trois fonctions de distance utilisées dans la présente recherche répondent à ces conditions. Comme les corrections avec chacune des trois méthodes étaient presque identiques, la méthode multiplicative a seule été retenue aux fins des évaluations ultérieures. Nous donnons à cette méthode le nom de correction par la méthode du quotient.

3.1.3 Distributions des corrections de la non-réponse

Les corrections correspondant à chacune des six méthodes décrites plus haut ont été calculées pour les données de l'enquête SIPP de 1987. Nous présentons au tableau 2 les distributions des corrections du poids de non-réponse ainsi obtenues. Seules les corrections y sont indiquées; il n'est pas question des poids tirés de ces corrections ni des poids de la vague 1. Le tableau 2 est divisé en

Tableau 2

Distribution des paramètres d'ajustement du poids de non-réponse du panel

	Moyenne	Minimum	Médiane	Maximum	$1 + CV^2$
Correction logistique					
Taux de réponse prévus	1.26	1.04	1.20	4.28	1.02
Taux de réponse mixtes	1.26	1.00	1.20	4.28	1.03
Cellules groupées	1.26	1.00	1.20	3.43	1.02
CHAID 1	1.26	1.02	1.22	3.49	1.03
CHAID 2	1.26	1.01	1.19	13.93	1.04
Méthode du quotient	1.26	0.91	1.23	2.51	1.02

Corrélations

	Correction logistique			CHAID 1	CHAID 2	Méthode du quotient
	Taux de réponse prévus	Taux de réponse mixtes	Cellules groupées			
Correction logistique						
Taux de réponse prévus	1.00	0.96	0.73	0.73	0.63	0.95
Taux de réponse mixtes		1.00	0.73	0.72	0.63	0.90
Cellules groupées			1.00	0.69	0.58	0.75
CHAID 1				1.00	0.81	0.73
CHAID 2					1.00	0.63
Méthode du quotient						1.00

deux portions: la portion supérieure donne la moyenne, la médiane et les valeurs extrêmes de chaque distribution de corrections, ainsi que la valeur $(1 + CV^2)$, où CV désigne le coefficient de variation de chaque correction. La valeur $(1 + CV^2)$ est un indicateur de l'augmentation de la variance des estimations introduites par l'utilisation de facteurs variables de la correction du poids de non-réponse (voir Kish 1992). La seconde portion du tableau donne les corrélations qui existent entre les diverses formes de correction étudiées.

Puisque le taux de réponse global pondéré du panel est de 0.794, la correction moyenne globale de la non-réponse serait de $1/(0.794) = 1.26$ si la même correction était utilisée pour toutes les personnes. Les corrections de pondération moyennes pour les trois méthodes de correction de la pondération qui utilisent l'inverse des taux de réponse des cellules (méthode de correction logistique en cellules groupées, CHAID 1 et CHAID 2) sont nécessairement égales à la correction globale de la non-réponse de 1.26. Les corrections de pondération moyennes des autres méthodes ne diffèrent que de façon minime de la correction moyenne globale du poids de non-réponse.

Pour les six méthodes, les distributions présentent une asymétrie positive, quelques cas présentant des poids importants. Par leur nature même, les méthodes fondées sur la régression logistique et les modèles CHAID ne peuvent avoir de corrections inférieures à 1.00, contrairement à ce qui se passe dans le cas de l'algorithme de la méthode du quotient. Les poids médians sont comparables pour toutes les méthodes, mais les poids maximums ne le sont pas. La méthode CHAID 2 présente une cellule avec un taux de réponse de seulement 7%, ce qui donne le poids maximal le plus élevé, soit 13.93. Le poids maximal le moins élevé (2.51) s'observe avec la méthode du quotient.

Les valeurs de $(1 + CV^2)$ sont assez uniformes d'une méthode de correction à l'autre. La méthode CHAID 2 présente la valeur de $(1 + CV^2)$ la plus grande, surtout par suite de la présence d'un plus grand nombre de valeurs aberrantes (comme le maximum de 13.93). Toutefois, même pour cette méthode, l'augmentation approximative de la variance des estimations de l'enquête n'est que de 4%. La correction par la méthode du quotient présente la plus petite augmentation de variance (2%), mais cette augmentation n'est pas très différente de celles observées avec les autres méthodes.

Les corrélations par paires entre les six séries de poids varient de 0.58 à 0.96. On constate comme prévu une corrélation très forte entre les poids du modèle de correction logistique des taux de réponse prévus et du modèle de correction logistique des taux de réponse mixtes. Compte tenu de la similitude entre la méthode de régression logistique des effets majeurs prévus et de la méthode du quotient, il n'est pas surprenant d'observer une très grande corrélation entre les poids obtenus avec ces deux méthodes. La corrélation relativement grande observée entre les poids de la méthode du quotient et ceux de la méthode CHAID 1 et de la méthode de correction logistique en cellules groupées confirme l'absence de termes d'interaction importants constatée antérieurement. Les poids de la méthode CHAID 2 présentent les corrélations les plus faibles avec ceux des autres méthodes, sauf avec ceux de la méthode CHAID 1. Ce résultat est probablement dû à la grande variabilité des poids de la méthode CHAID 2 qui découle d'un nombre élevé de cellules de correction (142).

3.2 Poids finals du panel

Les poids de non-réponse du panel examinés dans la section précédente découlent des corrections apportées aux poids de la vague 1 afin de compenser la non-réponse du panel. Les poids finals qui peuvent servir dans l'analyse des données de l'enquête par panel SIPP s'obtiennent en multipliant les poids de non-réponse du panel par les poids de la vague 1, puis en procédant à une stratification a posteriori pour faire en sorte que les totaux pondérés de l'échantillon soient conformes à ceux tirés principalement de l'étude de la population actuelle (Current Population Survey ou CPS). Nous avons effectué cette opération pour chacune des six méthodes de correction du poids de non-réponse du panel décrites plus haut.

Tableau 3

Corrélations entre les poids issus de la stratification a posteriori et mesures de l'augmentation de la variance

	Panel SIPP	Correction logistique			CHAID 1	CHAID 2	Méthode du quotient
		Taux de réponse prévus	Taux de réponse mixtes	Cellules groupées			
Panel SIPP	1.00	0.75	0.74	0.75	0.71	0.68	0.77
Correction logistique							
taux de réponse prévus		1.00	0.99	0.91	0.90	0.86	0.98
taux de réponse mixtes			1.00	0.91	0.90	0.86	0.97
cellules groupées				1.00	0.89	0.85	0.93
CHAID 1					1.00	0.94	0.91
CHAID 2						1.00	0.87
Méthode du quotient							1.00
$1 + CV^2$	1.08	1.09	1.09	1.08	1.09	1.10	1.08

La méthode de stratification a posteriori retenue était équivalente à celle utilisée dans le cadre de l'enquête SIPP, sauf que dans ce dernier cas, la stratification s'effectue par groupes de renouvellement (dans le cas des méthodes de pondération de rechange, la stratification a posteriori s'effectue sur l'ensemble des groupes de renouvellement combinés). Cette différence ne devrait pas avoir un effet appréciable. Après la stratification a posteriori, les six séries de poids finals et les poids du panel de l'enquête SIPP s'additionnent pour donner les même totaux de contrôle.

Pour comparer les poids finals obtenus à l'aide des six méthodes de correction, d'abord entre eux puis aux poids de l'enquête par panel SIPP, nous avons calculé les corrélations entre les poids ainsi que la mesure de la variabilité utilisée antérieurement, $(1 + CV^2)$. Les résultats sont présentés au tableau 3. Les estimations de la variabilité due à la pondération $(1 + CV^2)$ laissent constater des augmentations semblables oscillant entre 8% et 10% pour les variances des estimations de l'enquête de chacune des méthodes de pondération. Celles établies entre les séries de poids finals de l'enquête par panel sont toutes égales ou supérieures à 0.85. Si on compare ces corrélations à celles présentées au tableau 2, il paraît évident que les corrélations entre les poids finals sont sensiblement plus élevées que celles établies entre les poids de non-réponse du panel. Les corrélations entre le poids de l'enquête par panel SIPP et les poids finals des méthodes de rechange sont régulièrement inférieures à toutes les autres, ce qui s'explique probablement du fait que les variables utilisées aux fins des corrections du poids de non-réponse pour le premier différaient de ceux utilisés pour les seconds. Les variables utilisées dans les méthodes de rechange et qui n'étaient pas utilisées dans la pondération de l'enquête par panel SIPP étaient l'âge, le lien de parenté avec la personne de référence, le nombre d'items imputés, la catégorie d'emploi et l'admissibilité au programme de coupons alimentaires. La taille du ménage était la seule variable autre que les renseignements sur la MSA (exclus pour des raisons de confidentialité) prise en compte dans les poids de l'enquête par panel SIPP et ignorée dans les méthodes de rechange parce que son lien avec les taux de réponse n'avait pas été jugé significatif.

4. COMPARAISON DES ESTIMATIONS OBTENUES À L'AIDE DES DIVERSES MÉTHODES DE PONDÉRATION

Dans la section précédente, nous avons décrit l'élaboration des séries de poids de rechange qui peuvent être utilisées pour l'analyse des données de l'enquête par panel SIPP. Toutes les méthodes de pondération de rechange examinées incorporent des corrections pour tenir compte des probabilités de sélection inégales, de la non-réponse de la vague initiale, de la non-réponse du panel et de la stratification a posteriori dans les totaux de contrôle externes. Dans la présente section, nous comparons les estimations obtenues à l'aide des méthodes de pondération de rechange, d'abord l'une à l'autre puis aux estimations correspondantes obtenues avec les poids de l'enquête SIPP. En outre, le cas échéant, les diverses estimations de l'enquête sont également comparées aux estimations externes provenant d'autres sources. Certaines des estimations externes sont des estimations de référence obtenues à partir des données administratives de l'Étude de la population actuelle. D'autres estimations externes proviennent de la vague 1 de l'enquête par panel SIPP de 1989. Les données recueillies à la vague 7 de l'enquête par panel SIPP de 1987 portent sur la même période que celles recueillies à la vague 1 de l'enquête SIPP de 1989, ce qui signifie que les estimations provenant de ces deux sources de données devraient être comparables.

Pour faire les comparaisons avec les estimations de référence, il convient de reconnaître que toute différence relevée peut être due à une variété de facteurs parmi lesquels la non-réponse du panel ne constitue qu'une possibilité. Par exemple, les erreurs de réponses et les différences entre les définitions peuvent expliquer les différences observées entre les estimations de l'enquête SIPP et les estimations de référence. Il convient donc d'user de prudence lors des comparaisons avec les estimations de référence. Comme les estimations de l'enquête par panel SIPP de 1989 sont fondées sur les données de la vague 1, elles sont à l'abri du problème de la non-réponse du panel. Ainsi, les différences observées entre les estimations obtenues à partir des enquêtes par panel SIPP de 1987 et de 1989 sont peut-être celles qui risquent le plus d'être causées par l'incapacité des corrections du poids de non-réponse du panel à compenser entièrement le biais de non-réponse du panel. Toutefois, même dans ce cas, d'autres explications possibles telles que le conditionnement du panel pourraient contribuer aux différences observées (même si Pennell et Lepkowski (1992) montrent que le conditionnement du panel ne constitue pas un facteur important dans la plupart des estimations des enquêtes SIPP).

Nous présentons au tableau 4 une gamme d'estimations tirées des données d'enquêtes par panel SIPP de 1987 qui utilisent les poids de l'enquête SIPP ainsi que ceux obtenus à l'aide des six méthodes de pondération de rechange, ainsi que les estimations de référence correspondantes et les estimations tirées de l'enquête par panel SIPP de 1989, lorsqu'elles étaient disponibles. Les estimations sont données en pourcentages, sauf dans le cas des estimations du nombre moyen de mois sans assurance-maladie, du revenu médian des ménages et des salaires annuels. Les estimations portent sur la population totale, sauf dans le cas des estimations de la situation d'emploi (pourcentage des personnes occupées, des chômeurs et des inactifs), qui correspondent à des sujets âgés de plus de 15 ans, et dans le cas des salaires annuels, qui correspondent à des personnes âgées de plus de 14 ans. Les estimations correspondent à trois périodes différentes: juin 1987, janvier 1989 et l'année civile 1987. Par exemple, les trois premières estimations du tableau 4 sont les pourcentages estimés de personnes participant au programme AFDC (Aid to Families with Dependent Children)

Tableau 4

Estimations pour la population totale des données de l'enquête par panel SIPP de 1987 avec les méthodes de pondération de rechange, et estimations provenant d'autres sources

	Panel SIPP	Correction logistique			CHAID 1	CHAID 2	Méthode du quotient	Panel SIPP 1989	Référence
		Taux de réponse prévus	Taux de réponse mixtes	Cellules groupées					
AFDC - Juin 1987	3.73	3.70	3.74	3.72	3.71	3.60	3.69		4.28[1]
AFDC - Janvier 1989	3.10	3.12	3.14	3.12	3.14	3.02	3.10	3.56	4.24[2]
AFDC - Année 1987	4.85	4.78	4.82	4.81	4.80	4.69	4.78		
Coupons alimentaires - Juin 1987	7.43	7.26	7.30	7.34	7.38	7.20	7.21		7.35[3]
Coupons alimentaires - Janvier 1989	6.71	6.63	6.67	6.64	6.70	6.59	6.58	6.30	7.29[4]
Coupons alimentaires - Année 1987	10.30	10.11	10.16	10.18	10.24	10.05	10.06		
Medicaid - Janvier 1989	6.77	6.78	6.81	6.75	6.81	6.68	6.76	6.97	
Medicaid - Année 1987	9.21	9.21	9.24	9.21	9.25	9.09	9.21		
SSI - Juin 1987	1.68	1.70	1.69	1.67	1.69	1.65	1.69		1.68[3]
SSI - Janvier 1989	1.65	1.67	1.66	1.64	1.66	1.61	1.66	1.65	1.74[3]
SSI - Année 1987	1.80	1.82	1.82	1.80	1.82	1.78	1.82		
Sécurité sociale - Janvier 1989	14.92	14.87	14.87	14.89	14.88	14.89	14.85	15.14	
Taux de pauvreté - Juin 1987	10.88	10.75	10.79	10.76	10.79	10.69	10.74		
Taux de pauvreté - Janvier 1989	12.91	12.98	13.02	12.97	12.99	12.91	12.93	14.46	
Sous le seuil de pauvreté 1987/1988	2.25	2.31	2.32	2.30	2.29	2.32	2.31		
Au-dessus du seuil de pauvreté 1987/1988	2.69	2.63	2.64	2.60	2.62	2.63	2.63		
Nombre moyen de mois sans assurance-maladie - 1987	1.66	1.69	1.70	1.67	1.67	1.69	1.69		
Revenu médian du ménage - Janvier 1989	2,601	2,600	2,597	2,607	2,607	2,607	2,602	2,550	
Salaires annuels 1987 (billions $)	1.93	1.94	1.93	1.94	1.94	1.94	1.94		2.22[4]
Employés - Janvier 1989	62.74	62.36	62.34	62.43	62.42	62.52	62.42	61.60	
Chômeurs - Janvier 1989	3.57	3.64	3.63	3.60	3.58	3.60	3.63	4.52	
Inactifs - Janvier 1989	33.69	34.01	34.03	33.96	34.01	33.88	33.95	33.88	
Mariés en 1987	1.39	1.41	1.40	1.39	1.39	1.39	1.41		1.86[5]
Divorcés en 1987	0.51	0.50	0.50	0.49	0.50	0.51	0.49		0.90[6]
Changement d'adresse en 1987	12.88	13.32	13.32	13.19	13.36	13.37	13.33		17.99[6]

[1] Social Security Bulletin, Volume 52, N° 3.
[2] Social Security Bulletin, Volume 51, N° 7.
[3] USDA Food and Nutrition Service, données inédites.
[4] U.S. Bureau of the Census, Current Population Reports, Consumer Income, P-60, N° 174.
[5] National Center for Health Statistics: Vital Statistics of the U.S., 1987, Volume III, Marriage and Divorce, DHHS Pub. N° (PHS) 91-1103.
[6] U.S. Bureau of the Census, Current Population Reports, Population Characteristics, P-20, N° 473.

Tableau 5

Différences normalisées entre les estimations de l'enquête par panel SIPP de 1987 et les estimations repères

	Estimations repères	Panel SIPP	Correction logistique			CHAID 1	CHAID 2	Méthode du quotient
			Taux de réponse prévus	Taux de réponse mixtes	Cellules groupées			
Estimations, Panel SIPP 1989								
AFDC	3.56	−1.58	−1.52	−1.43	−1.52	−1.44	−1.84	−1.57
Coupons alimentaires	6.30	1.02	0.82	0.92	0.86	1.01	0.73	0.69
Medicaid	6.97	−0.50	−0.47	−0.40	−0.53	−0.39	−0.70	−0.51
SSI	1.65	0.05	0.11	0.08	−0.03	0.07	−0.15	0.09
Sécurité sociale	15.14	−0.38	−0.46	−0.46	−0.42	−0.44	−0.42	−0.50
Taux de pauvreté	14.46	−2.77	−2.64	−2.57	−2.67	−2.63	−2.78	−2.74
Revenu médian	2,550	2.05	2.01	1.89	2.30	2.30	2.29	2.09
Employés	61.60	2.42	1.60	1.56	1.76	1.72	1.95	1.73
Chômeurs	4.52	−4.93	−4.59	−4.59	−4.76	−4.90	−4.78	−4.60
Inactifs	33.88	−0.42	0.28	0.32	0.18	0.28	−0.01	0.15
Autres estimations repères								
AFDC - Juin 1987	4.28	−2.55	−2.66	−2.49	−2.59	−2.65	−3.14	−2.71
AFDC - Janvier 1989	4.24	−5.71	−5.62	−5.49	−5.63	−5.51	−6.10	−5.70
Coupons alimentaires - Juin 1987	7.35	0.27	−0.31	−0.16	−0.04	0.11	−0.50	−0.48
Coupons alimentaires - Janvier 1989	7.29	−2.04	−2.32	−2.17	−2.26	−2.06	−2.44	−2.50
SSI - Juin 1987	1.68	0.00	0.13	0.08	−0.03	0.08	−0.20	0.11
SSI - Janvier 1989	1.74	−0.57	−0.48	−0.53	−0.67	−0.54	−0.84	−0.50
Salaires annuels 1987	2.22	−16.12	−15.94	−16.38	−15.66	−15.61	−15.60	−15.78
Mariés en 1987	1.86	−5.11	−4.93	−4.98	−5.11	−5.10	−5.07	−4.95
Divorcés en 1987	0.90	−7.15	−7.37	−7.36	−7.40	−7.32	−7.20	−7.40
Changement d'adresse en 1987	17.99	−11.49	−10.50	−10.51	−10.80	−10.42	−10.40	−10.49

en juin 1987, en janvier 1989 et n'importe quand pendant l'année civile 1987. On ne dispose d'une estimation comparable de l'enquête par panel SIPP de 1989 que pour la période de janvier 1989.

Il importe avant tout de noter, dans le tableau 4, la similitude entre les estimations calculées avec les diverses méthodes de pondération à partir des données du SIPP de 1987. Les pourcentages estimatifs du tableau 4 sont en fait présentés avec une précision de 2 décimales puisque la présentation classique à une décimale ne laisserait souvent voir aucune différence entre les valeurs obtenues. La différence la plus importante s'observe dans le cas du pourcentage de personnes détenant un emploi en janvier 1989, où l'estimation utilisant le poids de l'enquête SIPP est de 62.7%, alors que celle obtenue à l'aide du poids de la méthode logistique à taux de réponse mixtes est de 62.3%. Même cette différence, qui est la plus grande observée, est relativement petite, surtout si on songe que l'erreur-type estimée correspondante n'est que de 0.3%.

Lorsqu'on compare les estimations de l'enquête par panel SIPP de 1987 aux estimations externes provenant de l'enquête par panel SIPP de 1989 ou d'autres sources, certaines des différences observées sont plus grandes et beaucoup plus importantes. Pour les examiner d'une manière plus détaillée, nous avons calculé les différences normalisées entre les estimations des méthodes de rechange et les estimations de référence. Les résultats de ces opérations sont présentés au tableau 5. On entend par différence normalisée la différence entre les estimations des méthodes de rechange et l'estimation externe, divisée par l'erreur-type de la différence.

La portion supérieure du tableau 5 présente les différences normalisée obtenues lorsque les données de l'enquête SIPP de 1989 sont utilisées pour produire l'estimation externe. Les différences normalisées pour la plupart des estimations sont inférieures à 2.0 en valeur absolue, ce qui signifie qu'elles peuvent être entièrement attribuables à des erreurs d'échantillonnage. Toutefois, les différences normalisées dans le cas du pourcentage de chômeurs et du taux de pauvreté sont supérieures à 2.0 et donc hautement significatives. Ainsi, les méthodes de rechange pour la correction de la pondération ne permettent pas d'aligner les estimations de l'enquête de 1987 sur celles de l'enquête de 1989 pour toutes les caractéristiques examinées.

La portion inférieure du tableau 5 montre les différences normalisées obtenues lorsque d'autres estimations repères sont utilisées. Ces différences normalisées sont généralement grandes et, dans beaucoup de cas, très grandes. Seules quelques-unes sont inférieures à 2.0, et beaucoup sont supérieures à 10.0. Étant donné les différences normalisées beaucoup plus petites notées dans la portion supérieure du tableau 5 pour des statistiques similaires, il

nous paraît vraisemblable que des facteurs autres que le biais de non-réponse du panel soient largement responsables de l'ampleur de ces différences. Les différences normalisées fondées sur ces sources de données principalement administratives mettent peut-être en lumière certains problèmes importants liés à la qualité des données (des enquêtes SIPP, des sources de données de référence ou des deux), mais elles nous sont de peu d'utilité pour l'évaluation de l'efficacité des méthodes de correction du poids de non-réponse dans la réduction du biais de non-réponse du panel.

5. DISCUSSION

On fait grand usage des poids de non-réponse afin de compenser la non-réponse des unités dans les enquêtes par sondages. L'exigence fondamentale de cette forme de pondération est la disponibilité des informations portant sur une ou plusieurs variables auxiliaires, tant pour les répondants que pour les non-répondants. Dans beaucoup d'enquêtes, ces informations ne sont disponibles que pour un petit nombre de variables auxiliaires (p. ex., UPÉ et strate à partir desquelles les unités ont été sélectionnées). Dans de telles enquêtes, les poids de non-réponse peuvent souvent prendre la forme de corrections de pondérations pour un ensemble de catégories, fondés sur la distribution croisée des variables auxiliaires.

Il existe toutefois des enquêtes dans lesquelles on peut compter sur un grand nombre de variables auxiliaires aux fins de la détermination possible de poids de non-réponse. Cette situation survient souvent lorsqu'on utilise un système de classement des dossiers administratifs en guise de base de sondage de l'enquête, l'ensemble des informations du système étant alors disponible pour la réalisation des corrections du poids de non-réponse. Tel est le cas également lorsque la collecte des données s'effectue en deux phases ou plus (c'est-à-dire, une entrevue initiale de présélection suivie d'une entrevue détaillée ou d'une forme quelconque de collecte de données à une date ultérieure) et lorsque les corrections du poids de non-réponse sont nécessaires pour les phases ultérieures. Dans ce dernier cas, les données des phases antérieures de collecte peuvent servir à la correction de la non-réponse au cours des phases ultérieures. C'est la même situation dans le cas des enquêtes par panel lorsqu'il devient nécessaire de procéder à des corrections pour la non-réponse au cours des vagues ultérieures de l'enquête, comme nous le décrivons dans le présent article.

Lorsqu'on dispose d'un grand nombre de variables auxiliaires pour chaque unité échantillonnée, deux choix doivent être faits. Premièrement, il faut choisir les variables auxiliaires à utiliser pour la correction. Deuxièmement il faut choisir la méthode de correction.

Dans la présente étude, la méthode fondamentale retenue pour le choix des variables auxiliaires à utiliser pour la correction du poids de non-réponse a consisté à déterminer la série de variables qui représentait de bons prédicteurs de la non-réponse du panel. Compte tenu de la gamme de variables auxiliaires disponibles, la première étape consistait à déterminer et à éliminer les variables qui n'étaient pas liées de près au taux de non-réponse du panel. Ensuite, des modèles de régression logistique utilisant les variables prédictives retenues à la présélection ont été examinés afin de déterminer les séries de variables à retenir aux fins de la correction de la pondération. Peu importe que le nombre de variables auxiliaires soit réduit à une taille raisonnable à l'aide de cette méthode ou d'une autre (p. ex., en utilisant l'algorithme CHAID) cette réduction risque d'être une première étape nécessaire lorsqu'on dispose d'un grand nombre de variables auxiliaires possibles.

Après la sélection du sous-ensemble de variables auxiliaires à utiliser, on peut choisir entre diverses méthodes pour la réalisation des corrections du poids de non-réponse. Nous avons examiné les corrections fondées sur les modèles de régression logistique, les modèles de recherches catégoriques et la méthode du quotient généralisée fondée sur l'échantillon. Les poids finals découlant de ces méthodes de correction montraient une très étroite corrélation entre eux et donnaient des estimations très semblables. Aucune des méthodes n'a produit des estimations meilleures que les autres pour la réduction du biais.

Les corrélations étroites observées entre les poids finals générés à l'aide des diverses méthodes de correction expérimentées peuvent s'expliquer en partie par les similitudes entre plusieurs de ces méthodes. Elles peuvent également s'expliquer en partie par la pondération finale de la stratification a posteriori qui a augmenté les corrélations entre les poids. Elle peut finalement s'expliquer en partie par l'absence d'effets d'interactions importants entre les variables auxiliaires. S'il y avait des effets d'interactions importants non inclus dans la modélisation logistique, on pourrait s'attendre à des différences plus grandes entre les poids obtenus par la méthode du quotient généralisée et par la méthode de correction logistique des taux de réponse prévus d'une part, et par la méthode CHAID, la méthode de correction des taux de réponses mixtes et la méthode de correction logistique en cellules groupées d'autre part. Ainsi, la similitude des poids produits à l'aide des diverses méthodes de pondération pour les données du SIPP pourrait ne pas être aussi importante en d'autres circonstances.

Lorsqu'on utilise de nombreuses variables auxiliaires pour ajuster les poids, il arrive souvent que les résultats soient hautement variables, entraînant ainsi une sérieuse perte de précision des estimations de l'enquête. Ce problème ne s'est toutefois pas posé avec les méthodes que nous avons évaluées. La variabilité des poids obtenue avec l'ensemble des méthodes testées s'est avérée semblable, compte tenu du fait que des précautions raisonnables avaient été prises lors de la réalisation des corrections.

Même si les résultats empiriques ne laissent pas voir de différences appréciables entre les estimations produites à l'aide des diverses méthodes de pondération testées et celles produites à l'aide des poids de l'enquête par panel SIPP, les corrélations entre les poids ajustés des méthodes de rechange et celui de l'enquête par panel SIPP se sont avérées inférieures aux corrélations calculées entre les

poids des diverses méthodes de rechange. Ce résultat laisse conclure que le choix des variables auxiliaires est important, et qu'il l'est probablement plus que le choix de la méthode de pondération. Même si les méthodes plus systématiques utilisées dans la présente recherche pour le choix des variables auxiliaires n'ont pas conduit à des améliorations importantes par rapport aux méthodes courantes des enquêtes SIPP, un choix des variables auxiliaires fondé sur l'analyse risque d'être plus productif dans d'autres études.

Lorsqu'on dispose d'un nombre important de variables auxiliaires liées à la propension à répondre, il paraît sage d'en utiliser le plus grand nombre possible aux fins de la correction de la non-réponse, en guise de sauvegarde pour la correction du biais de non-réponse. Cette stratégie générale devrait toutefois être tempérée par une évaluation soigneuse de la variation des poids obtenus, afin d'éviter une trop grande perte de précision dans les estimations de l'enquête. En outre, il conviendrait de tenir compte de la facilité de mise en vigueur des méthodes de pondération. Si, comme c'est le cas dans la présente étude, les méthodes de pondération de rechange donnent des poids et des estimations très semblables, on jugera peut-être préférable de choisir la plus simple.

REMERCIEMENTS

Nous remercions les membres du comité de lecture pour les commentaires utiles qu'ils ont formulés sur la version antérieure du présent article. Cet article porte sur des recherches effectuées pour le U.S. Bureau of the Census. Les opinions exprimées sont celles des auteurs et ne reflètent pas nécessairement celles du Bureau of the Census.

BIBLIOGRAPHIE

CHAPMAN, D.W., BAILEY, L., et KASPRZYK, D. (1986). Méthodes de compensation de la non-réponse au U.S. Bureau of the Census. *Techniques d'enquête*, 12, 161-180.

DEMING, W.E., et STEPHAN, F.F. (1942). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.

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

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

JABINE, T.B., KING, K.E., et PETRONI, R.J. (1990). *Survey of Income and Program Participation (SIPP): Quality Profile*. Washington, DC: U.S. Bureau of the Census.

KALTON, G. (1983). *Compensating for Missing Survey Data*. Ann Arbor, MI: Survey Research Center, University of Michigan.

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

KALTON, G., LEPKOWSKI, J.M., MONTANARI, G.E., et MALIGALIG, D. (1990). Characteristics of second wave nonrespondents in a panel survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 462-467.

KALTON, G., LEPKOWSKI, J., et LIN, T. (1985). Compensating for wave nonresponse in the 1979 ISDP Research Panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 372-377.

KASS, G.V. (1980). An exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, 29, 119-127.

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

LEPKOWSKI, J., KALTON, G., et KASPRZYK, D. (1989). Weighting adjustments for partial nonresponse in the 1984 SIPP Panel. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 296-301.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54, 2, 137-139.

NELSON, D., McMILLEN, D., et KASPRZYK, D. (1985). *An Overview of the SIPP, Update 1*. SIPP Working Paper No. 8401. Washington, DC: U.S. Bureau of the Census.

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

PENNELL, S.G., et LEPKOWSKI, J.M. (1992). Panel conditioning effects in the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 556-571.

RIZZO, L., KALTON, G., et BRICK, M. (1994). Weighting adjustments for panel nonresponse in the SIPP. Rapport final présenté au U.S. Bureau of the Census pour le SIPP Panel Nonresponse Project.

Techniques d'enquête, juin 1996
Vol. 22, n° 1, pp. 55–64
Statistique Canada

Estimation de la population par échantillonnage multiple et sous-dénombrement lors du recensement en présence d'erreurs d'appariement

YE DING et STEPHEN E. FIENBERG[1]

RÉSUMÉ

Les auteurs examinent le recensement par saisie-resaisie multiples en assouplissant l'hypothèse classique d'un appariement parfait. Ils proposent des modèles avec erreur d'appariement permettant de caractériser les méthodes d'appariement sujettes à des erreurs. Les données observées prennent la forme d'un tableau de contingence 2^k auquel manque une cellule et suivent une distribution multinomiale. Les auteurs proposent une méthode pour estimer la population. Cette approche s'applique à la fois aux modèles log-linéaires habituels pour les tableaux de contingence et aux modèles log-linéaires de l'hétérogénéité du potentiel de saisie. Enfin, les auteurs illustrent leur méthode et procèdent à une estimation en recourant à une répétition générale du recensement de 1990, effectuée en 1988 par le U.S. Bureau of the Census.

MOTS CLÉS: Recensement par saisie-resaisie; estimations de la population totale; modèles log-linéaires; erreurs d'appariement; recensement par resaisie multiple.

1. INTRODUCTION

La technique du recensement par resaisie multiple a été appliquée dans de nombreux domaines pour estimer l'importance d'une population fermée. Ainsi, Cormack (1968) et Seber (1982) offrent une excellente revue de nombreuses techniques utilisées. Nous envisagerons ici une séquence d'échantillons, $s_1, \ldots, s_k$, pour laquelle les sujets du i-ième échantillon sont identifiés de façon unique, par exemple par étiquetage ou marquage, avant d'être remis dans la population (Darroch 1958). Les méthodes de recensement par resaisie multiple habituelles reposent sur les hypothèses que voici:

(1) **Appariement idéal.** Les sujets d'une liste (source d'information, échantillon) peuvent être couplés à ceux d'une autre liste sans erreur. En d'autres termes, aucune erreur de classification ne survient lorsqu'on s'efforce de déterminer si un sujet donné a été enregistré par les deux sources d'information ou par une seule d'entre elles.

(2) **Indépendance.** Les listes sont indépendantes l'une de l'autre, c'est-à-dire que la probabilité qu'un sujet fasse partie d'une liste ne dépend pas de l'inclusion du sujet aux listes antérieures.

(3) **Homogénéité (potentiel de saisie identique).** Tous les sujets de la population à l'étude ont la même probabilité d'être observés (saisis) dans une liste (échantillon) quelconque.

(4) **Finitude.** La population est «fermée». Elle ne subit aucun changement consécutivement aux naissances, aux décès, à l'émigration ou à l'immigration au cours de la période durant laquelle s'effectue l'échantillonnage.

Darroch (1958) a examiné le recensement par resaisie multiple sous l'angle de ces quatre hypothèses. De son côté, Fienberg (1972) s'est servi d'un modèle log-linéaire pour autoriser une dépendance statistique entre des types spécifiques de l'échantillon. Il a donc abandonné l'hypothèse d'indépendance. Darroch, Fienberg, Glonek et Junker (1993) ont pour leur part mis au point un modèle log-linéaire élargi qui autorise l'hétérogénéité au niveau individuel et la dépendance, mais exige au moins trois échantillons, à savoir $k = 3$. Plusieurs auteurs se sont penchés sur les problèmes de couplage attribuables aux erreurs d'appariement et de non-appariement dans le contexte du recensement à deux échantillons utilisé par le U.S. Bureau of the Census pour évaluer la couverture du recensement. Ding et Fienberg (1994), par exemple, ont envisagé la modélisation des erreurs d'appariement dans le recensement à deux échantillons et mis au point une méthode systématique pour estimer la population totale. L'inclusion d'un troisième échantillon (tiré des dossiers administratifs) au modèle et l'estimation de la couverture du recensement avaient déjà été envisagées par le Bureau of the Census dans le passé. On n'a pas écarté cette possibilité pour étendre et évaluer l'approche du système dual. Dans le présent article, nous examinerons les modèles d'erreur d'appariement relatifs au problème du recensement à échantillon multiple tolérant les principes de dépendance et d'hétérogénéité.

Nous supposons ici que les observations tirées des données d'un recensement par saisie multiple font partie d'une classification croisée 2^k, avec ou sans le i-ième échantillon définissant la catégorie de la i-ième dimension. En vertu d'un telle classification, la cellule qui correspond

[1] Ye Ding, Research Scientist, Bureau of Biometrics, New York State Health Department, Concourse, Room C-144, Empire State Plaza, Albany, New York 12237, U.S.A.; Stephen E. Fienberg, Maurice Falk Professor of Statistics and Social Science, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.

à l'absence des k échantillons n'existe pas. L'objectif consiste à estimer le nombre de sujets de la population qui ne sont pas observés, donc qui correspondent au chiffre de la cellule manquante dans le tableau de contingence 2^k. À la partie 2, nous examinerons les effets des erreurs d'appariement sur le tableau 2^k incomplet observé. La partie 3 propose quelques modèles pour les erreurs d'appariement en mesure de caractériser une méthode d'appariement sujette aux erreurs. À partir de ces modèles et des hypothèses (3) et (4), nous élaborerons une procédure pour estimer la taille de la population à partir d'un modèle log-linéaire. Enfin, à la partie 5, nous nous servirons des méthodes envisagées pour analyser les données de la répétition générale de 1988 du recensement entreprise par le U.S. Bureau of the Census.

2. LES ERREURS D'APPARIEMENT DANS LES RECENSEMENTS À ÉCHANTILLON MULTIPLE

Nous commencerons par répartir les erreurs d'appariement en deux grandes catégories: les faux appariements et les non-appariements erronés. Pour bien comprendre la nature de ces erreurs dans un recensement à échantillon multiple, prenons le cas d'un recensement à trois échantillons. Supposons qu'aucune donnée ne manque ou qu'il n'y ait aucune erreur dans la saisie de l'information sur un sujet de la population et qu'on prélève trois échantillons de la population, s_1, s_2 et s_3. Supposons, par exemple, qu'on retrouve les sujets 1, 3, 4 et 7 dans l'échantillon s_1, les sujets 3, 4 et 8 dans l'échantillon s_2 et les sujets 4, 9 et 10 dans l'échantillon s_3. Sous sa forme vectorielle, cette situation s'exprime comme suit: $s_1 = (1, 3, 4, 7)$, $s_2 = (3, 4, 8)$ et $s_3 = (4, 9, 10)$. Il n'y a pas d'erreurs d'appariement, pourvu que les renseignements soient complets et exacts. On obtient le tableau incomplet 2^3 qui suit au moyen des trois échantillons:

Tableau 1

Tableau original sans erreur d'appariement

	s_1			
	Présent		Absent	
	s_2		s_2	
s_3	Présent	Absent	Présent	Absent
Présent	1	0	0	2
Absent	1	2	1	–

Supposons par ailleurs qu'en raison de données manquantes ou de renseignements incorrects, on observe en réalité ce qui suit:

$$s_1 = (1, 3, 4, 7), \quad s_2 = (3^*, 4^*, 8), \quad s_3 = (4, 9, 10),$$

où 3* et 4* représentent les sujets 3 et 4, mais associés à des renseignements incorrects, ce qui entraîne deux non-appariements erronés lors du couplage des échantillons. Partant de l'hypothèse que tous les appariements sont exacts, on obtient le tableau 2^3 incomplet que voici:

Tableau 2

Observations avec erreurs d'appariement

	s_1			
	Présent		Absent	
	s_2		s_2	
s_3	Présent	Absent	Présent	Absent
Présent	0	1	0	2
Absent	0	3	3	–

Lorsqu'on compare les deux tableaux, il est facile de voir l'effet des erreurs d'appariement:

(i) Le nombre d'observations peut augmenter dans certaines cellules et diminuer dans d'autres, si bien que les totaux marginaux, en particulier le nombre total de sujets différents observés dans les trois échantillons, peuvent varier, sous réserve de la contrainte que le nombre total d'observations dans les échantillons, x_{1++}, x_{+1+}, et x_{++1}, reste le même. Une modification du nombre total de sujets différents dans les échantillons soulèverait un problème distinct du problème habituel posé par une erreur de classification dans l'analyse des données catégoriques, où on examine la possibilité qu'une erreur se glisse dans la répartition des sujets entre les différentes catégories (p. ex., lire Chen 1979).

(ii) Parallèlement, la probabilité dans certaines cellules peut changer pourvu que la probabilité p_{1++}, p_{+1+}, et p_{1++}, d'être saisie dans un échantillon demeure la même.

En raison de la complexité des erreurs d'appariement du cas à trois échantillons, nous avons besoin d'une terminologie particulière pour faciliter la description. Par rapport à l'échantillon s_1, nous dirons donc qu'un sujet se trouve à l'état 1 si on l'observe dans l'échantillon et à l'état 0 dans le cas contraire. Le triplet (i,j,k), $0 \leq i, j, k \leq 1$ indiquera un sujet à l'état i, j et k, pour les échantillons s_1, s_2 et s_3, respectivement. Ainsi, (1,0,0) dénotera un sujet qu'on observe uniquement dans s_1 et (1,1,1), un sujet saisi dans les trois échantillons. Le niveau d'un sujet (i,j,k) sera $i + j + k$, soit le nombre d'échantillons dans lequel se retrouve le sujet. Il existe donc quatre niveaux différents, 0, 1, 2 et 3. Un sujet n'aura le niveau 0 que s'il n'apparaît dans aucun échantillon, alors qu'il sera de niveau 3 quand il est présent dans les trois échantillons. Si le couplage est incorrect selon la règle d'appariement, le sujet (1,1,0) se décomposera en «deux sujets différents», (1,0,0) et (0,1,0), en supposant qu'il n'y ait pas de faux appariement. Par ailleurs, le sujet (1,0,0) apparié par

erreur au sujet (0,1,0) donnera une seule observation (1,1,0). Pour plus de commodité, nous appellerons une telle décomposition ou combinaison une *transition*. Les transitions ne peuvent aller que du niveau 3 ou du niveau 2 au même niveau (s'il n'y a pas d'erreur d'appariement) ou à un niveau inférieur, s'il n'y a pas de faux appariements. Plus précisément, un sujet (1,1,1) peut faire la transition vers l'un des 5 jeux possibles de sujets que voici:

$$\{(1,1,1)\}, \quad \{(1,0,0), (0,1,1)\}, \quad \{(0,1,0), (1,0,1)\}$$

$$\{(0,0,1), (1,1,0)\}, \quad \{(1,0,0), (0,1,0), (0,0,1)\}.$$

Au niveau 2, un sujet (1,1,0) pourra se décomposer en $\{(1,0,0),(0,1,0)\}$ ou demeurer $\{(1,1,0)\}$. Il en ira autant des sujets $\{(0,1,1)\}$ et $\{(1,0,1)\}$. De ce qui précède, on peut résumer l'effet d'une erreur d'appariement au moyen du diagramme suivant:

Tableau 1 → {Processus d'appariement} → Tableau 2

où le tableau 1 représente le tableau 2^k incomplet original, sans erreur d'appariement, et le tableau 2 correspond au tableau 2^k incomplet observé, avec erreur d'appariement. Par conséquent, la probabilité par cellule et le chiffre prévu d'une cellule du tableau 1 seront notés par $\{r_{ijk}\}$ et $\{l_{ijk}\}$, et par $\{p_{ijk}\}$ et $\{m_{ijk}\}$ pour le tableau 2, où $1 \leq i,j,k \leq 2$.

3. QUELQUES MODÈLES POUR LES ERREURS D'APPARIEMENT

Nous présenterons maintenant des modèles qui décrivent les erreurs d'appariement. Chacun d'eux nous permet de formuler la réattribution de la probabilité et du chiffre prévu des cellules associés au tableau 1.

Modèle (1). Outre les hypothèses d'homogénéité et de finitude expliquées en §1, nous supposons: (i) que le couplage n'entraîne pas de faux appariements; (ii) que la probabilité qu'un sujet garde son état original est égale à θ et que la probabilité d'effectuer une transition vers un ensemble possible de sujets est égale à $(1 - \theta)/(m - 1)$, où m représente le nombre de jeux possibles de sujets pour lesquels la transition est autorisée. Pour le sujet (1,1,1) examiné à la dernière partie, par exemple, $m = 5$.

Un tel modèle permet d'exprimer les probabilités du tableau avec erreur d'appariement, $\{p_{ijk}\}$, d'après les probabilités du tableau sans erreur d'appariement, $\{r_{ijk}\}$, pour le recensement à trois échantillons de la façon suivante:

$$p_{111} = \theta r_{111},$$

$$p_{112} = \frac{1-\theta}{4} r_{111} + \theta r_{112},$$

$$p_{121} = \frac{1-\theta}{4} r_{111} + \theta r_{121},$$

$$p_{211} = \frac{1-\theta}{4} r_{111} + \theta r_{211},$$

$$p_{122} = \frac{1-\theta}{2} r_{111} + (1-\theta) r_{112} + (1-\theta) r_{121} + r_{122},$$

$$p_{212} = \frac{1-\theta}{2} r_{111} + (1-\theta) r_{112} + (1-\theta) r_{211} + r_{212},$$

$$p_{221} = \frac{1-\theta}{2} r_{111} + (1-\theta) r_{211} + (1-\theta) r_{121} + r_{221}.$$

Soit

$$\vec{p} = (p_{111}, p_{112}, p_{121}, p_{211}, p_{122}, p_{212}, p_{221})^T,$$

et

$$\vec{r} = (r_{111}, r_{112}, r_{121}, r_{211}, r_{122}, r_{212}, r_{221})^T,$$

donc

$$\vec{p} = M_1 \times \vec{r}. \tag{1}$$

Dans ce cas, M_1 représente une matrice 7×7 définie par les sept équations qui précèdent, issues du modèle (1). Il est assez facile de vérifier que la probabilité de saisir un sujet dans chaque échantillon est fixe, à savoir $p_{1++} = r_{1++} = p_1, p_{+1+} = r_{+1+} = p_2, p_{++1} = r_{++1} = p_3$. Il doit en être ainsi puisque la probabilité de saisie dans l'échantillon ne dépend pas de la façon dont fonctionne la méthode d'appariement.

On peut aisément généraliser la formule pour k échantillons. Les opérations algébriques deviennent néanmoins fort fastidieuses quand k est élevé. Il est possible de simplifier le modèle en exigeant que les transitions descendent d'au moins un niveau, ce qui nous amène au modèle (2).

Modèle (2). En plus des hypothèses d'homogénéité et de finitude données en §1, on suppose: (i) que le couplage n'entraîne pas de faux appariements; (ii) qu'une transition ne peut se faire que vers le bas, d'au moins un niveau; (iii) que la probabilité qu'un sujet garde son état original est θ et que la probabilité d'une transition vers un ensemble possible de sujets est $(1 - \theta)/(m' - 1)$, où m' représente le nombre de combinaisons de sujets vers lesquels la transition est possible et autorisée.

Examinons d'abord le cas à trois échantillons. Un sujet (1,1,1) peut se décomposer en trois individus, soit $(1,1,1) \mapsto \{(1,0,0), (0,1,0), (0,0,1)\}$ (le signe « $\mapsto$ » indique qu'il y a décomposition), si les trois appariements présumés ne se concrétisent pas. L'hypothèse (ii) du modèle (2) établit que cette triple erreur a une probabilité négligeable de se reproduire, comparativement à la probabilité d'une transition où un seul appariement n'est pas effectué, si bien que $(1,1,1) \mapsto \{(1,1,0),(0,0,1)\}$, ou $(1,1,1) \mapsto \{(1,0,1),(0,1,0)\}$, ou $(1,1,1) \mapsto \{(1,1,0),(0,0,1)\}$.

Le modèle paramétrique qui exprime $\{p_{ijk}\}$ en termes de $\{r_{ijk}\}$ pour le cas à trois échantillons est le suivant:

$$p_{111} = \theta r_{111},$$

$$p_{112} = \frac{1-\theta}{3} r_{111} + \theta r_{112},$$

$$p_{121} = \frac{1-\theta}{3} r_{111} + \theta r_{121},$$

$$p_{211} = \frac{1-\theta}{3} r_{111} + \theta r_{211},$$

$$p_{122} = \frac{1-\theta}{3} r_{111} + (1-\theta) r_{112} + (1-\theta) r_{121} + r_{122},$$

$$p_{212} = \frac{1-\theta}{3} r_{111} + (1-\theta) r_{112} + (1-\theta) r_{211} + r_{212},$$

$$p_{221} = \frac{1-\theta}{3} r_{111} + (1-\theta) r_{211} + (1-\theta) r_{121} + r_{221}.$$

Par conséquent

$$\vec{p} = M_2 \times \vec{r}, \tag{2}$$

où M_2 est une matrice 7×7 définie par les sept équations qui précèdent, dérivées du modèle (2). Encore une fois, la probabilité de saisie reste la même, c.-à-d. $p_{1++} = r_{1++} = p_1$, $p_{+1+} = r_{+1+} = p_2$, $p_{++1} = r_{++1} = p_3$.

Pour le problème à k échantillons, soit $p_{\bar{1}}$, la probabilité d'être saisi dans chaque échantillon, à savoir $p_{\bar{1}} = p_{111\ldots1}$, et soit $p_{\bar{1},\bar{2}(h_1,h_2)}$, la probabilité par cellule correspondant à une absence dans le h_1-ième et le h_2-ième échantillons, et d'une présence dans les autres échantillons, *etc.* En vertu du modèle (2), $p_{\bar{1}} = \theta r_{\bar{1}}$. Lorsque $i \leq k-2$, la probabilité de ne pas apparaître dans les h_1-ième, h_2-ième, ..., et h_i-ième échantillons et saisis dans les autres échantillons est égale à

$$p_{\bar{1},\bar{2}(h_1,h_2,\ldots,h_i)} = \theta r_{\bar{1},\bar{2}(h_1,h_2,\ldots,h_i)} + \frac{1-\theta}{k-i+1} \sum_{j=1}^{i} r_{\bar{1},\bar{2}(\{h_1,h_2,\ldots,h_i\}\setminus h_j)}.$$

Quand $i = k-1$, le sujet ne fait partie que d'un échantillon. Par exemple, la probabilité de n'être saisi que dans le premier échantillon est égale à

$$p_{1,\bar{2}} = r_{1,\bar{2}} + (1-\theta) \sum_{h \neq 1} r_{1,1(h),\bar{2}} + \frac{(1-\theta)}{3} \sum_{h_1,h_2 \geq 2} r_{1,1(h_1,h_2),\bar{2}} + \sum_{j=3}^{k-1} \sum_{h_1,h_2,\ldots,h_j \geq 2} \frac{(1-\theta)}{(j+1)} r_{1,1(h_1,h_2,\ldots,h_j),\bar{2}},$$

où $r_{1,1(h_1,h_2,\ldots,h_j),\bar{2}}$ représente la probabilité par cellule du tableau original, correspondant à une présence dans le premier h_1-ième, h_2-ième, ..., h_j-ième échantillon et à une absence dans les autres échantillons. Pour des raisons de symétrie, on peut écrire l'expression pour $p_{1(h),\bar{2}}$, c.-à-d. la probabilité d'être observé dans le h-ième échantillon uniquement et pas dans les autres.

Il est possible d'améliorer le modèle (2) en supposant des taux d'appariement inégaux. Examinons, par exemple, deux décompositions: $(1,1,1) \mapsto \{(1,1,0),(0,0,1)\}$ et $(1,1,0) \mapsto \{(0,1,0),(1,0,0)\}$. Dans les deux cas, le point commun est qu'un appariement présumé ne s'est pas réalisé. La différence réside dans le fait que, dans le premier cas, il y a couplage de deux sources d'information, comparativement à une seule dans le second. On peut raisonnablement supposer que la probabilité d'une erreur d'appariement diffère dans les deux cas, contrairement à ce que suppose le modèle (2). Ceci nous amène donc au:

Modèle (3). En plus des hypothèses (i) et (iii) du modèle (2), on suppose

$$(1,1,1) \mapsto \begin{cases} (1,1,1) & \text{avec une probabilité } \alpha_1 \\ \{(1,1,0),(0,0,1)\} & \text{avec une probabilité } (1-\alpha_1)/3 \\ \{(0,1,1),(1,0,0)\} & \text{avec une probabilité } (1-\alpha_1)/3 \\ \{(1,0,1),(0,1,0)\} & \text{avec une probabilité } (1-\alpha_1)/3 \end{cases}$$

$$(1,1,0) \mapsto \begin{cases} (1,1,0) & \text{avec une probabilité } \alpha_2 \\ \{(0,1,0),(1,0,0)\} & \text{avec une probabilité } 1-\alpha_2 \end{cases}$$

$$(1,0,1) \mapsto \begin{cases} (1,0,1) & \text{avec une probabilité } \alpha_2 \\ \{(1,0,0),(0,0,1)\} & \text{avec une probabilité } 1-\alpha_2 \end{cases}$$

$$(0,1,1) \mapsto \begin{cases} (0,1,1) & \text{avec une probabilité } \alpha_2 \\ \{(0,1,0),(0,0,1)\} & \text{avec une probabilité } 1-\alpha_2 \end{cases}$$

et (1,0,0), (0,1,0) et (0,0,1) ne varient pas avec une probabilité égale à un.

Selon ce modèle, on peut exprimer la probabilité par cellule $\{p_{ijk}\}$ du tableau 2 en fonction de α_1, α_2 et de la probabilité par cellule du tableau 1, $\{r_{ijk}\}$. Pour cela, il faut tenir compte de toutes les transitions possibles qui donneront un sujet dans la cellule (i,j,k) du tableau 2. Par exemple, prenons le sujet observé (1,0,0). Ce sujet se retrouve dans la cellule (1,2,2) du tableau 2. Soit F, l'observation d'un sujet de statut (1,0,0) et soit E_{ijk}, l'inclusion d'un sujet dans la cellule (i,j,k) du tableau 1. Alors,

$$F = \bigcup_{\{i,j,k\}} (E_{ijk} \cap F).$$

Selon le modèle (3), il n'existe que quatre transitions en vertu desquelles F se réalise, en l'occurrence:

$$(1,1,1) \mapsto \{(1,0,0),(0,1,1)\},$$

$$(1,1,0) \mapsto \{(1,0,0),(0,1,0)\},$$

$$(1,0,1) \mapsto \{(1,0,0),(0,0,1)\},$$

$$(1,0,0) \mapsto \{(1,0,0)\}.$$

Par conséquent,

$$F = (E_{111} \cap F) \cup (E_{112} \cap F) \cup (E_{121} \cap F) \cup (E_{122} \cap F).$$

Étant donné la définition de la probabilité par cellule des deux tableaux, $p(F) = p_{122}$ et $p(E_{ijk}) = r_{ijk}$. En vertu des hypothèses du modèle (3), $p(F \mid E_{111}) = (1 - \alpha_1)/3$, $p(F \mid E_{112}) = p(F \mid E_{121}) = \alpha_2$ et $p(F \mid E_{122}) = 1$.

Puisque $E_{111} \cap F$, $E_{112} \cap F$, $E_{121} \cap F$ et $E_{122} \cap F$ sont quatre possibilités qui s'excluent mutuellement pour que F se réalise,

$$\begin{aligned} p_{122} &= p(E_{111} \cap F) + p(E_{112} \cap F) \\ &\quad + p(E_{121} \cap F) + p(E_{122} \cap F) \\ &= p(F \mid E_{111}) \cdot p(E_{111}) + p(F \mid E_{112}) \cdot p(E_{112}) \\ &\quad + p(F \mid E_{121}) \cdot p(E_{121}) + p(F \mid E_{122}) \cdot p(E_{122}) \\ &= \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{112} + (1 - \alpha_2) r_{121} + r_{122}. \end{aligned}$$

On peut dériver les autres probabilités par cellule du tableau 2 de la même manière, si bien qu'on obtient

$$p_{111} = \alpha_1 r_{111},$$

$$p_{112} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{112},$$

$$p_{121} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{121},$$

$$p_{211} = \frac{1 - \alpha_1}{3} r_{111} + \alpha_2 r_{211},$$

$$p_{122} = \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{112} + (1 - \alpha_2) r_{121} + r_{122},$$

$$p_{212} = \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{112} + (1 - \alpha_2) r_{211} + r_{212},$$

$$p_{221} = \frac{1 - \alpha_1}{3} r_{111} + (1 - \alpha_2) r_{211} + (1 - \alpha_2) r_{121} + r_{221}.$$

Donc

$$\vec{p} = M_3 \times \vec{r}, \tag{3}$$

où M_3 représente une matrice 7×7 définie par les sept équations qui précèdent, issues du modèle (3).

Lorsque $\alpha_1 = \alpha_2 = \theta$, on obtient la même formule qu'avec le modèle (2). Dans le cas particulier où $\alpha_1 = \alpha_2 = 1$, $p_{ijk} = r_{ijk}$, ce qui nous ramène au problème classique. Encore une fois, les probabilités de saisie demeurent les mêmes, soit $p_{1++} = r_{1++}$, $p_{+1+} = r_{+1+}$, $p_{++1} = r_{++1}$.

4. ESTIMATION DE LA TAILLE DE LA POPULATION

4.1 Formulation du modèle log-linéaire

Aux fins d'exposition, nous nous limiterons au cas du recensement à trois échantillons, bien que l'extension à k échantillons où $k > 3$ se fasse simplement. Comme nous l'avons fait précédemment, soient l_{ijk} et m_{ijk}, le chiffre par cellule prévu pour les tableaux 1 et 2, respectivement. La relation entre la probabilité par cellule et le chiffre prévu par cellule est $l_{ijk} = r_{ijk}N$ et $m_{ijk} = p_{ijk}N$. Soient

$$\vec{m} = (m_{111}, m_{112}, m_{121}, m_{211}, m_{122}, m_{212}, m_{221})^T,$$

et

$$\vec{l} = (l_{111}, l_{112}, l_{121}, l_{211}, l_{122}, l_{212}, l_{221})^T.$$

Puisque, pour chacun des modèles proposés à la dernière partie, il existe une matrice M dont les inscriptions dépendent des paramètres de la probabilité d'appariement pour le modèle retenu, ce qui $\vec{p} = M \times \vec{r}$, multiplié par N donne

$$\vec{m} = M \times \vec{l}. \tag{4}$$

La paramétrisation de m_{ijk} est simple avec tous les modèles log-linéaires indiqués pour le tableau 1. Ainsi, pour chacun des modèles suggérés par Fienberg (1972), on peut exprimer le chiffre prévu au moyen des fonctions des paramètres du terme u:

$$l_{ijk} = g_{ijk}(u, u_1(i), u_2(j), u_3(k), u_{12}(ij), u_{13}(ik), u_{23}(jk)), \tag{5}$$

On peut ensuite dériver les paramètres de $\{m_{ijk}, (ijk) \neq (222)\}$ à partir de (4).

4.2 Estimation de la taille de la population

Examinons maintenant le taux d'appariement des divers modèles. Pour estimer la population, on procède de la façon suivante. Tout d'abord, à l'instar de Sanathanan (1972), on estime le maximum de vraisemblance des paramètres du terme u au moyen de l_c, la vraisemblance conditionnelle associée au tableau 2 étant donné n,

$$l_c = n! \prod_{\{(ijk)\neq(222)\}} \frac{(q_{ijk})^{x_{ijk}}}{x_{ijk}!},$$

où $n = \sum_{\{(ijk)\neq(222)\}} x_{ijk}$ et $q_{ijk} = m_{ijk}/n$. Sanathanan (1972) montre que dans des conditions de régularité convenables, les estimations conditionnelles et inconditionnelles du maximum de vraisemblance sont à la fois cohérentes et présentent la même distribution normale asymptotique. Quand on supprime les paramètres redondants du terme u au moyen des contraintes liées au modèle log-linéaire du tableau 1, la difficulté consiste à établir le maximum de l_c, sous réserve de la seule contrainte que voici:

$$\sum_{\{(ijk)\neq(222)\}} m_{ijk} = n.$$

Sur le plan numérique, il s'agit d'un problème d'optimisation sans limite linéaire. Rao (1957) a examiné les conditions de régularité qui débouchent sur une estimation unique du maximum de vraisemblance pour les paramètres, avec une distribution multinomiale. Les conditions qu'il a établies sont satisfaites par la paramétrisation de $\{q_{ijk}\}$. Une fois que les estimations conditionnelles du maximum de vraisemblance des paramètres du terme u sont connues, on se sert du modèle log-linéaire établi pour le tableau 1 afin d'obtenir les estimations conditionnelles du maximum de vraisemblance pour $\{l_{ijk}\}$, soit le chiffre prévu par cellule du tableau 1, y compris le chiffre prévu pour la cellule manquante. L'estimation de N est donc

$$\hat{N} = \sum_{\{ijk\}} \hat{l}_{ijk}.$$

S'il n'y a pas d'erreur d'appariement, puisque $\alpha_1 = \alpha_2 = 1$ au modèle (3), $m_{ijk} = l_{ijk}$. Par conséquent,

$$\hat{N} = n + \hat{m}_{222},$$

ce qui revient à la méthode d'estimation du problème classique du recensement à resaisie multiple énoncé par Fienberg (1972) avec les modèles log-linéaires proposés par cet auteur.

Comme on l'a indiqué précédemment, on a spécifié un modèle log-linéaire pour le tableau 1 et on considère que les observations s'inscrivent au tableau 2, dont le modèle paramétrique pour le chiffre prévu par cellule est défini par le modèle log-linéaire et par le modèle retenu pour les erreurs d'appariement. Pour savoir si un modèle log-linéaire spécifié pour le tableau 1 convient, on peut appliquer les tests habituels d'adéquation de Pearson et du ratio de vraisemblance, X^2 et G^2, décrits dans Fienberg (1972), au tableau 2. Chaque statistique se caractérise par une distribution χ^2 asymptotique selon l'hypothèse nulle que le modèle convient avec $2^k - 1 -$ (nombre de paramètres indépendants du modèle) degrés de libertés.

5. ANALYSE DES DONNÉES DE LA RÉPÉTITION GÉNÉRALE DU RECENSEMENT DE 1988 ST. LOUIS

Le U.S. Bureau of the Census se sert de l'estimation à système dual (ESD) articulée sur le recensement type à deux échantillons pour évaluer la couverture du recensement depuis 1950. En 1988, le Bureau a procédé à une répétition générale en prévision du recensement décennal de 1990 à trois endroits: St. Louis (Missouri), Columbia (Missouri) et l'ouest de l'État de Washington. Zaslavsky et Wolfgang (1993) présentent les données d'un sous-groupe de la population extrait de l'enquête postcensitaire (EPC) de la répétition générale de St. Louis. Ce sous-groupe est essentiellement composé d'adultes de race noire et de sexe masculin vivant en milieu urbain, dont on croit la population sous-estimée par les méthodes à système dual. Les données résultantes viennent de trois sources: l'échantillon C est celui du recensement proprement dit; l'échantillon P dérive de l'EPC et on a recouru à une troisième source d'information, en l'occurrence la liste administrative complémentaire (LAC) tirée des dossiers administratifs de divers services d'État et services fédéraux (y compris sécurité de l'emploi, permis de conduire, impôt, service sélectif et administration des anciens combattants) avant le recensement. L'échantillon C et l'échantillon P procurent les données nécessaires à l'ESD habituelle ou à l'approche par saisie et resaisie. On peut combiner les données de la LAC à celles du recensement et de l'échantillon P pour effectuer une analyse sous l'angle d'un triple échantillon, même si le but original de l'exercice était d'améliorer la couverture de l'échantillon P. Le tableau 3 présente les données à trois échantillons relatives à la strate d'échantillonnage 11 de l'EPC pour St. Louis, obtenues par groupement des données originales du tableau 1 de Zaslavsky et Wolfgang (1993) pour quatre strates a posteriori définies par les paramètres propriétaire/locataire × 20-29, 30-44 ans.

Tableau 3

Données à trois échantillons pour la strate 11 de St. Louis

LAC	Recensement			
	Présent échantillon P		Absent échantillon P	
	Présent	Absent	Présent	Absent
Present	300	51	53	180
Absent	187	166	76	–

On peut analyser les données d'un tel système triple au moyen du modèle (2) des erreurs d'appariement et des données d'une étude sur l'erreur d'appariement (EEP ou étude de réappariement) distincte, associées aux mêmes strates échantillonnées a posteriori. L'EEP est l'une des opérations effectuées par le Bureau of the Census pour évaluer l'EPC. On l'applique typiquement à un échantillon de cas en recourant à des procédures plus développées et à du personnel hautement qualifié, et en procédant à de nouvelles interviews afin d'estimer le biais associé au processus de couplage antérieur. Dans leur analyse de l'étude sur l'erreur d'appariement entreprise lors d'un recensement expérimental, en 1986, à Los Angeles, Hogan et Wolter (1988) signalent que le réappariement avait été effectué indépendamment de l'appariement original et qu'on avait établi l'écart entre les résultats de l'appariement et du réappariement. À cause de l'approche intensive utilisée, les auteurs estiment que les résultats du réappariement illustrent un couplage véritable, et que les différences entre les données des deux exercices représentent le biais qui fausse les résultats de l'appariement original. Par conséquent, les données de l'EEP permettent d'estimer le taux d'erreur lors de l'appariement initial. Mulry, Dajani et Biemer (1989) ont rédigé un rapport sur l'EEA pour la répétition générale de 1988, dans lequel ils fournissent les résultats du réappariement pour les trois sites d'essai. Le tableau 4 reproduit les données du rapport qui nous intéressent.

Tableau 4

Étude de réappariement de St. Louis: échantillon *P*
Source: Mulry, Dajani et Biemer (1989)

Classification à l'appariement initial	Classification au réappariement			
	Apparié	Non apparié	Indéterminé	Total
Apparié	2,667	7	8	2,682
Non apparié	9	427	30	466
Indéterminé	0	7	20	27
Total	2,676	441	58	3,175

Soit α, le taux d'appariement entre les échantillons *C* et *P* et $\gamma = 1 - \alpha$, le taux d'erreur du non-appariement. Nous présumons que le réappariement ne comporte aucune erreur. À partir des données du tableau 4, on peut estimer α avec $\hat{\alpha} = 2667/(2667 + 9) = 99.6637\%$ et γ avec $\hat{\gamma} = 1 - \hat{\alpha} = .3363\%$. Le paramètre θ représente le taux d'appariement à trois échantillons pour l'échantillon *C*, l'échantillon *P* et la LAC. Deux appariements sont nécessaires à une classification correcte (1,1,1) des trois échantillons, par exemple le premier entre l'échantillon *C* et l'échantillon *P* et le second entre l'échantillon *P* et la LAC. Puisque l'appariement entre le recensement et la LAC n'a pas été évalué, nous supposons que les deux couplages sont indépendants l'un de l'autre et que le taux d'appariement entre l'échantillon *P* et la LAC est identique à celui entre l'échantillon *C* et l'échantillon *P*. On peut donc utiliser $\theta = \alpha^2$ et $\hat{\theta} = \hat{\alpha}^2 = 99.3285\%$. D'après d'autres renseignements qualitatifs, le taux d'appariement paraît anormalement élevé. En outre, le taux d'erreur de l'appariement entre le recensement et la LAC est sans doute plus élevé que le taux d'erreur de l'appariement entre le recensement et l'échantillon *P*. Faute de meilleurs renseignements quantitatifs, nous avons néanmoins décidé de les utiliser dans les calculs qui suivent.

Tableau 5

Estimations selon divers modèles

Modèle log-linéaire	EP ordinaire		EP avec modèle à erreur d'appariement (2)	
	$\hat{N}$ (E.-T.)	Ajustement (d.l.)	$\hat{N}$ (E.-T.)	Ajustement (d.l.)
[C] [P] [A]	1091.48 (11.24)	248.31 (3)	1083.58 (10.93)	244.56 (3)
[CP] [A]	1204.14 (23.31)	90.60 (2)	1194.73 (22.86)	87.30 (2)
[PA] [C]	1108.34 (13.77)	247.93 (2)	1100.03 (13.40)	244.53 (2)
[CA] [P]	1068.87 (10.47)	230.66 (2)	1061.09 (10.10)	226.42 (2)
[CP] [CA]	1271.11 (52.55)	87.16 (1)	1256.77 (50.97)	84.37 (1)
[CP] [PA]	1598.88 (106.26)	17.55 (1)	1585.03 (104.93)	15.88 (1)
[CA] [PA]	1080.47 (13.38)	230.43 (1)	1072.19 (12.88)	226.44 (1)
[CP] [CA] [PA]	2360.82 (363.25)	– (0)	2309.55 (352.36)	– (0)

Le tableau 5 donne l'estimation de la taille de la population pour divers modèles log-linéaires, avec l'estimation de l'erreur-type et des statistiques d'adéquation. On a calculé l'erreur-type selon la méthode delta envisagée par Fienberg (1972). L'hypothèse d'indépendance entre le recensement et l'échantillon *P* a été contestée dans le cadre de l'ESD. La méthode du système dual ne nous permet de vérifier cette hypothèse et d'ajuster la dépendance potentielle que de façon limitée, alors que ces deux aspects peuvent être réglés au moyen de modèles log-linéaires pour trois échantillons ou plus. Quatre modèles énumérés au tableau 5 supposent l'indépendance entre le recensement et l'échantillon *P*: le modèle de totale indépendance [C] [P] [A], [PA] [C], [CA] [P], et [CA] [PA]. Aucun d'eux ne permet un bon ajustement des données. Les trois modèles avec terme d'interaction entre le recensement et l'échantillon *P*, [CP] [A], [CP] [CA] et [CP] [PA], assurent un meilleur ajustement. L'addition d'un terme d'interaction reliant le recensement à la LAC n'améliore que légèrement l'ajustement du modèle [CP] [CA] par rapport à [CP] [A], signe que le recensement et l'échantillon *P* sont presque indépendants de la LAC. Le modèle [CP] [PA] est celui qui procure le meilleur ajustement, ce qui semble indiquer que l'hypothèse habituelle d'indépendance de l'ESD n'est pas valable et qu'il existe un lien entre l'échantillon *P* et la LAC. L'ajustement obtenu avec le modèle (2) de l'erreur d'appariement est légèrement meilleur pour les sept modèles log-linéaires insaturés, mais à peine, en raison du fort taux d'appariement des données de la répétition générale du recensement américain de 1988. Pour le modèle [CP] [PA],

on observe un écart de 0.8738% dans l'estimation de N combiné à un taux de non-appariement de 0.3363%. En supposant que la différence dans l'estimation de N varie de façon à peu près linéaire avec le taux de non-appariement, si le taux de non-appariement avait été de 10%, à savoir s'il y avait eu appariement à 90%, l'écart entre l'estimation habituelle du maximum de vraisemblance de N et notre propre estimation aurait été de 26%.

Tableau 6

Données du système dual pour la strate 11 de St. Louis

Échantillon P	Recensement		
	Présent	Absent	Total
Présent	487	129	616
Absent	217	-	
Total	704		

Le tableau 6 présente les données habituelles du système dual pour la strate 11 de St. Louis. Le nombre de gens dénombrés lors du recensement et dans l'échantillon P est $y_{11} = 300$, le nombre de personnes dans le recensement seulement est $y_{12} = 217$ et le nombre de sujets dans l'échantillon P seulement est $y_{21} = 129$. Le dénombrement total s'élève à $y_{1+} = y_{11} + y_{12} = 704$ pour le recensement et à $y_{+1} = y_{11} + y_{21} = 616$ pour l'échantillon P. L'estimation du système dual est $\widehat{\text{DSE}} = y_{1+}y_{+1}/y_{11} = 893$ (p. 232, Bishop, Fienberg et Holland 1975) et la variance estimée de $\widehat{\text{DSE}}$ est $\text{Var}(\widehat{\text{DSE}}) = y_{1+}y_{+1}y_{12}y_{21}/y_{11}^3 = 105.4$ (p. 233, Bishop et coll. 1975). L'erreur-type s'établit à $\text{SE}(\widehat{\text{DSE}}) = 10.27$.

Le sous-dénombrement pour l'estimation de la population $\widehat{\text{DSE}}$ est $(\widehat{\text{DSE}} - y_{1+})/\widehat{\text{DSE}} \times 100\% = 21.17\%$. Avec le modèle permettant le meilleur ajustement, le sous-dénombrement s'élève à $(\hat{N} - y_{1+})/\hat{N} = 55.97\%$ pour l'estimation $\hat{N} = 1599$, en supposant l'absence d'erreur d'appariement, et à 55.58% pour $\hat{N} = 1585$, selon le modèle (2) de l'erreur d'appariement. Il existe donc un biais à la hausse de $55.97\% - 55.58\% = 0.39\%$ quand on ne tient pas compte des erreurs d'appariement. Ce résultat s'approche beaucoup des 0.37% calculés par Ding et Fienberg (1994) à partir des données du recensement expérimental de 1986 à Los Angeles au moyen d'un taux d'appariement pour deux échantillons de 99.4734%. Le taux d'appariement utilisé ici avec les données de St. Louis est de 99.6637%. Nos estimations révèlent que la population d'adultes de sexe masculin et de race noire vivant en milieu urbain, choisie pour la répétition générale de St. Louis, a été l'objet d'un sous-dénombrement important lors du recensement et que l'estimateur habituel à système dual ou à saisie et resaisie sous-estime gravement le sous-dénombrement. Un troisième échantillon, qualitativement différent, pourrait donner de bons résultats pour ce groupe démographique.

Des probabilités de saisie homogènes font partie des hypothèses de l'approche type à l'estimation d'une population fermée. Darroch et ses collaborateurs (1993) ont mis au point un modèle quasi-symétrique et un modèle à quasi-symétrie partielle permettant de faire varier le potentiel de saisie des sujets. Le premier modèle suppose que les trois échantillons présentent le même genre d'hétérogénéité, tandis que le modèle à quasi-symétrie partielle présume que deux échantillons ont la même hétérogénéité mais pas le troisième. Il s'agit d'un modèle raisonnable puisque le troisième échantillon est qualitativement très différent du recensement et de l'EPC. Ce modèle équivaut à une combinaison de la dépendance et de l'hétérogénéité. Pour les probabilités par cellule multinomiale, y compris la cellule manquante $R = (r_{111}, r_{112}, \ldots, r_{222})$, il s'agit de deux modèles log-linéaires de type $R = A\beta$ assortis d'une matrice de plan d'expérience A appropriée et d'un vecteur des paramètres β. Darroch et ses collaborateurs (1993) donnent les matrices de plan d'expérience pour les deux modèles.

Tableau 7

Modèles à potentiel de saisie hétérogène

Modèle log-linéaire	EP de Darroch et coll. (1993)		EP du modèle à erreur d'appariement (2)	
	$\hat{N}$ (E.-T.)	Ajustement (d.l.)	$\hat{N}$ (E.-T.)	Ajustement (d.l.)
Quasi-symétrie totale	1923.63 (216.84)	133.54 (2)	1906.61 (213.47)	133.50 (2)
Quasi-symétrie partielle	2576.54 (413.28)	11.70 (1)	2557.08 (409.39)	11.72 (1)

La méthode que nous proposons intègre aisément le potentiel de saisie hétérogène, ce qui permet d'estimer la taille de la population, en supposant un modèle d'hétérogénéité pour le tableau 1 et en adoptant l'estimation conditionnelle de vraisemblance (Sanathanan 1972). Le tableau 7 présente les estimations obtenues après ajustement du modèle quasi-symétrique et du modèle à quasi-symétrie partielle aux données de la strate 11. Encore une fois, les erreurs d'appariement de l'analyse n'ont pas d'effet sensible, en raison du taux d'appariement élevé. Le modèle à quasi-symétrie partielle permet un meilleur ajustement que le modèle quasi-symétrique, signe que l'hétérogénéité est plausible et que la LAC paraît avoir une hétérogénéité différente. Le fait que le modèle à indépendance totale s'ajuste mal pourrait aussi s'expliquer en partie par la dépendance entre les échantillons (en particulier entre le recensement et l'échantillon P), et en partie par le potentiel de saisie hétérogène.

Le modèle à quasi-symétrie partielle intègre la dépendance [CP], si bien qu'il constitue une solution de rechange au modèle [CP][PA] du tableau 5. Les deux modèles permettent un ajustement analogue des données, mais aboutissent à des estimations radicalement différentes de N, le modèle qui incorpore l'hétérogénéité produisant une estimation nettement plus élevée, assortie d'une estimation de l'erreur-type estimative beaucoup plus importante. Il se pourrait donc que les paramètres d'hétérogénéité soient très instables et, bien que les deux modèles ne soient pas

hiérarchiques, donc ne puissent se comparer directement, il semble raisonnable de retenir l'estimation la plus faible et la plus stable qui n'intègre pas l'hétérogénéité.

Darroch et coll.(1993) ont analysé quatre sous-strates de la strate 11. Les deux variables de classification croisée des quatre sous-strates O2, R2, O3 et R3 étaient: propriétaire ou locataire d'une maison, et 20 à 29 ans ou 30 à 44 ans. Les données des quatre sous-strates apparaissent au tableau 8, où 1 indique la présence et 0 l'absence dans l'échantillon. Nous avons réanalysés ces données aux fins de comparaison. Le tableau 9 et le tableau 10 donnent une estimation des deux modèles d'hétérogénéité. Comme on l'a indiqué plus haut, le taux d'appariement élevé débouche sur des estimations et un ajustement analogues pour les modèles qui incorporent l'erreur d'appariement. Le modèle à quasi-symétrie partielle donne un ajustement sensiblement plus précis que le modèle quasi-symétrique. Les meilleurs ajustements concernent les sous-strates R2 et R3. Si on additionne les estimations de N pour les quatre sous-strates, le total pour la version à quasi-symétrie partielle du modèle à erreur d'appariement donne $\hat{N} = 2980.8$, soit 16% de plus que l'estimation obtenue avec le modèle groupé du tableau 7. Bien sûr, l'erreur-type de l'estimation augmente du même ordre de grandeur.

Tableau 8

Données à trois échantillons pour quatre sous-strates de la strate 11
Source: tableau 2, Darroch et coll. (1993)

Échantillon			Sous-strate			
C	P	A	O2	R2	O3	R3
0	0	1	59	43	35	43
0	1	0	8	34	10	24
0	1	1	19	11	10	13
1	0	0	31	41	62	32
1	0	1	19	12	13	7
1	1	0	13	69	36	69
1	1	1	79	58	91	72

Tableau 9

Estimations pour une quasi-symétrie totale

Sous-strate	EP de Darroch et coll. (1993)		EP avec modèle à erreur d'appariement (2)	
	$\hat{N}$ (E.-T.)	Ajustement (d.l.)	$\hat{N}$ (E.-T.)	Ajustement (d.l.)
O2	780.83 (294.81)	11.70 (2)	777.98 (293.99)	11.69 (2)
R2	394.34 (56.45)	41.09 (2)	391.14 (55.29)	41.02 (2)
O3	765.45 (254.57)	25.99 (2)	759.97 (252.44)	25.98 (2)
R3	361.83 (47.33)	59.31 (2)	358.71 (46.20)	59.22 (2)

Tableau 10

Estimations pour une quasi-symétrie partielle

Sous-strate	EP de Darroch et coll. (1993)		EP du modèle à erreur d'appariement (2)	
	$\hat{N}$ (E.-T.)	Ajustement (d.l.)	$\hat{N}$ (E.-T.)	Ajustement (d.l.)
O2	605.66 (212.63)	7.51 (1)	601.44 (210.93)	7.52 (1)
R2	652.34 (205.12)	0.04 (1)	646.59 (202.58)	0.04 (1)
O3	1124.00 (473.26)	8.27 (1)	1126.90 (476.54)	8.22 (1)
R3	611.78 (200.82)	2.92 (1)	605.91 (198.26)	2.92 (1)

6. RÉCAPITULATION

Dans le présent article, nous proposons des modèles pour les erreurs d'appariement et des modèles pour l'estimation de la population totale et du sous-dénombrement du recensement dans un recensement à échantillon multiple. Nous avons illustré les méthodes proposées en procédant à une nouvelle analyse des données sur la couverture du recensement obtenues lors de la répétition générale de 1988 à St. Louis. Nous envisageons pour cela deux sources d'information, soit les données d'une étude sur l'erreur d'appariement (EEA) et celles d'un système triple où chaque sujet a fait l'objet d'une classification croisée selon sa présence ou son absence dans un des trois échantillons suivants: recensement, enquête postcensitaire (échantillon P) et liste administrative complémentaire. Nous avons intégré la formule type du modèle log-linéaire de Fienberg (1972) à notre méthode d'estimation pour tenir compte de la dépendance statistique et des erreurs d'appariement, et afin de disposer d'un test d'adéquation formelle pour les divers modèles. Notre méthode est valable pour n'importe quel modèle de type log-linéaire et nous avons montré comment les modèles d'hétérogénéité peuvent être incorporés à cette approche afin de tenir compte à la fois des erreurs d'appariement et d'un potentiel de saisie hétérogène.

Nos modèles de l'erreur d'appariement supposent des faux appariements négligeables. L'analyse de sensibilité effectuée par Ding (1990) montre que lorsque le taux de non-appariements erronés et le taux de faux appariements ont le même ordre de grandeur, le biais d'appariement résulte principalement du taux de faux non-appariements (lire aussi Fay, Passel, Robinson et Cowan 1988, p. 53). On le doit à une forte probabilité de saisie pour le recensement et l'enquête postcensitaire, si bien qu'une variation analogue des taux de non-appariements erronés et de faux appariements exerce sensiblement plus d'influence sur les non-appariements erronés que sur les faux appariements. Dans le cas des données censitaires de l'essai effectué en 1986 à Los Angeles, Ding et Fienberg (1994) estiment le taux de non-appariements erronés et le taux de faux appariements à environ 0.5% et 0.8%, respectivement. Compte tenu de ces résultats empiriques, il y a lieu de croire que

nos modèles pour les erreurs de non-appariements erronés donnent une approximation raisonnable de la réalité, du moins dans l'application censitaire dont il est question ici.

Nous avons analysé les données du système triple de St. Louis en estimant le taux d'appariement de l'EEA. Les taux d'appariement pourraient ne pas être homogènes avec différentes strates de la population. C'est pourquoi nous suggérons l'utilisation des données de l'EEA associées à la même strate d'échantillonnage. En §3, nous avons élaboré une formule pour un recensement à k échantillons et cette approche peut s'appliquer à un recensement à k échantillons où $k \geq 4$.

BIBLIOGRAPHIE

BISHOP, Y.M.M., FIENBERG, S.E., et HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: M.I.T. Press.

CHEN, T.T. (1979). Log-linear models for categorical data with misclassification and double sampling. *Journal of American Statistical Association*, 74, 481-488.

CORMACK, R.M. (1968). The statistics of capture-recapture methods. *Oceanography and Marine Biology, Annals Review*, 6, 455-506.

DARROCH, J.N. (1958). The multiple-recapture census, I: estimation of a closed population. *Biometrika*, 45, 343-359.

DARROCH, J.N., FIENBERG, S.E., GLONEK, G.F.V., et JUNKER, B.W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of American Statistical Association*, 88, 1137-1148.

DING, Y. (1990). Capture-recapture census with uncertain matching. Thèse de doctorat, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania.

DING, Y., et FIENBERG, S.E. (1994). Estimation de système dual du sous-dénombrement dans le recensement lorsqu'il y a erreur d'appariement. *Techniques d'enquête*, 20, 155-165.

FAY, R.E., PASSEL, J.S., ROBINSON, J.G., et COWAN, C.D. (1988). The coverage of population in the 1980 census. Bureau of the Census, U.S. Department of Commerce.

FIENBERG, S. E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, 59, 591-603.

HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête post-censitaire. *Techniques d'enquête*, 14, 105-124.

MULRY, M.H., DAJANI, A., et BIEMER, P. (1989). The Matching Error Study for the 1988 Dress Rehearsal. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 704-709.

RAO, C.R. (1957). Maximum likelihood estimation for the multinomial distribution. *Sankhyā*, 18, 139-148.

SANATHANAN, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43, 142-152.

SEBER, G.A.F. (1982). *The Estimation of Animal Abundance and Related Parameters.* New York: MacMillan.

ZASLAVSKY, A.M., et WOLFGANG, G.S. (1993). Triple System Modeling of Census, Post-Enumeration Survey and Administrative List Data. *Journal of Business and Economic Statistics*, 11, 279-288.

Techniques d'enquête, juin 1996
Vol. 22, n° 1, pp. 65–75
Statistique Canada

Utilisation de la méthode de Lavallée et Hidiroglou pour le calcul des limites de stratification aux fins de l'enquête annuelle sur les dépenses en capital du Bureau of the Census

JOHN G. SLANTA et THOMAS R. KRENZKE[1]

RÉSUMÉ

On a utilisé la méthode de Lavallée-Hidiroglou (L-H) pour le calcul des bornes de stratification dans le cadre de l'enquête annuelle du Bureau of the Census sur les dépenses en capital (Annual Capital Expenditures Survey, ou ACES), afin de stratifier une portion de l'univers de cette enquête aux fins de l'étude pilote et des enquêtes préliminaires subséquentes. Cette méthode itérative minimise la taille de l'échantillon tout en fixant le niveau souhaité de fiabilité en établissant des bornes appropriées. Toutefois, deux problèmes se sont posés lors de son application. D'abord, des bornes de départ différentes ont engendré des bornes finales différentes. Deuxièmement, la convergence vers les bornes localement optimales a été lente; autrement dit, le nombre d'itérations était élevé et la convergence n'était pas garantie. Le présent article décrit les difficultés rencontrées avec la méthode L-H et montre comment elles ont été résolues de manière à ce que cette méthode fonctionne correctement pour l'ACES. Nous décrivons notamment comment les tracés de contours ont été établis et utilisés afin d'illustrer l'importance des problèmes relevés lors de l'utilisation de la méthode L-H. Nous décrivons en outre les changements apportés à la méthode L-H afin d'en faire un moyen utile de détermination des bornes de stratification pour l'ACES.

MOTS CLÉS: Convergence; tracés de contours; enquêtes économiques.

1. INTRODUCTION

Le plan de sondage de l'enquête annuelle du Bureau of the Census sur les dépenses en capital (Annual Capital Expenditures Survey ou ACES) a pour objectifs principaux de procurer les niveaux de fiabilité souhaités à l'aide de méthodes faisables au plan opérationnel tout en respectant les contraintes budgétaires imposées. Pour répondre à ces objectifs, nous avons mis au point un plan d'échantillonnage aléatoire simple stratifié fondé sur une version modifiée de la méthode de détermination des bornes de stratification de Lavallée et Hidiroglou (L-H) (1988). Cette méthode de stratification pour les populations asymétriques donne des bornes optimales en minimisant la taille de l'échantillon total, compte tenu d'un coefficient de variation (c.v.) souhaité. Les personnes qui sont chargées d'une enquête spécialisée et qui ont accès à une variable de stratification unique peuvent tirer parti de la facilité d'exécution de cette méthode et des économies qu'elle permet de réaliser.

Nous avons examiné plusieurs articles qui portaient sur d'autres méthodes de détermination des bornes des strates fondées sur la taille. Hess, Sethi et Balakrishnan (1966) ont comparé plusieurs méthodes de stratification. La populaire méthode de Dalenius et Hodges (Cochran 1977, p. 129) a été jugée facile à mettre en oeuvre dans notre cas, mais nous l'avons rejetée au départ puisqu'elle ne tient pas compte des strates à tirage complet au moment de sa mise au point. La méthode de Sethi (1963) utilisant les distributions normales n'a pas été retenue parce que nous avons jugé qu'il serait encombrant d'identifier la distribution et que l'utilisation des distributions normales pour chacune des 80 types d'industrie examinés dans le cadre de l'enquête ne constituerait pas une solution optimale. La règle de l'égalité du produit des poids et de la portée des strates proposée par Eckman (1959) nous a semblé exiger des calculs plutôt rébarbatifs.

La méthode L-H est donc apparue la plus propice à nos besoins. Conçue précisément pour les populations asymétriques, une situation qu'on observe souvent avec des enquêtes économiques, cette méthode calcule une borne qui définit la strate à tirage complet, ainsi que les bornes optimales pour les strates à tirage partiel. En outre, elle crée parfois une strate à tirage complet supplémentaire si, sous l'effet de la répartition de Neyman, la taille de l'échantillon de la strate est plus grande ou égale à la taille de la strate.

La méthode L-H utilise un algorithme itératif qui débute avec le calcul ou la détermination arbitraire des bornes initiales des strates. Ensuite, on procède au calcul des paramètres tels que la taille, la moyenne et la variance des strates. Ces paramètres sont incorporés dans les formules de calcul des bornes, elles-mêmes dérivées d'une minimisation de la taille de l'échantillon en fonction d'une valeur souhaitée du c.v. Si les nouvelles bornes ne convergent pas, les paramètres des strates sont calculés à nouveau pour la strate nouvellement définie. Le cycle se poursuit jusqu'à ce que les bornes convergent.

Schneeberger (1979) a abordé le problème de la détermination des bornes de stratification optimales. Il a montré dans son article que lorsqu'on aborde ce problème sous l'angle d'un programme non linéaire, résolu par une méthode

[1] John G. Slanta, Manufacturing and Construction Division; Thomas R. Krenzke, Decennial Statistical Studies Division, U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A.

du gradient, la solution obtenue peut être une valeur minimale, maximale, tant locale que globale, ou encore un point de selle de la variance de la moyenne de l'échantillon. Detlefsen et Veum (1991) ont jugé que cette caractéristique de la méthode L-H constituait un inconvénient lorsqu'ils en ont étudié l'application pour l'enquête mensuelle sur le commerce au détail du Bureau of the Census. Ils ont observé que les bornes obtenues avec la méthode L-H étaient souvent sensiblement différentes des bornes fixées au départ, de sorte que la taille minimale de l'échantillon obtenu correspondait à un minimum local. Représentée géométriquement, la taille de l'échantillon en fonction de deux bornes de stratification prend l'aspect d'un paysage marqué d'une ou plusieurs vallées en forme de cuvettes. La méthode L-H commence dans une région et descend jusqu'au point le plus bas. S'il existe plus d'un minimum, elle ne poursuit pas sa recherche du minimum global. En conséquence, un des objectifs visés consiste à avoir des bornes initiales qui se situent au voisinage du minimum global. Le recours à des bornes de départ découlant d'une méthode comme celle de Dalenius et Hodges pourrait permettre de répondre à cette exigence.

Detlefsen et Veum (1991) ont également relevé des cas de convergence lente ou de non-convergence. Toutefois, ils ont également noté que la convergence est plus rapide lorsque le nombre de strates est réduit et lorsque les bornes de départ sont les mêmes que les bornes de sélection d'échantillon de l'enquête précédente. Afin de nous prémunir contre un nombre de boucles infinies dans le programme informatique ou contre un nombre très élevé d'itérations, nous avons décidé de faire deux choses: premièrement, nous avons mis en oeuvre un plan d'échantillonnage dans lequel la méthode L-H créerait des ensembles de seulement trois strates fondées sur la taille. Deuxièmement, nous avons décidé de mettre en oeuvre des règles d'interruption qui feraient en sorte d'interrompre le programme lorsque le taux de convergence semblait diminuer.

Dans le présent article, nous fournissons des informations de base sur l'ACES et nous décrivons brièvement notre utilisation de la méthode L-H. Nous montrons comment les tracés de contours et les graphiques tridimensionnels justifient l'utilisation de la méthode L-H pour obtenir les bornes finales. Nous expliquons comment les tracés de contours permettent de résoudre le problème de la convergence en montrant comment les contraintes peuvent être réglées de manière à être prises en compte après chaque itération. Ce procédé nous protège contre la convergence trop lente ou la non-convergence, lorsqu'on présume que les avantages attendus ne valent pas les efforts supplémentaires requis.

2. CONTEXTE DE L'ACES

L'ACES de 1992 a été conçue par le Bureau of the Census pour devenir un test opérationnel des méthodes d'échantillonnage, de traitement, de programmation, de saisie des données, de révision et d'estimation qui se prolongerait au-delà d'une étude pilote conduite en 1991, en prévision de la tenue de l'enquête définitive, en 1993. Les valeurs estimées des dépenses en capital correspondant aux activités intérieures ont été publiées par groupes de branches à partir de l'enquête de 1992. En outre, les enquêtes préliminaires de 1991 et de 1992 ont fourni de précieuses données sur les dépenses en capital qui serviront aux améliorations futures des plans d'échantillonnage.

L'unité de sondage de l'ACES était l'entreprise, qui pouvait comprendre plusieurs établissements. La population échantillonnée comprenait toutes les entreprises actives comptant cinq employés ou plus et oeuvrant dans tous les principaux secteurs industriels, exception faite du gouvernement. Ces secteurs comprenaient les mines, la construction, le secteur manufacturier, le transport, le commerce en gros et au détail, les finances, les services ainsi qu'une portion du secteur agricole comprenant les services agricoles, la foresterie, la pêche, la chasse et le trappage. Seules les entreprises nationales ont été incluses dans la base de sondage. Le personnel du service de recherche et de méthodologie de la Division de l'industrie du Bureau of the Census s'est chargé d'établir la base de sondage, de choisir l'échantillon et de produire les estimations.

La base d'échantillonnage de l'ACES a été construite à partir de la Standard Statistical Establishment List (SSEL) du Bureau of the Census en novembre 1992, en utilisant les données finales de 1991 pour les établissements à une seule unité (Single Unit ou SU) et les données de 1990 pour les établissements associés à des entreprises à unités multiples (multiunit ou MU). Ont été exclus de la base de sondage les services gouvernementaux, le service postal, les établissements internationaux, les établissements de Porto Rico, de Guam, des îles Vierges et des îles Mariannes. Les données des registres secondaires d'identification des employeurs, qui correspondent à des établissements SU figurant sur la liste SSEL, mais associés à des établissements MU, les établissements associés à la production agricole et les ménages privés ont également été exclus de la base de sondage.

Le fichier fondé sur les établissements a été fondu en un fichier fondé sur les entreprises. En outre, les codes à quatre chiffres de la classification type des industries (Standard Industrial Classification ou SIC) correspondant à chaque entreprise ont été remplacés par des codes de catégorie ACES. Les 80 catégories ACES étaient constituées soit de codes SIC à trois chiffres ou de combinaison de ces codes. La base de sondage de l'ACES comprenait environ deux millions d'entreprises.

3. APPLICATION DE LA MÉTHODE L-H À L'ACES

L'univers de l'enquête a été divisé en deux strates principales. La strate I a été définie arbitrairement comme une strate à tirage complet constituée de grandes entreprises comptant plus de 500 employés et disposant d'un actif de plus de 100 millions de dollars. Les entreprises de la strate I n'ont pas été classées dans une industrie ACES. Pour les totaux estimés de la paye utilisés dans le calcul de la taille des échantillons du niveau industriel, les entreprises de la strate I pouvaient contribuer à plus d'une industrie ACES, selon le nombre des branches ACES différentes pour lesquelles elles possédaient des feuilles de paye, tel qu'indiqué dans la liste SSEL.

La strate II contenait les entreprises comptant entre cinq et 500 employés. Les entreprises de la strate II ont été classées en une seule industrie, même si elles s'adonnaient à plus d'une activité. Chaque entreprise possédait des informations de base pour chacune des branches ACES dans lesquelles elle était active. Toutefois, la feuille de paye d'une entreprise donnée contribuait uniquement à la paye totale estimée versée dans la branche dans laquelle elle était classée. Par la suite, à l'intérieur de la strate II, pour chaque catégorie d'industrie ACES, trois strates fondées sur la taille ont été créées en fonction de la paye totale annuelle de l'entreprise à l'aide de la méthode L-H.

Ce plan de sondage présente un risque de mauvaise classification des entreprises par suite de la mesure de la taille choisie. Nous avons classé chaque entreprise de la strate II dans la branche où elle versait la paye la plus élevée. Toutefois, les entreprises indiquent elles-mêmes leurs dépenses en capital consacrées aux branches ACES dans le questionnaire ACES. Elles peuvent indiquer des dépenses dans de multiples branches. Si trop d'entreprises signalent des dépenses dans des branches autres que celle dans laquelle elles ont été classées, on perd alors toute emprise sur la fiabilité des estimations.

On s'inquiète par ailleurs d'un écart possible entre la variation de la paye et la variation des dépenses. Comme la taille de l'échantillon est directement liée à la variance, il se peut que les tailles d'échantillons soient différentes de ce qui est réellement nécessaire. En conséquence, comme la corrélation entre la paye et les dépenses n'est pas élevée, les possibilités de répondre aux exigences de fiabilité s'amenuiseront.

L'application de la méthode L-H au plan de sondage préliminaire de l'ACES de 1992 exigeait la séparation de la strate II en une strate à tirage complet et en deux strates à tirage partiel pour chaque industrie de l'ACES. Les bornes ont été dérivées pour chaque industrie en prenant la dérivée partielle de la taille de l'échantillon pour une des bornes, tout en fixant l'autre. Toutefois en pratique, nous avons permis aux deux bornes de se déplacer simultanément. Nous avons ainsi obtenu un processus itératif de minimisation de la taille de l'échantillon pour chaque industrie, compte tenu des contraintes imposées par le c.v. À l'intérieur de la strate II, pour chaque industrie de l'ACES, et en tenant pour acquise la répartition de Neyman (Detlefsen et Veum 1991), l'équation de la taille de l'échantillon qui est minimisée se présente comme suit:

$$n = n_{TA} + \frac{N\left(\sum_{j=1}^{2} W_j S_j\right)^2}{\frac{cv^2 Y^2}{N} + \sum_{j=1}^{2} W_j S_j^2}, \qquad (1)$$

où n_{TA} désigne le nombre d'entreprises dans la strate à tirage complet à l'intérieur de la strate II définie par la méthode L-H, N désigne le nombre d'entreprises de la strate II dans la branche ACES qui nous intéresse, $W_j = N_j/N$ désigne la proportion de la strate, N_j désigne le nombre d'entreprises de la strate II pour la strate j, c.v. désigne le coefficient de variation désiré pour la branche ACES qui nous intéresse, Y désigne la paye totale pour les strates I et II de la branche ACES qui nous intéresse, définie par

$$Y = \sum_{k=1}^{N_I} y_k + \sum_{j=1}^{3} \sum_{i=1}^{N_j} y_{ji},$$

N_I désigne le nombre d'entreprises de la strate I et S_j désigne l'écart-type de la paye tiré de la liste SSEL pour la strate j dans la strate II définie par

$$S_j = \sqrt{\frac{\sum_{i=1}^{N_j} (y_{ji} - \bar{Y}_j)^2}{N_j - 1}},$$

où y_{ji} désigne la valeur de la paye de l'entreprise i appartenant à la strate j pour la branche ACES qui nous intéresse et $\bar{Y}_j$ désigne la moyenne de la paye pour la strate j.

Le taux de fiabilité pour chaque industrie correspondait à une valeur attendue du c.v. de 5% sur la paye. Nous ignorons toutefois quelles seraient les valeurs de l'écart-type découlant des dépenses en capital, puisqu'il n'existait pas de données sur les dépenses en capital pour la base de sondage. Les entreprises répondant dans des branches ACES différentes de celles pour lesquelles elles avaient contribué dans le plan de sondage ont également causé des fluctuations des valeurs du c.v. Le nombre total d'entreprises sélectionnées pour l'enquête préliminaire de l'ACES de 1992 était de 11,194, soit 1,500 entreprises de la strate I et 9,694 entreprises de la strate II.

4. CONVERGENCE EN VOISINAGES

Il peut parfois arriver que la méthode L-H exige un grand nombre d'itérations avant que les bornes ne convergent; il arrive même que ces bornes ne convergent jamais. En règle générale, après quelques itérations seulement, une vaste proportion des améliorations possibles de la taille de l'échantillon ont déjà été obtenues. Notre but était de mettre en pratique des règles d'interruption qui permettraient d'arrêter le processus dès qu'on atteindrait une certaine superficie autour d'un minimum local. C'est ce qui nous a poussés à utiliser des tracés de contours dans notre analyse de l'effet des bornes sur la taille de l'échantillon obtenue. Cela nous a également permis d'obtenir un aperçu graphique des voisinages des minimum locaux. Nous utiliserons deux distributions pour illustrer les avantages d'une révision des tracés de contours.

4.1 Distribution symétrique

Notre premier exemple est une distribution symétrique tirée de l'article de Schneeberger. Cette distribution est symétrique au point $x = 1$, comme l'illustre la figure 1.

$$f(x) = \begin{cases} 0 & x \le 0 \\ 2x & 0 < x \le 0.5 \\ 2(1-x) & 0.5 < x \le 1 \\ 2(x-1) & 1 < x \le 1.5 \\ 2(2-x) & 1.5 < x \le 2 \\ 0 & 2 < x \end{cases}$$

Schneeberger cherchait à définir les bornes de trois strates à tirage partiel en utilisant une méthode du gradient. Les résultats obtenus à l'aide de la fonction objective de $z = (\sum W_h \sigma_h)^2$ sont présentés au tableau 1.

Tableau 1

Bornes optimales pour une distribution non asymétrique

	b_1	b_2	Point optimal
(2a)	.50241	1.03985	Point minimal
(2b)	.70910	1.29090	Point de selle
(2c)	.96015	1.49759	Point minimal

Source: Schneeberger (1979).

Nous avons généré cinq ensembles de données de tailles différentes (N = 50, 100, 200, 1,000 et 5,000) en utilisant la formule $F(x) = (j - 1/2)/N$. Pour cet exemple, nous avons adapté la méthode L-H afin d'établir trois strates à tirage partiel et aucune à tirage complet et de pouvoir ainsi comparer nos résultats avec ceux de l'article de Schneeberger. Dans notre application de l'estimation des totaux, avec minimisation de la taille de l'échantillon sous réserve d'une valeur du c.v. de 5%, nous utilisions la méthode L-H pour chacune des cinq tailles de population en utilisant trois méthodes différentes de démarrage. Les résultats de cette opération sont présentés au tableau 2.

Les informations du tableau 2 appellent trois commentaires principaux. Premièrement, la convergence des algorithmes dépend de la taille de la population. La théorie sous-jacente de la méthode L-H est fondée sur des distributions continues. Nos exemples, à l'image de toute application de sondage, sont faits de données discrètes tirées de populations finies. On observe également qu'à mesure que la valeur de N augmente, les bornes obtenues se rapprochent de la valeur minimale correspondant à une population de taille infinie. La figure 2 montre la rugosité de la surface de la taille de l'échantillon lorsque la valeur de N est petite (N = 50). La représentation tridimensionnelle de la figure 2 montre bien la forme en selle de la surface obtenue. Dans ce graphique, les axes représentent les bornes inférieure et supérieure et la surface représente les tailles d'échantillons obtenues. Ce graphique montre le point de selle, les deux minimum locaux, et donne également un aperçu de l'importance des réductions de taille de l'échantillon découlant d'un déplacement des bornes. Par comparaison, la figure 3 montre que la surface obtenue lorsque la valeur de N est élevée (N = 5000) est lisse. Ces résultats portent à conclure que la rugosité de la surface de la taille de l'échantillon et, en conséquence, la taille de la population auront un effet sur le lieu de convergence des bornes.

Cet exemple nous permet en outre de rappeler encore une fois que les bornes finales dépendent des bornes initiales. Pour cet exemple, Schneeberger indique qu'avec un point de départ symétrique à $x = 1$, où $b_1 = 1 - \lambda$ et $b_2 = 1 + \lambda (0 < \lambda < 1)$ et qui définit la ligne $b_2 = 2 - b_1$, la méthode du gradient déplace le gradient sur la ligne $b_2 = 2 - b_1$ jusqu'au point d'équilibre. Lorsque nous posons les bornes de départ sur cette ligne, comme cela s'est produit lorsque nous avons commencé avec $N_1 = N_2 = N_3$, la méthode L-H converge également au point d'équilibre (voir tableau 1). Avec les bornes de départ de

Tableau 2

Bornes L-H de trois strates à tirage partiel pour une distribution symétrique

N	Méthode de démarrage	1ère itération			Itération à moins de 5% de la taille de l'échantillon				Itération finale			
		b_1	b_2	n	b_1	b_2	n	itér. n°	b_1	b_2	n	itér. n°
50	$N_1 = N_2 = N_3$	.59	1.41	10.89	.66	1.34	9.98	2	.70	1.31	9.77	4
100	$N_1 = N_2 = N_3$	.59	1.41	12.60	.66	1.34	10.91	2	.70	1.30	10.55	5
200	$N_1 = N_2 = N_3$	.59	1.41	13.42	.66	1.34	11.43	2	.71	1.29	10.99	6
1000	$N_1 = N_2 = N_3$	.59	1.41	13.85	.66	1.34	11.75	2	.71	1.29	11.37	7
5000	$N_1 = N_2 = N_3$	.59	1.41	14.12	.66	1.34	11.84	2	.71	1.29	11.45	9
50	Dalenius-Hodges	.70	1.40	10.09	.70	1.40	10.09	1	.77	1.37	9.63	4
100	Dalenius-Hodges	.70	1.40	10.90	.84	1.40	10.14	7	.93	1.47	9.65	13
200	Dalenius-Hodges	.70	1.40	11.42	.83	1.40	10.44	7	.95	1.49	9.96	17
1000	Dalenius-Hodges	.70	1.40	11.86	.86	1.42	10.67	8	.96	1.50	10.27	23
5000	Dalenius-Hodges	.70	1.40	11.95	.86	1.42	10.74	8	.96	1.50	10.34	28
50	Hors ligne	.50	1.30	10.87	.57	1.20	9.43	3	.55	1.11	9.11	6
100	Hors ligne	.50	1.30	11.95	.57	1.18	10.04	3	.53	1.07	9.65	8
200	Hors ligne	.50	1.30	12.64	.56	1.14	10.28	4	.51	1.05	9.96	12
1000	Hors ligne	.50	1.30	13.24	.56	1.14	10.59	4	.50	1.04	10.27	18
5000	Hors ligne	.50	1.30	13.37	.56	1.14	10.67	4	.50	1.04	10.34	24

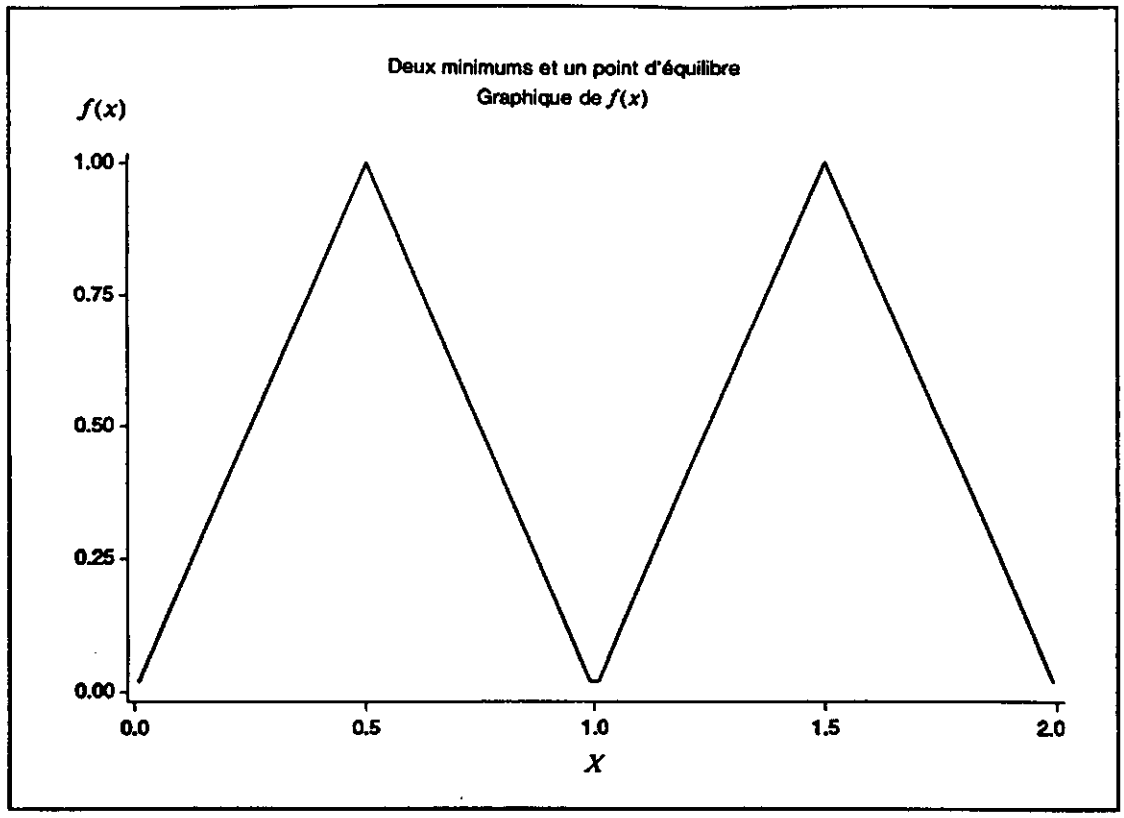

Figure 1. Graphique d'une distribution symétrique.

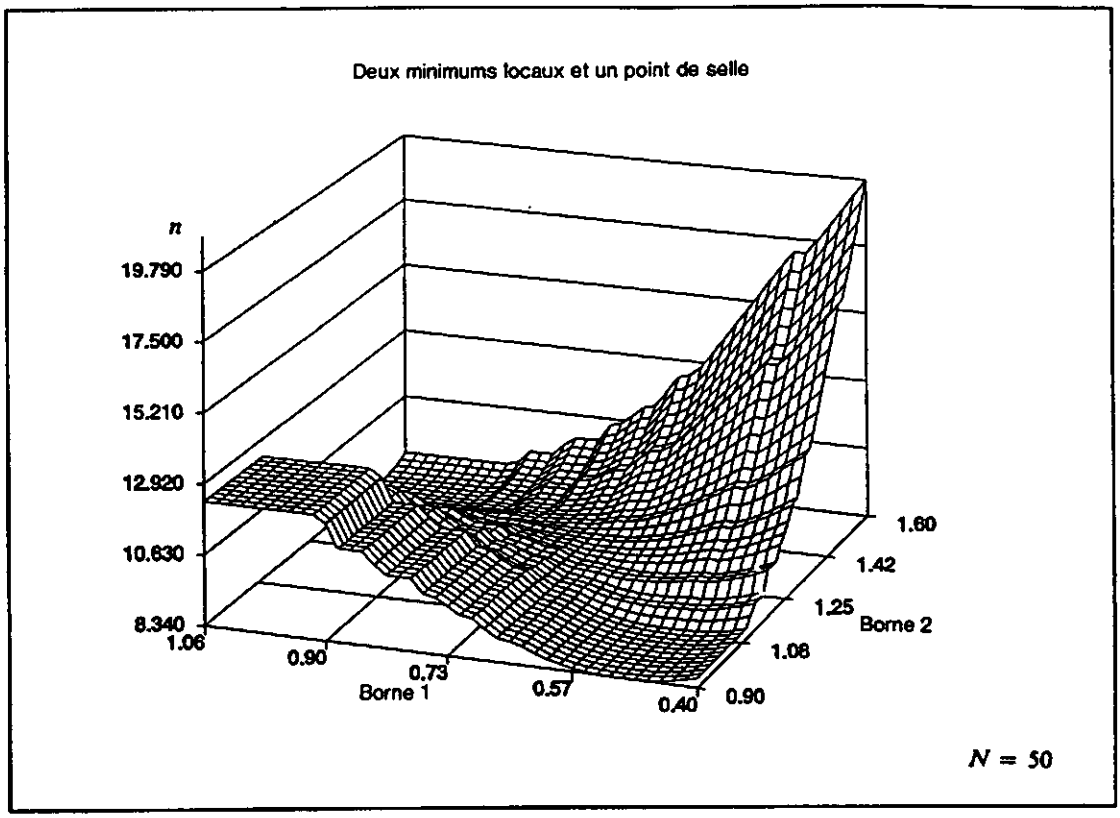

Figure 2. Surface de la taille de l'échantillon pour une distribution symétrique ($N = 50$).

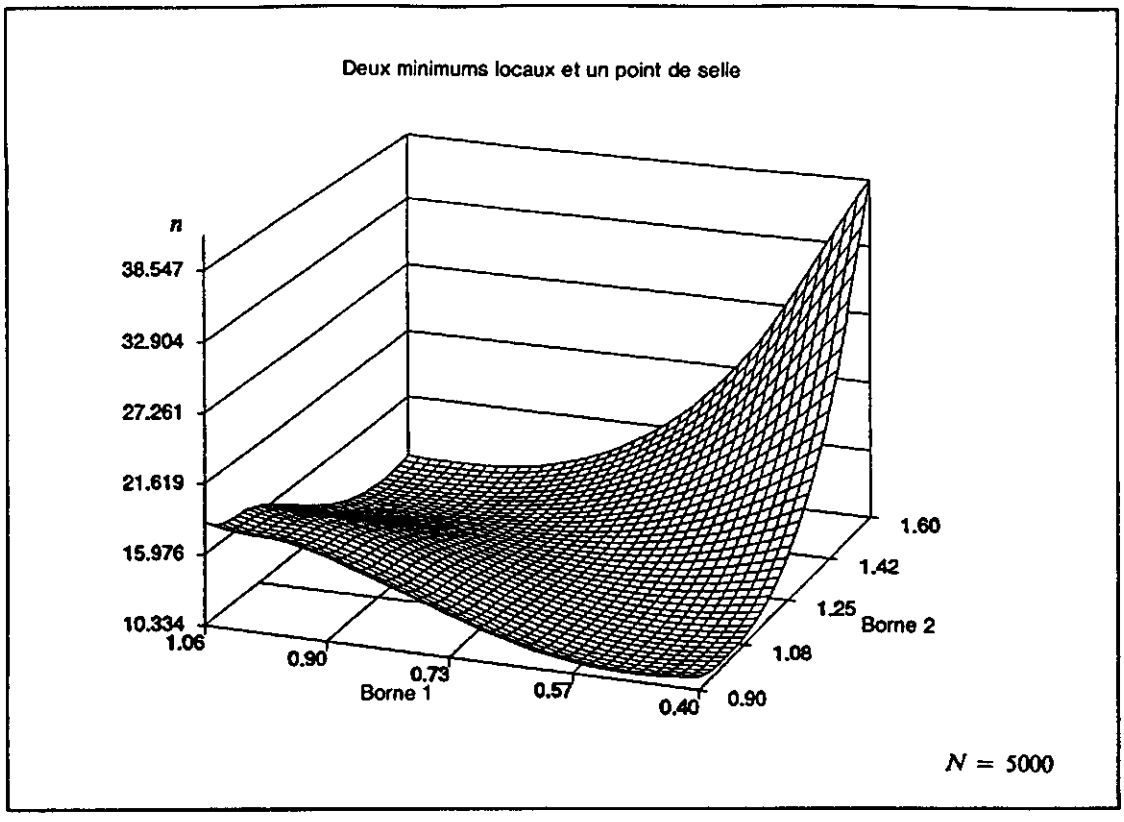

Figure 3. Surface de la taille de l'échantillon pour une distribution symétrique ($N = 5000$).

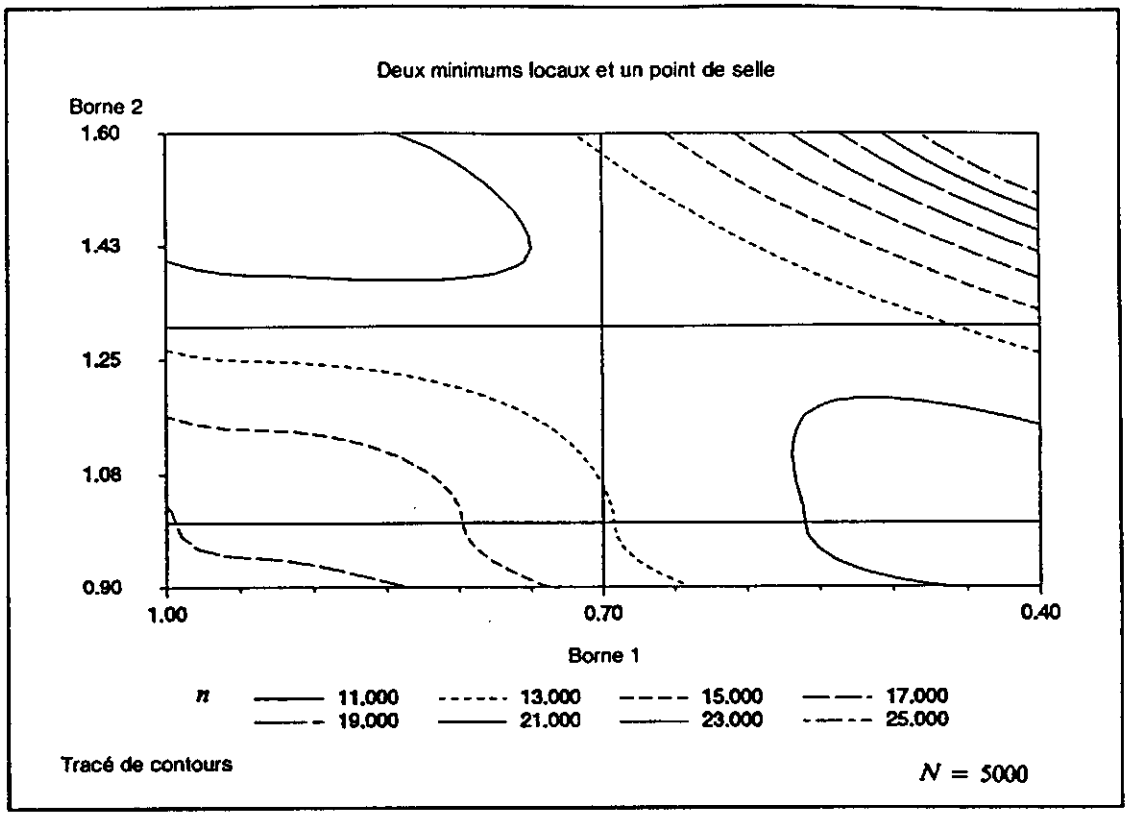

Figure 4. Tracé de contours pour une distribution symétrique ($N = 5000$).

la méthode de Dalenius-Hodges, lesquelles ne se situent pas sur la ligne lorsque $b_2 > 2 - b_1$, la méthode L-H converge vers un minimum (2c). La méthode de Dalenius-Hodges fonctionne bien dans cet exemple à cause des trois strates à tirage partiel. Avec des bornes de départ qui ne se situent pas sur la ligne dans le cas où $b_2 < 2 - b_1$ (dans ce cas-ci, $b_1 = .5$ et $b_2 = 1.3$), la méthode L-H converge vers un minimum différent (2a). Ce problème n'est pas unique à la méthode L-H puisque Schneeberger indique que les bornes obtenues par la méthode du gradient sont également dépendantes des bornes de départ.

On observe finalement dans cet exemple que les quelques premières itérations donnent des réductions relativement grandes de la taille de l'échantillon et que les itérations qui suivent donnent plusieurs réductions plus petites de cette taille. Nous présentons au tableau 2 les résultats de l'itération dans laquelle l'algorithme produit une taille de l'échantillon qui s'écarte de moins de 5% de la taille finale. Ceci signifie que l'algorithme L-H tend rapidement vers un voisinage autour de la borne optimale. Lorsque nous sommes près d'une taille optimale de l'échantillon, il semble exister une vaste gamme de bornes produisant un nombre limité de tailles d'échantillon. L'application de règles d'interruption peut effectivement réduire le temps de calcul sans nous conduire à renoncer à une véritable réduction de la taille de l'échantillon puisque les valeurs de ce paramètre sont données en nombres entiers.

Nous présentons à la figure 4 un tracé de contours de la surface illustrée à la figure 3. Ici encore, les axes représentent les bornes inférieures et supérieures des strates et la surface est définie par la taille d'échantillon qui en découle. Les lignes du tracé représentent une valeur de taille de l'échantillon. L'espace délimité correspond à une gamme de valeurs de la taille de l'échantillon. Par exemple, la ligne continue représente une taille d'échantillon de 11 et la ligne en tirets courts représente une taille d'échantillon de 13. La superficie délimitée par la ligne continue et la ligne en tirets courts contient les tailles d'échantillon comprises entre 11 et 13. Ce tracé de contours illustre une amélioration marginale de la taille de l'échantillon et montre que lorsqu'on atteint la portion inférieure de la surface, il n'est plus nécessaire de poursuivre l'opération. À partir de ce point, la plupart des améliorations de la taille de l'échantillon obtenue d'une itération à l'autre sont inférieures à l'unité. Il devient évident qu'après les quelques premières itérations, l'amélioration obtenue dans la taille de l'échantillon d'une itération à l'autre diminue rapidement. Par exemple, dans le tableau 2, lorsque $N = 5000$ et lorsqu'on a utilisé la méthode de Dalenius-Hodges pour déterminer les bornes de départ, les huit premières itérations ont donné 74% de la réduction totale de la taille de l'échantillon obtenue entre la première et la 28e et dernière itération.

4.2 Distribution asymétrique

Les données économiques sont habituellement hautement asymétriques, ce qui nous porte à accorder notre préférence à une strate à tirage complet. Le prochain exemple est tiré de la distribution de Pareto, une distribution extrêmement typique des univers économiques où il existe un grand nombre de petites entreprises et un petit nombre de grandes entreprises.

La fonction de la distribution de Pareto est définie par $F(x) = 1 - 1/(1 + x)^b$, $0 \leq x < \infty$. De cette fonction, nous avons encore une fois tiré cinq ensembles de données de tailles différentes en utilisant la formule $F(x) = (j - 1/2)/N$. Nous avons laissé les valeurs de b changer à mesure que la taille de la population évoluait.

Tableau 3

Bornes L-H d'une distribution asymétrique (une strate à tirage complet, deux strates à tirage partiel)

N	Méthode de démarrage	1ère itération					Itération à moins de 5% de la taille de l'échantillon					Itération finale					
		b	b_1	b_2	n_{TA}	n	b_1	b_2	n_{TA}	n	itér. n°	b	b_1	b_2	n_{TA}	n	itér. n°
50	$N_1 = N_2 = N_3$	.80	.63	2.81	17	17.2	1.66	10.20	7	9.6	5	.80	2.44	11.81	7	9.4	9
100	$N_1 = N_2 = N_3$	.90	.56	2.33	34	34.3	1.61	10.29	11	15.8	5	.90	2.58	12.44	10	15.1	12
200	$N_1 = N_2 = N_3$	.90	.56	2.36	67	67.2	2.35	17.04	15	21.8	6	.90	3.61	20.46	13	20.9	13
1000	$N_1 = N_2 = N_3$	1.00	.50	2.00	333	334.2	3.35	30.58	32	53.0	7	1.00	4.93	36.32	27	51.3	18
5000	$N_1 = N_2 = N_3$	1.05	.47	1.85	1665	1667.2	4.67	64.33	62	113.5	7	1.05	7.39	79.38	50	108.8	22
50	Dalenius-Hodges	.80	1.25	8.04	9	10.5	1.76	10.37	7	9.5	3	.80	2.44	11.81	7	9.4	6
100	Dalenius-Hodges	.90	1.39	8.98	13	16.6	1.62	10.16	11	15.8	2	.90	2.58	12.44	10	15.1	9
200	Dalenius-Hodges	.90	1.82	11.66	20	24.3	2.45	17.29	15	21.7	3	.90	3.61	20.46	13	20.9	10
1000	Dalenius-Hodges	1.00	2.37	17.28	55	65.6	3.15	29.70	33	53.5	3	1.00	4.93	36.32	27	51.3	15
5000	Dalenius-Hodges	1.05	3.09	26.27	155	175.0	4.98	66.28	60	112.3	4	1.05	7.39	79.38	50	108.8	19
50	Hidiroglou 1986	.80	.94	6.50	10	11.3	1.58	10.02	7	9.6	3	.80	2.44	11.81	7	9.4	7
100	Hidiroglou 1986	.90	.74	6.17	17	19.6	1.66	10.38	11	15.8	4	.90	2.58	12.44	10	15.1	11
200	Hidiroglou 1986	.90	1.39	9.55	24	27.2	2.50	17.58	14	21.5	4	.90	3.61	20.46	13	20.9	10
1000	Hidiroglou 1986	1.00	2.02	15.13	62	71.3	3.34	30.54	32	53.0	4	1.00	4.93	36.32	27	51.3	15
5000	Hidiroglou 1986	1.05	3.24	28.72	142	164.1	5.11	67.05	59	112.0	4	1.05	7.39	79.38	50	108.8	19

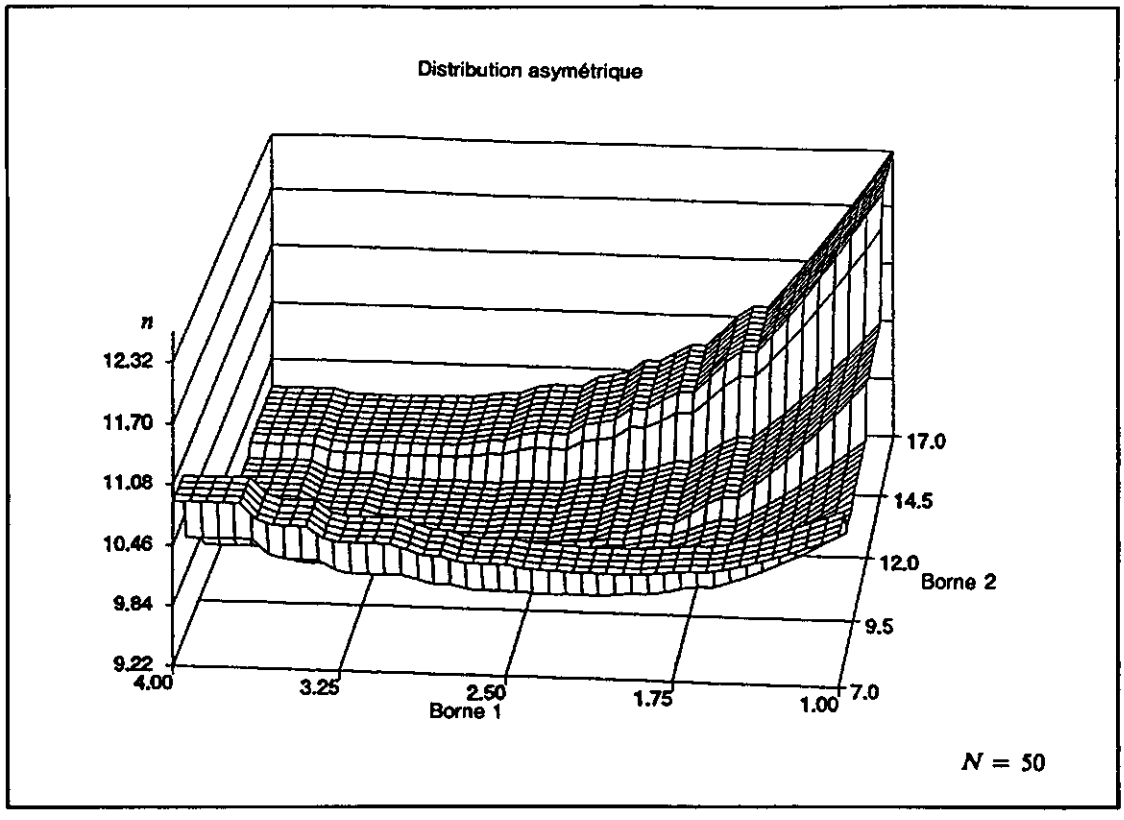

Figure 5. Surface de la taille de l'échantillon pour une distribution asymétrique ($N = 50$).

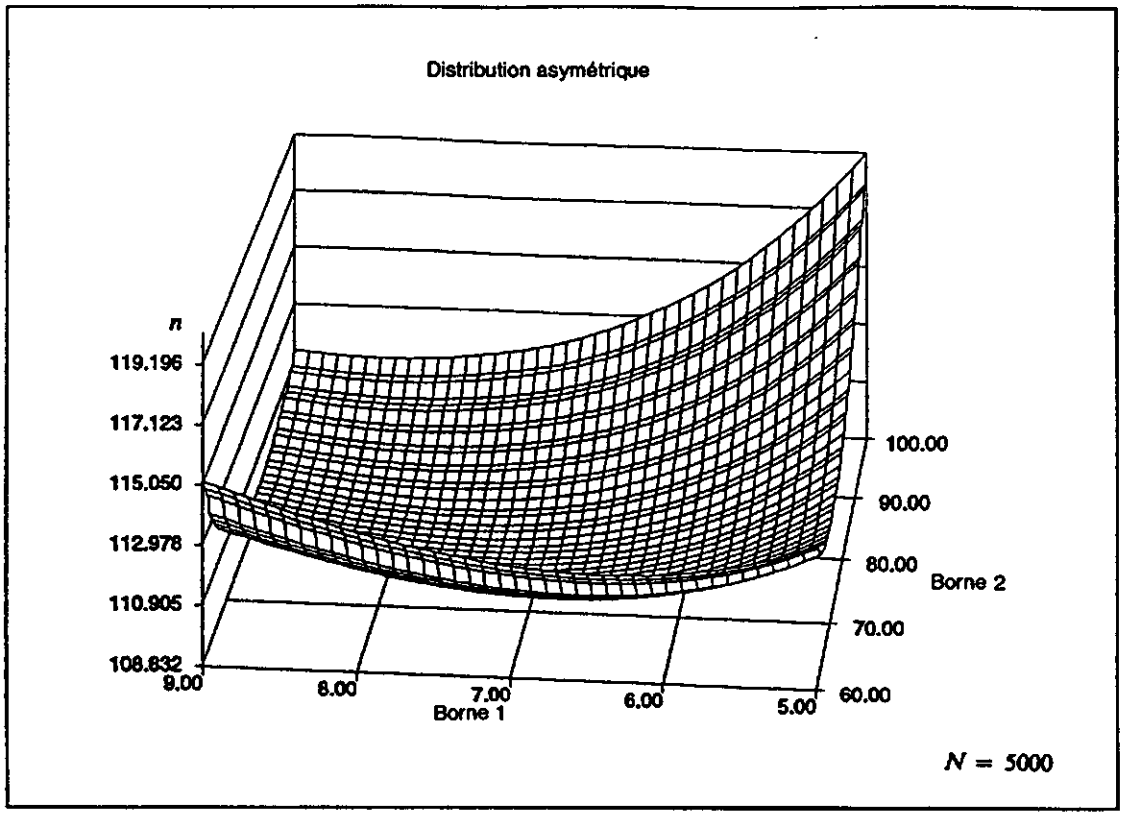

Figure 6. Surface de la taille de l'échantillon pour une distribution asymétrique ($N = 5000$).

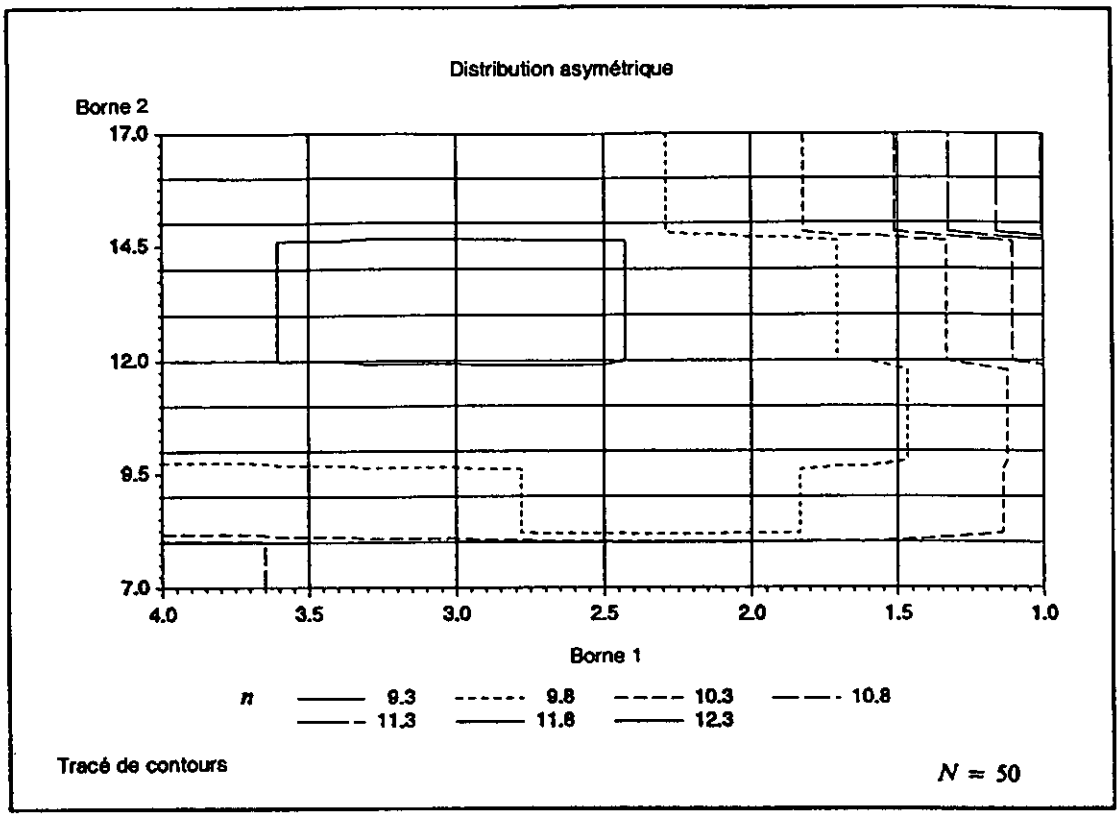

Figure 7. Tracé de contours d'une distribution asymétrique ($N = 50$).

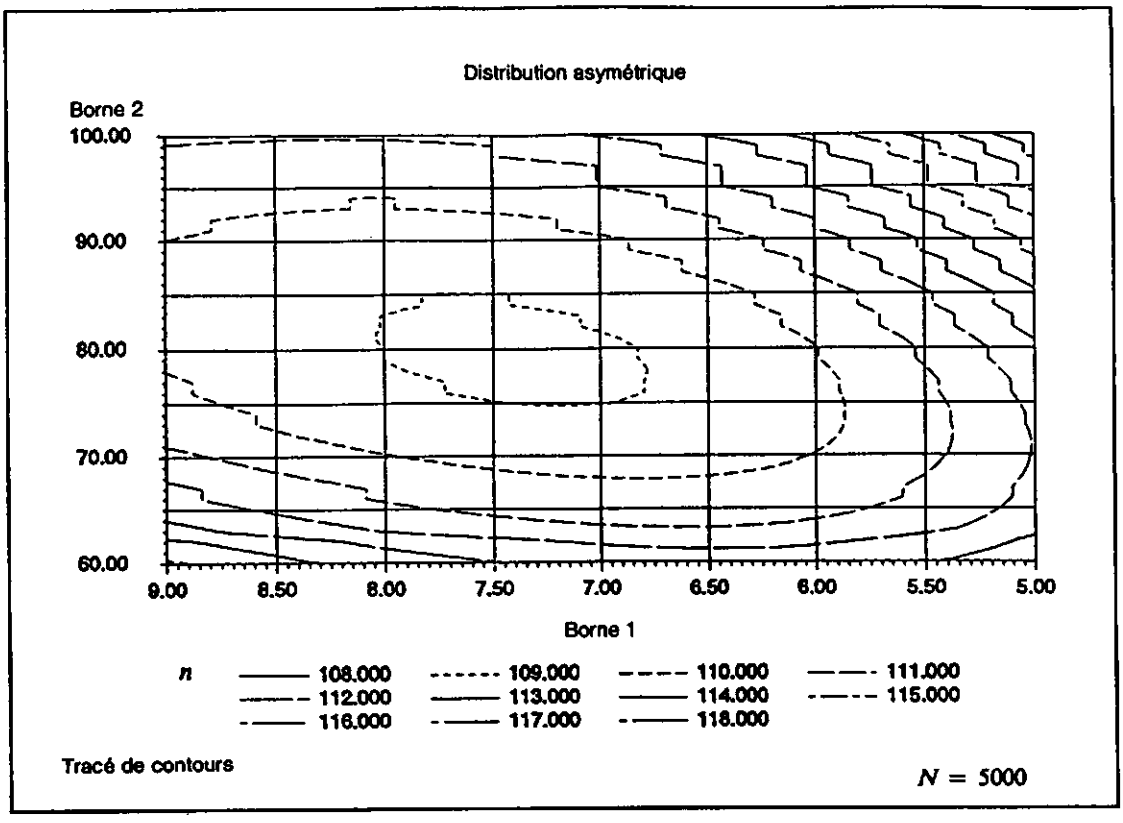

Figure 8. Tracé de contours d'une distribution asymétrique ($N = 5000$).

Nous avons procédé ainsi pour maintenir l'extrémité supérieure de la distribution discrète finie proportionnellement comparable, en gros, à la population entière pour chaque taille de population. À cette fin, nous avons choisi le paramètre b de manière qu'une proportion d'environ 90% de la somme totale soit prise en compte par les 20% supérieurs de l'ensemble des unités de sondage possibles. Puisque les ensembles de données contiennent un nombre fini de valeurs discrètes, la dérivation des variances des strates différentes lorsque les valeurs de b étaient inférieures à 2 n'a pas posé de problème.

Nous présentons au tableau 3 les résultats de l'application de la méthode L-H pour des tailles de population et des points de départ différents. Le premier groupe utilise des valeurs de départ qui donnent des populations de strates égales ($N_1 = N_2 = N_3$). Le deuxième groupe utilise la méthode de Dalenius-Hodges pour obtenir l'ensemble des bornes initiales. Le troisième groupe obtient les bornes initiales en utilisant une méthode de détermination de la borne de la strate à tirage complet proposée par Hidiroglou (1986) et utilise la méthode de Dalenius-Hodges pour les autres bornes. Ici encore, on peut observer que la surface représentant la taille de l'échantillon est beaucoup plus rugueuse pour les petites populations (voir figure 5). Par exemple, lorsque $N = 50$ et que b_1 est fixe, on n'obtient qu'une seule taille de l'échantillon lorsque b_2 varie entre 11.8 et 14.7. Ceci s'explique par l'absence de valeurs comprises à l'intérieur de cette gamme dans la population. À mesure que la taille de la population augmente, les valeurs des données se rapprochent et la surface de l'échantillon s'adoucit (voir figure 6).

Le tracé de contours correspondant à $N = 50$ (figure 7) présente des formes irrégulières délimitées par des lignes droites. Par contre, le tracé de contours correspondant à $N = 5000$ (figure 8) présente des formes elliptiques concentriques presque lisses. Une forme comparable et concentrique des tracés de contours semblerait être une qualité souhaitable puisque cela signifierait que le minimum global est le seul minimum local.

Le tracé de contours correspondant à $N = 50$ illustre le cas où la méthode L-H n'a pas donné une convergence vers des bornes optimales. Puisque, dans cet exemple, nous avons laissé le programme L-H tourner jusqu'à la convergence, il est permis de se demander pourquoi la méthode L-H n'a pas abouti à une convergence aux bornes optimales. Ce problème s'explique plus facilement à l'examen de la figure 5. Nous constatons en effet que lorsque la taille de la population est petite, la surface de la taille de l'échantillon n'est pas aussi lisse que dans la figure 6. Nous distinguons dans la figure 5 des crêtes importantes causées par les vastes lacunes dans les données discrètes asymétriques ($x_{43} = 9.71, x_{44} = 11.81, x_{45} = 14.79, x_{46} = 19.29$). Ceci signifie que pour une valeur donnée b_1, toute valeur de b_2 comprise entre 11.81 et 14.79 donnerait la même taille d'échantillon. Lorsque nous avons utilisé le programme L-H avec des bornes de départ différentes autres que les trois énumérées au tableau 3, nous avons obtenu les bornes finales du tableau 3 ainsi que d'autres bornes avec leurs valeurs correspondantes de la taille de l'échantillon. Il semble que la méthode L-H converge vers une région inférieure sur l'une des crêtes principales, à condition que cette région se trouve dans le voisinage des bornes optimales. La taille d'échantillon minimale est de 9.22 et la méthode L-H du tableau 3 donne une taille d'échantillon de 9.36. La valeur entière la plus petite de la taille de l'échantillon pour chaque résultat qui répond à la contrainte ou qui la dépasse est 10. Ici encore, nous constatons que la méthode L-H fonctionne exceptionnellement bien même avec des distributions discrètes dont les populations sont petites, puisque les bornes convergent à l'intérieur du voisinage contenant la solution optimale.

Il convient en outre de signaler que les bornes peuvent prendre une large gamme de valeurs tout en maintenant la même valeur entière de la taille de l'échantillon. À mesure que la taille du voisinage s'élargit, la gamme des bornes augmente également. Signalons finalement que même si la gamme des valeurs b_1 pour un voisinage donné est plus petite que la gamme des valeurs b_2, il existe beaucoup plus d'unités de sondage dans la gamme des b_1 que dans celle des b_2 à cause de l'asymétrie de la distribution.

5. CONCLUSION

Les graphiques que nous avons présentés montrent qu'une vaste gamme de bornes donne une gamme étroite de valeurs de la taille de l'échantillon lorsqu'elle se trouve au voisinage d'une valeur optimale (portion inférieure en forme de cuvette des graphiques). Une amélioration additionnelle de la taille de l'échantillon (p. ex., un léger gain marginal) risque de ne pas justifier les efforts supplémentaires nécessaires pour l'obtenir. Il n'est même pas sûr qu'un tel gain marginal contribuera à améliorer la taille de l'échantillon puisque cette dernière est en réalité une valeur entière et que le gain marginal en question pourrait n'être qu'une petite fraction. La méthode L-H s'est avérée très efficace pour le calcul des bornes dans un voisinage souhaité autour d'une valeur optimale, et elle s'est également avérée relativement rapide.

En mesurant le taux de convergence à l'aide de la taille de l'échantillon au lieu des bornes, nous avons été en mesure de mieux déterminer le moment où on atteignait un voisinage souhaitable autour d'une valeur optimale. Les bornes varient en effet grandement dans un tel voisinage tandis que la taille de l'échantillon (le paramètre qui nous intéresse surtout) ne varie que légèrement. Lorsque l'amélioration obtenue dans la taille de l'échantillon d'une itération à l'autre était devenue minimale ou nulle, nous arrêtions immédiatement le programme en présumant que nous avions atteint le voisinage désiré. Les règles d'interruption suivantes sont recommandées. On interrompt le programme lorsque:

1) la différence entre la nouvelle borne supérieure et la borne supérieure de l'itération précédente est inférieure à un. Nous utilisons le nombre entier de un, dans notre cas, puisque les valeurs de la paye n'existent qu'en nombres entiers et qu'un changement des bornes d'une valeur inférieure à un n'influera sur aucune entreprise;
2) la différence entre la nouvelle borne inférieure et celle de l'itération précédente est inférieure à un;
3) la différence entre la nouvelle taille de l'échantillon et celle de l'itération antérieure est inférieure à une petite valeur arbitraire. Nous recommandons un nombre inférieur à un puisque les valeurs de la taille de l'échantillon sont habituellement arrondies et que toute amélioration fractionnaire de la taille de l'échantillon est négligeable. Il convient d'être prudent lors du choix de cette valeur puisqu'il est possible que le taux de réduction de la taille de l'échantillon augmente d'une itération à l'autre par suite d'un changement de la pente de la surface;
4) le programme atteint la 30^{e} itération. Il s'agit évidemment là d'une valeur arbitraire qui peut dépendre du nombre de fois (industries) où il convient d'utiliser la méthode L-H.

Il convient en outre de mentionner que les populations petites peuvent provoquer une convergence des bornes à un point sous-optimal, comme nous l'avons montré dans les exemples. Les graphiques de la taille de l'échantillon montrent une surface rugueuse pour les petites populations et une surface lisse pour les grandes. C'est cette surface rugueuse, due à la nature discrète de la petite population, qui contribue en partie à déterminer le lieu de convergence de la méthode L-H.

Soulignons finalement, à titre de conclusion, que la méthode de Dalenius-Hodges présume que toutes les strates obtenues seront échantillonnées. La méthode L-H est conçue pour établir une sous-strate analytique à tirage complet. Par conséquent, la strate supérieure élaborée par la méthode de Dalenius-Hodges au moment de la création des bornes initiales pour les branches ACES, sera déséquilibrée puisqu'elle ne sera pas échantillonnée. Dans une telle situation, nous avons observé des améliorations dans la taille de l'échantillon en passant de la méthode de Dalenius-Hodges à la première itération de la méthode L-H. Il peut cependant arriver que les bornes de départ conduisent à un minimum local qui ne représente pas la meilleure solution.

REMERCIEMENTS

Les auteurs tiennent à remercier M. Michel Hidiroglou pour ses commentaires utiles. Nous remercions également Carol (Veum) Caldwell, Easley Hoy, les critiques de la revue Techniques d'enquête et les gestionnaires du service de la recherche et de la méthodologie de la Manufacturing and Construction Division pour leurs commentaires utiles formulés au cours de l'examen du présent article.

BIBLIOGRAPHIE

COCHRAN, W.G. (1977). *Sampling Techniques*, (3^{e} éd.). New York: John Wiley and Sons.

DETLEFSEN, R., et VEUM, C. (1991). Design issues for the retail trade sample surveys of the U.S. Bureau of the Census. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 214-219.

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

HESS, I., SETHI, V.K., et BALAKRISHNAN, T.R. (1966). Stratification: A practical investigation. *Journal of the American Statistical Association*, 61, 74-90.

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

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

SCHNEEBERGER, H. (1979). Saddle-points of the variance of the sample mean in stratified sampling. *Sankhȳa*, Series C, 41, 92-96.

SETHI, V.K. (1963). A note on optimum stratification of populations for estimating the population means. *American Journal of Statistics*, 5, 20-23.

Techniques d'enquête, juin 1996
Vol. 22, n° 1, pp. 77–83
Statistique Canada

Nouvelle méthode visant à limiter le nombre d'ondulations indésirables et les corrections lors de l'estimation de la tendance-cycle au moyen du modèle X-11-ARMMI

ESTELA BEE DAGUM[1]

RÉSUMÉ

L'estimation de la tendance-cycle par la méthode X-11-ARMMI est souvent effectuée en appliquant le filtre de Henderson à 13 termes à des données désaisonnalisées, modifiées par des valeurs extrêmes. Cependant, ce filtre produit dans la courbe tendance-cycle finale ou «historique» un grand nombre d'ondulations indésirables qui sont interprétées, erronément, comme des points de retournement. L'utilisation d'un filtre de Henderson plus long, tel que celui à 23 termes, n'est pas une solution, car ce filtre retarde la détection des points de retournement, donc, ne convient pas pour les analyses économiques et commerciales courantes. L'auteure propose une nouvelle méthode d'utilisation du filtre de Henderson à 13 termes ayant l'avantage de i) diminuer le nombre d'ondulations indésirables, ii) réduire l'importance des corrections apportées aux valeurs préliminaires et iii) ne pas augmenter le décalage avec lequel sont décelés les points de retournement. Les résultats de l'application de la méthode à neuf indicateurs avancés de l'Indice composite canadien des indicateurs avancés sont présentés à titre d'illustration.

MOTS CLÉS: Tendance-cycle; X-11-ARMMI; points de retournement; indicateurs économiques avancés.

1. INTRODUCTION

L'estimation de la tendance-cycle par la méthode de désaisonnalisation X-11-ARMMI (Dagum 1980, 1988), ainsi que par la variante X-11 du U.S. Bureau of the Census (Shiskin, Young et Musgrave 1967), est effectuée au moyen de filtres linéaires élaborés par Henderson (1916). Les filtres sont appliqués à des séries désaisonnalisées dont les données irrégulières sont modifiées pour tenir compte des valeurs extrêmes. La longueur du filtre est sélectionnée automatiquement, d'après des valeurs spécifiques du ratio signal/bruit (I/S), le filtre à 13 termes étant choisi le plus couramment.

La question de l'estimation de la tendance-cycle a été étudiée par plusieurs auteurs, dont Rhoades (1980), Cholette (1981, 1982), Kenny et Durbin (1982), Castles (1987), Dagum et Laniel (1987), Cleveland, Cleveland, McRae et Terpenning (1990), Wallgren et Wallgren (1990), Gray et Thomson (1990), Findley et Monsell (1990), Scott (1990), et Kenny (1993). Cependant, la plupart des bureaux de la statistique (sauf l'Australian Bureau of Statistics) se concentrent avant tout sur la publication de données désaisonnalisées, quelques-uns seulement fournissant des renseignements sur la tendance-cycle, ordinairement sous forme graphique.

Plusieurs raisons poussent les bureaux de la statistique à limiter la publication d'estimations de la tendance-cycle. En général, les séries désaisonnalisées sont déjà suffisamment lisses pour donner une indication claire de la tendance à court terme. Dans le cas des séries très volatiles, qui exigent un lissage supplémentaire, les principales objections à l'estimation de la tendance-cycle sont: 1) l'ampleur de la correction des valeurs les plus récentes (en général, beaucoup plus importante que pour les estimations désaisonnalisées correspondantes) et 2) la présence de cycles courts ou ondulations (cycles de 9 et 10 mois) dans la courbe finale de la tendance-cycle quand on utilise le filtre de Henderson à 13 termes. À cet égard, Kenny (1993) argumente que la présence d'ondulations dans les estimations finales de la tendance-cycle provoque la détection d'un grand nombre de faux points de retournement, phénomène qui rend le filtre à 13 termes inapproprié pour la surveillance des renversements de cycle. Il propose l'utilisation du filtre de Henderson à 23 termes en vue d'obtenir une courbe beaucoup plus lisse. Cependant, il est bien connu que ce filtre plus long retarde la détection des points de retournement, donc, ne convient pas aux analyses économiques et commerciales courantes. Pour ce type d'analyse, le filtre de Henderson à 13 termes est préférable. Toutefois, il présente l'inconvénient de produire des ondulations susceptibles d'être interprétées erronément comme des points de retournement.

La présente étude vise principalement à proposer une méthode d'utilisation du filtre de Henderson à 13 termes ayant l'avantage de 1) réduire le nombre d'ondulations indésirables, 2) limiter l'ampleur des corrections apportées aux estimations les plus récentes quand de nouvelles observations sont ajoutées à la série et 3) ne pas augmenter le décalage avec lequel sont décelés les points de retournement.

2. FILTRES EN CASCADE POUR L'ESTIMATION DE LA TENDANCE-CYCLE

Le filtre de Henderson à 13 termes est celui choisi le plus fréquemment. Combiné aux filtres saisonniers type (moyennes mobiles à 5 et à 7 termes), il donne un filtre en

[1] Estela Bee Dagum, Faculté des sciences statistiques, Université de Bologne, Via delle Belle Arti 41, (40126) Bologne, Italie.

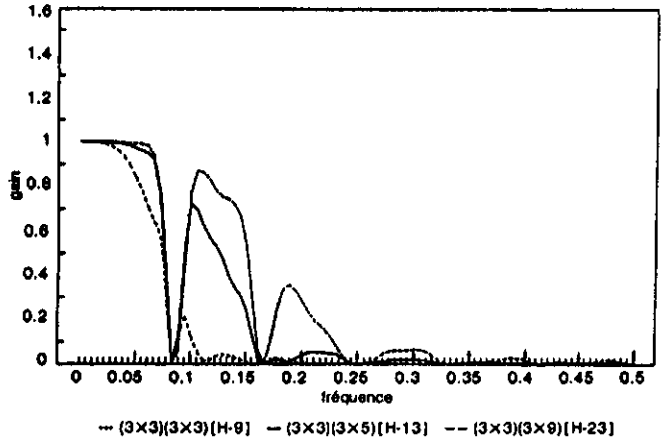

Figure 1. Filtres en cascade symétriques pour l'estimation de la tendance-cycle.

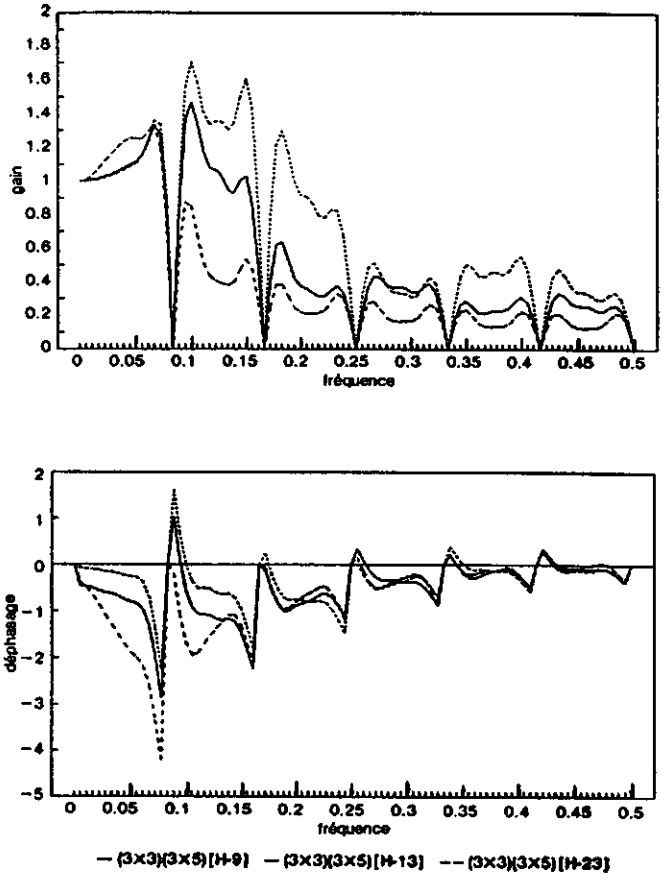

Figure 2. Filtres en cascade concourants de la tendance-cycle. Moyenne mobile saisonnière type combiné aux trois filtres de Henderson.

cascade symétrique applicable aux valeurs finales ou centrales (au moins quatre années à partir de chaque extrémité de la série) produisant une amélioration de l'ordre de celle illustrée à la figure 1.

La figure 1 montre aussi les améliorations obtenues avec d'autres convolutions de filtres, nommément 1) des filtres saisonniers courts combinés au filtre de Henderson à 9 termes et 2) des filtres saisonniers longs combinés au filtre de Henderson à 23 termes. Manifestement, aucun des filtres en cascade utilisés pour estimer la tendance-cycle, particulièrement ceux contenant le filtre de Henderson à 9 ou à 13 termes, ne supprime les cycles de 9 ou 10 mois (bande de fréquences de 0.08 à 0.16). En fait, le filtre en cascade symétrique obtenu au moyen du filtre de Henderson à 9 termes laisse passer environ 90% de l'amplitude de ces cycles courts; ceux contenant le filtre à 13 ou à 23 termes laissent passer 72% et 21%, respectivement.

Si on applique les filtres concourants de la tendance-cycle aux observations les plus récentes, on observe un sommet encore plus élevé dans la bande de fréquences qui correspond aux cycles de 9 et de 10 mois (figure 2). En outre, ces filtres asymétriques produisent tous un déphasage, pratiquement de deux mois pour le filtre à 23 termes, d'un mois pour celui à 13 termes et d'un demi-mois pour celui à 9 termes.

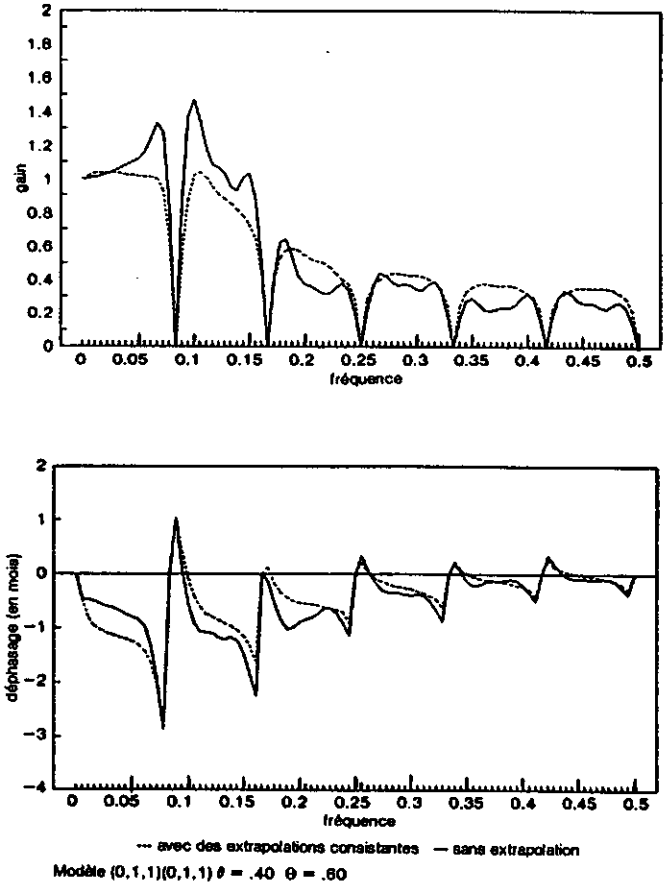

Figure 3. Filtres en cascade concourants de la tendance-cycle, (3×3) (3×5) [H − 13], avec et sans extrapolations ARMMI.

La figure 3 montre que, quand on utilise des extrapolations ARMMI, les gains obtenus au moyen du filtre en cascade concourant (contenant le filtre de Henderson à 13 termes) se rapprochent de ceux produits par les filtres symétriques, moyennant toutefois une légère augmentation du déphasage. Les extrapolations sont celles d'un modèle ARMMI $(0,1,1)$ $(0,1,1)_s$ où le paramètre ordinaire à moyennes mobiles est $\theta = 0.40$ et le paramètre saisonnier à moyennes mobiles est $\Theta = 0.60$.

Le gain et le déphasage obtenus avec ce filtre concourant de la tendance-cycle, qui ne sont pas illustrés faute d'espace, se situent entre ceux produits par les deux autres combinaisons.

Dans le cas d'extrapolations ARMMI, le gain produit par le filtre concourant converge rapidement vers la valeur finale. Dagum et Laniel (1987) montrent qu'après l'ajout de trois observations supplémentaires à la série, le gain obtenu avec le filtre asymétrique est très proche de celui obtenu avec le filtre symétrique. Les propriétés de ces filtres de la tendance-cycle sont également examinées en détail par Dagum, Chhab et Chiu (1993 et 1996).

Le filtre de Henderson à 13 termes produit des ondulations dans l'estimation finale de la tendance-cycle uniquement si les données entrées dans le filtre, c.-à-d. la série désaisonnalisée dont on a remplacé les valeurs extrêmes, présentent une certaine amplitude pour la bande de fréquences de 0.08 à 0.16.

Dans la plupart des cas empiriques, la présence d'ondulations indésirables s'observe pour des périodes de grande volatilité durant lesquelles les données observées sont influencées surtout par des valeurs aberrantes qui peuvent être interprétées erronément comme des points de retournement du cycle. Bien que les séries désaisonnalisées soient corrigées pour tenir compte des valeurs extrêmes, il est nécessaire de les lisser davantage, soit en appliquant un filtre de Henderson plus long, soit en effectuant plus strictement la correction des valeurs extrêmes. Afin de bénéficier de la plus grande rapidité de détection des points de retournement qu'offre un filtre court, la seconde méthode est celle choisie ici.

À l'heure actuelle, les limites implicites sur lesquelles on se fonde pour corriger les valeurs extrêmes qui influent sur la série désaisonnalisée sont fixées à ± 1.5 sigma et ± 2.5 sigma. On attribue un poids nul aux valeurs situées au-delà de ± 2.5 sigma et un poids égal à un à celles situées en deçà de ± 1.5 sigma. Aux valeurs irrégulières comprises entre ± 2.5 sigma et ± 1.5 sigma, on attribue un poids fractionnaire qui varie linéairement de zéro à un.

3. NOUVELLE MÉTHODE

La nouvelle méthode proposée consiste essentiellement 1) à étendre la série désaisonnalisée lissée (modifiée par des valeurs extrêmes ayant un poids nul) au moyen d'extrapolations ARMMI, et 2) à appliquer le filtre de Henderson à 13 termes à la série étendue en fixant des limites standardisées plus strictes pour repérer et corriger les valeurs extrêmes.

L'expérimentation avec des données réelles montre que le spectre d'amplitude de la série désaisonnalisée n'est réduit radicalement aux fréquences variant de 0.8 à 0.16 que si on fixe des limites standardisées strictes, telles que ± 0.7 sigma et ± 1.0 sigma. Dans ces conditions, la courbe de la tendance-cycle obtenue quand on applique le filtre de Henderson à 13 termes ne présente pas d'ondulations indésirables, mais permet toujours de déceler rapidement les points de retournement. Si on émet l'hypothèse que la distribution est normale, ces nouvelles limites standardisées impliquent que 48% des valeurs irrégulières seront modifiées, 32% recevront un poids nul, et 16%, un poids fractionnaire variant de zéro à un.

L'extension de la série désaisonnalisée lissée au moyen d'extrapolations ARMMI est nécessaire pour diminuer la correction des estimations les plus récentes de la tendance-cycle.

Si on se sert du modèle X-11-ARMMI ou X11, l'application de la nouvelle méthode doit comporter les deux étapes suivantes:

1) La meilleure série désaisonnalisée possible est produite en choisissant les options les plus appropriées pour l'estimation des composantes, c'est à dire la saisonnalité, la tendance-cycle, les variations des jours commerciaux et les effets de Pâques, plus les valeurs antérieures permanentes et temporaires, le cas échéant. Les valeurs désaisonnalisées sont présentées au tableau D11. La série désaisonnalisée, corrigée pour tenir compte des valeurs extrêmes de poids nul repérées conformément aux limites standardisées implicites, figure au tableau E2. Si on modifie les estimations de la série désaisonnalisée publiée pour l'année courante conformément à des méthodes de correction particulières, il faut traiter de nouveau la série révisée au moyen du programme X-11-ARMMI pour obtenir les données de sortie correspondant au tableau E2.
2) La série de données de sortie du tableau E2 est étendue d'une année grâce à des données extrapolées au moyen d'un modèle ARMMI. Le modèle (0,1,1) (0,0,1) se révèle adéquat pour nombre de séries réelles. Bien que les données du tableau E2 soient désaisonnalisées, l'emploi du paramètre saisonnier à moyennes mobiles (dont la valeur est souvent très faible) est nécessaire pour tenir compte de certaines autocorrélations saisonnières. On traite ensuite la série étendue au moyen du programme X-11-ARMMI, en choisissant l'option de mesures sommaires et en appliquant des limites standardisées strictes (± 0.7 sigma et ± 1.0 sigma) ainsi que le filtre de Henderson à 13 termes. Les estimations de la tendance-cycle ainsi obtenues figurent au tableau D12.

4. RÉSULTATS EMPIRIQUES

La nouvelle méthode d'estimation de la tendance-cycle est testée au moyen de neuf séries tirées de l'Indice composite canadien des indicateurs avancés. Dans la version dite «filtrée» de l'Indice composite des indicateurs avancés publiée par Statistique Canada, on lisse chaque série constitutive, ainsi que l'Indice, en appliquant aux données désaisonnalisées des filtres asymétriques fondés sur le modèle ARMMI élaboré par Rhoades (1980). Les propriétés spectrales de ces filtres ARMMI de la tendance-cycle sont similaires à celles des points extrêmes des filtres de Henderson à 9, à 13 ou à 23 termes, selon le modèle ARMMI choisi. La comparaison des résultats obtenus avec les filtres ARMMI à ceux observés en appliquant la

nouvelle méthode sort du cadre de la présente étude, mais il est vraisemblable que les seconds seraient également meilleurs (consulter Cholette 1982). La plupart des séries choisies sont extrêmement volatiles et mènent toutes à des points de retournement du cycle économique. Les séries étudiées sont les suivantes:

Indice du cours des actions TSE300
Indice du logement (IL)
Offre de monnaie (M1)
Emploi dans les services aux personnes et aux entreprises (ESPE)
Durée hebdomadaire du travail dans le secteur de fabrication (DHTF)
Ventes au détail de meubles et d'articles ménagers (VDMAM)
Ventes au détail d'autres biens durables (VDBD)
Nouvelles commandes de biens durables (NCBD)
Ratio des livraisons aux stocks (RLS).

Les avantages de la nouvelle méthode comparativement à la méthode X-11-ARMMI utilisée actuellement sont évalués en regard des points suivants.

4.1 Réduction du nombre d'ondulations dans les estimations finales de la tendance-cycle

Afin de calculer la réduction du nombre d'ondulations, nous commençons par donner la définition d'un point de retournement dans le contexte des données de la tendance-cycle. On définit ordinairement un point de retournement comme un point t dans le temps où une des observations d'une série, disons Y_t, a une valeur plus élevée (ou plus faible) que les k observations précédentes et les m observations subséquentes, ou est égale à celles-ci. Autrement dit,

$$Y_{t-k} \leq \ldots \leq Y_{t-1} > Y_t \geq Y_{t+1} \geq \ldots \geq Y_{t+m}$$

définit le passage à une phase descendante, et

$$Y_{t-k} \geq \ldots \geq Y_{t-1} < Y_t \leq Y_{t+1} \leq \ldots \leq Y_{t+m}$$

le passage à une phase ascendante.

En ce qui concerne les séries désaisonnalisées ou les données sur la tendance-cycle, il n'existe pas de consensus quant aux valeurs de k et de m qui définissent un point de retournement. Rhoades (1980) considère qu'un point de retournement est atteint quand $k = 1$ et $m = 0$; Wecker (1979) définit un point de retournement comme la deuxième d'au moins deux baisses ou hausses successives, c.-à-d. $k = 2$ et $m = 2$, et d'autres, dont Zellner, Hong et Min (1991), LeSage (1991), et Pfeffermann et Bleuer (1992) choisissent $k = 3$ et $m = 0$. Ces définitions des points de retournement ne correspondent pas nécessairement à celles utilisées pour l'analyse des cycles économiques, mais n'importe laquelle permet de calculer le nombre d'ondulations, à condition que deux points de retournement (un passage à la baisse et un passage à la hausse) surviennent durant une période de 10 mois ou moins. Nous considérons ici qu'un point de retournement survient quand $k = 3$ et $m = 0$, car les données sur la tendance-cycle sont déjà très lisses.

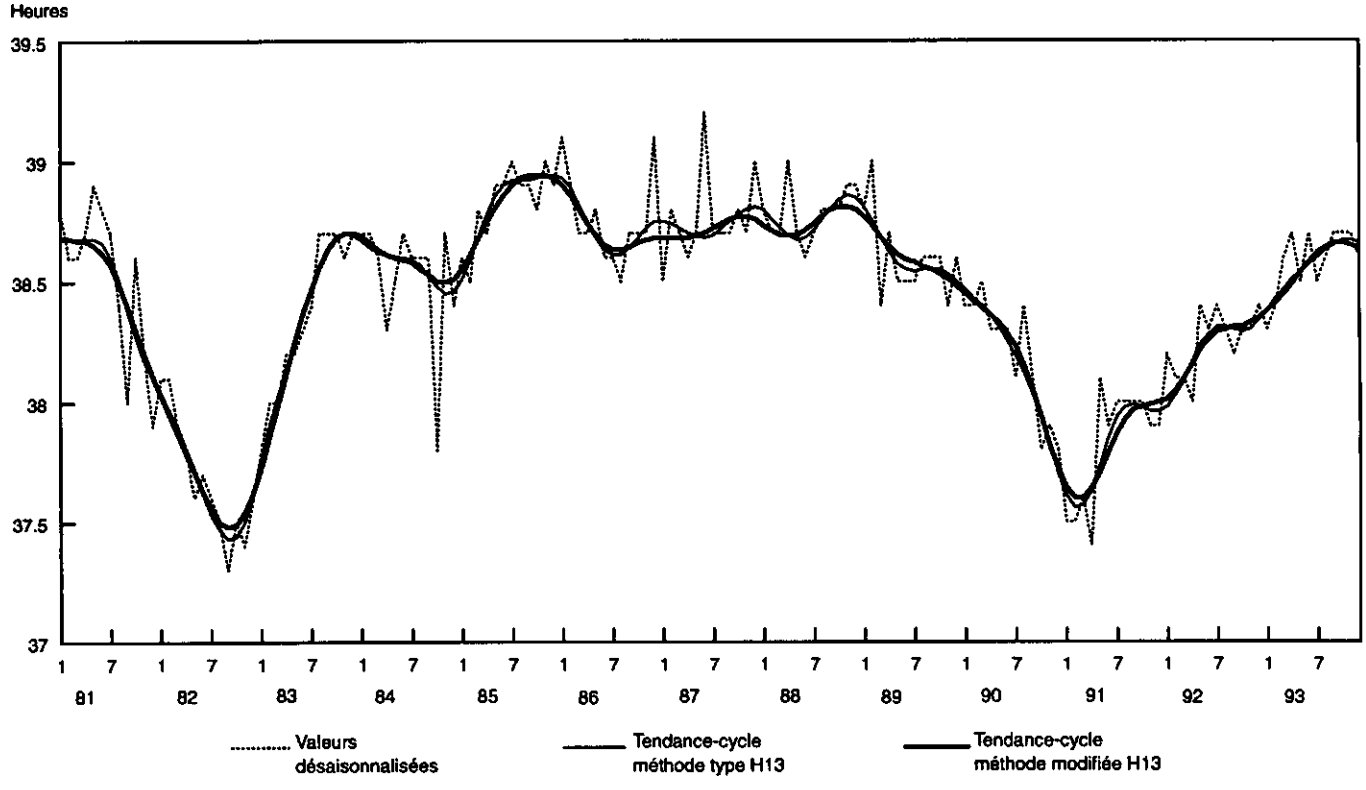

Figure 4. Durée hebdomadaire du travail dans le secteur de fabrication.

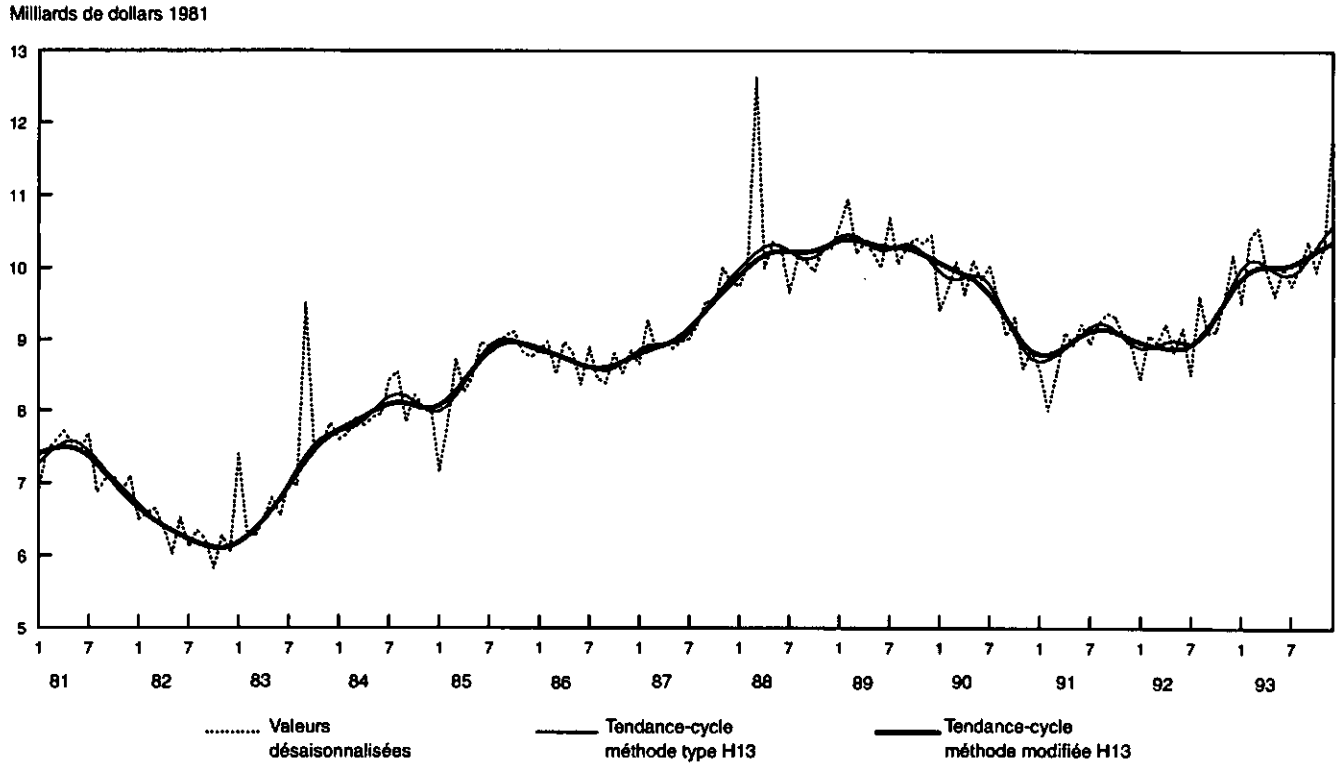

Figure 5. Nouvelles commandes de biens durables.

Le tableau 1 montre le nombre d'ondulations observé dans les estimations de la tendance-cycle quand on applique le filtre type ou le filtre de Henderson à 13 termes modifié pour la période de janvier 1981 à décembre 1993.

Tableau 1

Nombre d'ondulations indésirables dans les données sur la tendance-cycle obtenues en utilisant le filtre de Henderson à 13 termes pour la période de 1981 à 1993

Séries	Méthode type	Méthode modifiée
NCBD	9	2
IL	8	4
VDBD	8	4
ESPE	8	5
DHTF	7	1
RLS	5	1
TSE300	4	2
M1	4	2
VDMAM	4	0

Les résultats indiquent que, d'une part, la réduction du nombre d'ondulations est d'autant plus importante que ce nombre est élevé au départ et, d'autre part, qu'elle est très significative dans tous les cas.

À titre d'illustration, les figures 4 et 5, correspondant aux séries DHTF et NCBD respectivement, montrent les valeurs désaisonnalisées et les données sur la tendance-cycle obtenues en appliquant les méthodes type et modifiée. Ces figures montrent que la nouvelle méthode permet de réduire le nombre d'ondulations décelées dans la courbe de la tendance-cycle, comparativement à la méthode type. En fait, la tendance-cycle obtenue selon la méthode modifiée ressemble à celle produite au moyen du filtre de Henderson à 23 termes, quoiqu'elle coupe davantage les crêtes et les creux des cycles de longue période soit plus prononcée.

4.2 Détection des points de retournement

Il importe que la réduction du nombre d'ondulations dans les estimations finales de la tendance-cycle n'augmente pas le décalage avec lequel sont décelés les points de retournement, phénomène qui constitue la principale limitation du filtre de Henderson à 23 termes.

Pour examiner l'évolution de la correction de la tendance-cycle à n'importe quel point dans le temps, on a calculé les estimations pour tous les points finals et pour les points précédents dans le temps. La comparaison de l'évolution de la correction de l'estimation de la tendance-cycle selon la méthode modifiée et selon la méthode type montre qu'en moyenne, le décalage avec lequel les points de retournement du cycle sont décelés est similaire pour les deux méthodes. Selon la série examinée, l'écart observé entre les décalages est nul, ou de l'ordre de plus ou moins un mois. À titre d'illustration, la figure 6a. montre l'évolution de la correction de l'estimation, selon la méthode modifiée, de la tendance-cycle des Nouvelles commandes de biens durables pour le point de retournement du cycle de février 1991. Des mises à jour successives sont effectuées au moyen de données allant jusqu'à mars 1991, avril 1991

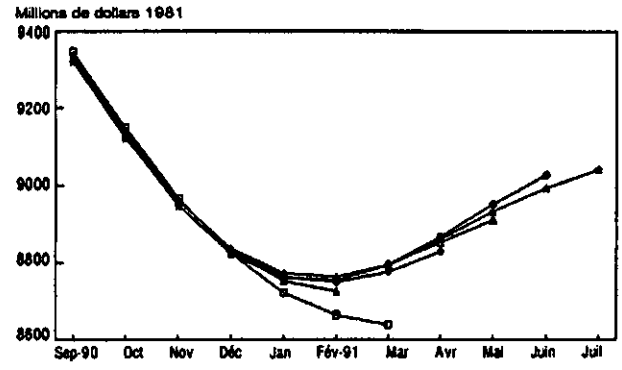

Figure 6a. Nouvelles commandes de biens durables. Évolution de la correction de l'estimation selon la méthode modifiée de la tendance-cycle H13.

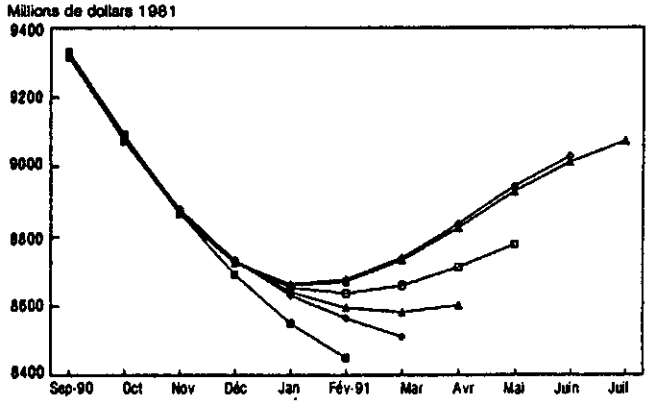

Figure 6b. Nouvelles commandes de biens durables. Évolution de la correction de l'estimation selon la méthode type de la tendance-cycle H13.

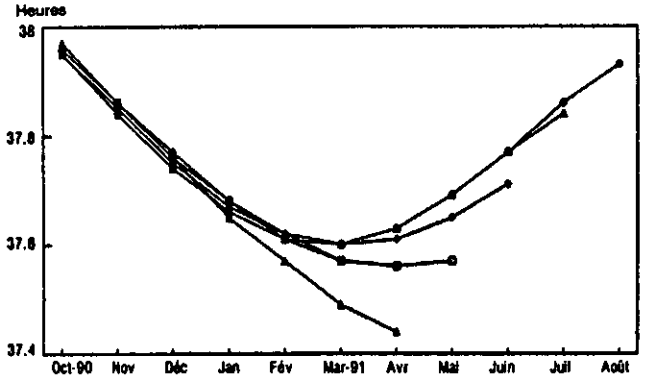

Figure 7a. Durée hebdomadaire de travail dans le secteur de fabrication. Évolution de la correction de l'estimation selon la méthode modifiée de la tendance-cycle H13.

et ainsi de suite. Le point de retournement est reconnu en avril, après deux mois, tandis qu'en appliquant la méthode type, sa détection prend trois mois, comme l'illustre la figure 6b. En outre, les figures 6a. et 7a. montrent que les estimations révisées successives de la tendance-cycle restent généralement très proches des valeurs finales. Les lignes qui font saillie, indiquant une correction importante, sont

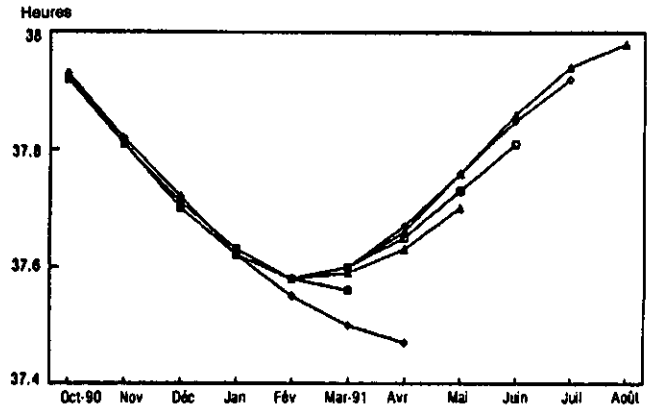

Figure 7b. Durée hebdomadaire de travail dans le secteur de fabrication. Évolution de la correction de l'estimation selon la méthode type de la tendance-cycle H13.

expliquées par les données sous-jacentes qui semblent signaler une baisse croissante contredite par les valeurs suivantes.

Les figures 7a. et 7b., qui correspondent à la Durée hebdomadaire du travail dans le secteur de la fabrication, indiquent que le point de retournement de février-mars 1991 est décelé avec un décalage de trois mois par les deux méthodes.

4.3 Diminution de la correction des estimations concourantes de la tendance-cycle

La réduction de la correction totale des estimations les plus récentes de la tendance-cycle, qui sont de nature provisoire, est un autre aspect important dont il faut tenir compte. Théoriquement, on obtient l'estimation finale de la tendance-cycle après avoir ajouté quatre années de données à la série, mais les corrections deviennent très petites après l'ajout des données de trois mois.

Le tableau 2 montre le pourcentage moyen absolu de correction des estimations concourantes de la tendance-cycle durant une période de quatre ans allant de janvier 1988 à décembre 1991. Les résultats indiquent que, dans six des neuf cas analysés, la correction totale des valeurs concourantes de la tendance-cycle est beaucoup plus faible pour la méthode modifiée que pour la méthode type, et qu'elle n'est légèrement plus importante que dans deux cas.

Tableau 2

Pourcentage moyen absolu de correction totale des valeurs concourantes de la tendance-cycle au moyen du filtre de Henderson à 13 termes

Séries	Méthode type (1)	Méthode modifiée (2)	Rapport (2)/(1)
NCBD	1.55	1.10	0.73
VDMAM	0.62	0.47	0.76
VDBD	0.77	0.62	0.80
RLS	0.87	0.70	0.80
DHTF	0.13	0.12	0.92
TSE300	1.12	1.07	0.95
M1	0.35	0.35	1.00
IL	2.09	2.20	1.05
ESPE	0.40	0.42	1.05

5. CONCLUSION

L'article vise à présenter une nouvelle méthode d'estimation de la tendance-cycle au moyen du filtre de Henderson à 13 termes ayant l'avantage de i) diminuer le nombre d'ondulations indésirables dans la courbe finale de la tendance-cycle, ii) réduire l'importance des corrections apportées aux valeurs concourantes préliminaires, et iii) ne pas augmenter le décalage de temps avec lequel sont décelés les points de retournement.

La nouvelle méthode consiste essentiellement à étendre la série désaisonnalisée lissée (modifiée par des valeurs extrêmes de poids nuls) d'une année au moyen d'extrapolations fondées sur le modèle ARMMI, puis, à appliquer le filtre de Henderson à 13 termes en fixant des limites standardisées strictes pour repérer et corriger les valeurs aberrantes.

L'application de la méthode à neuf séries tirées de l'Indice composite canadien des indicateurs avancés donne des résultats très satisfaisants.

REMERCIEMENTS

Le présent article se fonde sur des travaux financés par Statistique Canada. Toutefois, les opinions exprimées sont celles de l'auteure et ne reflètent pas nécessairement celles de l'organisme. J'exprime ma gratitude à Marietta Morry et à Norma Chhab, du Centre de recherche et d'analyse en séries chronologiques, pour les innombrables discussions enrichissantes que nous avons eues et pour leur collaboration aux travaux qui ont mené à la publication de cet article. Je remercie également un évaluateur anonyme pour les commentaires précieux qu'il a faits au sujet d'une version antérieure.

BIBLIOGRAPHIE

CASTLES, I. (1987). A Guide to Smoothing Time Series Estimates of Trend. N° 1316.0 au catalogue, Australian Bureau of Statistics.

CHOLETTE, P.A. (1981). A comparison of various trend-cycle estimators. Dans *Time Series Analysis*. (O.D. Anderson et M.R. Perryman, Éds). Amsterdam: North-Holland, 77-87.

CHOLETTE, P.A. (1982). Comparaison de deux estimateurs des cycles économiques. Document de recherche n° 82-09-OO1F, Centre de recherche et d'analyse en séries chronologiques, Statistique Canada.

CLEVELAND, R., CLEVELAND, W.S., McRAE, J.E., et TERPENNING, I. (1990). STL: A seasonal-trend decomposition procedure based on Loess. *Journal of Official Statistics*, 6, 3-33.

DAGUM, E.B. (1980). La méthode de désaisonnalisation X-11-ARMMI. N° 12-564F au catalogue. Statistique Canada.

DAGUM, E.B. (1988). The X-11-ARIMA/88 Seasonal Adjustment Method - Foundations and User's Manual. Centre de recherche et d'analyse en séries chronologiques, Statistique Canada.

DAGUM, E.B., et LANIEL, N. (1987). Revisions of trend-cycle estimators of moving average seasonal adjustment methods. *Journal of Business and Economic Statistics*, 5, 177-189.

DAGUM, E.B., CHHAB, N., et CHIU, K. (1993). Linear properties of the X-11-ARIMA seasonal adjustment method. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*.

DAGUM, E.B., CHHAB, N., et CHIU, K. (1996). Derivation and properties of the Census X-11 variant and the X-11-ARIMA linear filters. *Journal of Official Statistics*, (à paraître).

FINDLEY, D.F., et MONSELL, B.C. (1990). Comment (sur Cleveland et coll. 1990). *Journal of Official Statistics*, 6, 55-59.

GRAY, A.G., et THOMSON, P.J. (1990). Comment (sur Cleveland et coll. 1990). *Journal of Official Statistics*, 6, 47-54.

HENDERSON, R.(1916). Note on graduation by adjusted average. *Transactions of the Actuarial Society of America*, 17, 43-48.

KENNY, P. (1993). Trend presentation. T02919, SMQ, Branch, Central Statistical Office, Londres, Angleterre.

KENNY, P.B., et DURBIN, J. (1982). Local trend estimation and seasonal adjustment of economic and social time series. *Journal of the Royal Statistical Society*, Séries A, 145, 1-41.

LeSAGE, J.P. (1991). Analysis and development of leading indicators using a Bayesian turning-points approach. *Journal of Business and Economic Statistics*, 9, 305-316.

PFEFFERMANN, D., et BLEUER, S.R. (1992). Probabilistic detection of nonseasonal turning points in economic time series estimated from sample surveys. Rapport interne, Direction de la méthodologie, Statistique Canada, Ottawa.

RHOADES, D. (1980). La conversion de l'actualité en fiabilité des séries chronologiques économiques ou filtrage des séries chronologiques économiques occasionnant un déphasage minimum. *Revue statistique du Canada*, 6-13.

SCOTT, S. (1990). Comment (sur Cleveland et coll. 1990). *Journal of Official Statistics*, 6, 59-62.

SHISKIN, J., YOUNG, A.H., et MUSGRAVE, J.C. (1967). The X-11 Variant of Census Method II Seasonal Adjustment. Technical Paper No. 15, U.S. Bureau of the Census.

WALLGREN, B., et WALLGREN, A. (1990). Comment (sur Cleveland et coll. 1990). *Journal of Official Statistics*, 6, 39-46.

WECKER, W. (1979). Predicting the turning points of a series. *Journal of Business*, 52, 35-50.

ZELLNER, A., HONG, C., et MIN, C. (1991). Forecasting turning points in international output growth rates using Bayesian exponentially weighted autoregression, time-varying parameter, and pooling techniques. *Journal of Econometrics*, 48, 275-304.

Techniques d'enquête, juin 1996
Vol. 22, n° 1, pp. 85–94
Statistique Canada

Un algorithme de stratification mobile

YVES TILLÉ[1]

RÉSUMÉ

Un algorithme général à probabilités égales est présenté. On donne les probabilités d'inclusion d'ordre deux qui correspondent à cet algorithme. Celui-ci généralise la méthode de sélection-rejet permettant de constituer un échantillon au moyen d'un plan simple sans remise. Un autre cas particulier de cet algorithme baptisé «algorithme de stratification mobile» est discuté. Il permet d'obtenir un effet de stratification lissé en utilisant comme variable de stratification le numéro d'ordre des unités d'observation. On donne des approximations des probabilités d'inclusion d'ordre un et deux. Ces approximations permettent de construire un estimateur de la moyenne de la population ainsi qu'un estimateur de la variance de cet estimateur de moyenne. Cet algorithme est ensuite comparé à un plan stratifié classique avec allocation proportionnelle.

MOTS CLÉS: Algorithme de tirage; sondage à probabilités égales; strate.

1. INTRODUCTION

Quand un fichier est ordonné selon une variable auxiliaire proche de la variable d'intérêt, comment peut-on sélectionner un échantillon en tirant profit de cette information? Une solution à ce problème consiste à effectuer un tirage stratifié. Cependant, pour effectuer un tel tirage, il faut résoudre le délicat problème du découpage de la population en strates. Une autre solution simple, rapide et efficace consiste à recourir à un tirage systématique. L'algorithme s'écrit en quelques lignes. De plus, on sait que l'on profitera de la manière dont le fichier est ordonné. Le tirage systématique a cependant un défaut important: pour estimer les variances des estimateurs de totaux ou de moyennes, il est nécessaire d'énoncer une ou plusieurs hypothèses sur la population. Nous montrons qu'il existe un autre algorithme de tirage simple permettant de constituer un échantillon en un seul passage en tirant profit de l'ordonnancement du fichier. Pour cet algorithme, nous donnons un estimateur de la variance de l'estimateur d'un total ou d'une moyenne qui ne nécessite pas une modélisation de la population.

Dans la section 2, un algorithme général de tirage donnant des probabilités d'inclusion d'ordre un égales est présenté. On donne les probabilités d'inclusion d'ordre un et deux. Dans la section 3, on montre que l'algorithme présenté généralise la méthode de sélection-rejet permettant de constituer un échantillon simple sans remise et le plan stratifié avec allocation proportionnelle. Enfin, dans la section 4, on définit la méthode de la strate mobile et les conclusions sont données en section 5.

2. PRÉSENTATION DE L'ALGORITHME GÉNÉRAL

2.1 L'algorithme

Soit $U = \{1, \ldots, i, \ldots, N\}$ une population finie, on note $y_1, \ldots, y_i, \ldots, y_N$, les N valeurs prises par la variable y sur les N unités d'observation de U. La moyenne des valeurs prises par la variable y sur la population est notée

$$\bar{y} = \frac{1}{N} \sum_{i \in U} y_i.$$

Un échantillon aléatoire s de taille fixe n est prélevé dans cette population. Les variables aléatoires indicatrices de la présence des unités d'observation dans s sont notées I_i, $i \in U$. La probabilité d'inclusion d'ordre un est notée $\pi_i = \Pr(i \in s) = E(I_i)$, $i \in U$, et la probabilité d'inclusion d'ordre deux $\pi_{ik} = E(I_i I_k)$, $i \neq k \in U$. L'algorithme est très court. Il s'apparente aux algorithmes de Fan, Fuller et Rezucha (1962), Bebbington (1975), McLeod et Bellhouse (1983) et de Sunter (1977 et 1986). Il faut connaître seulement N, n et les b_i, $i = 0, \ldots, N-1$. Les autres variables sont des variables de travail.

Algorithme général

$j <= 0$;
$i <= 0$;
Répéter pour $i = 0, \ldots, N-1$
  $u <=$ un nombre aléatoire selon une loi uniforme [0,1];
  si $\dfrac{(b_i + i)n/N - j}{b_i} > u$ alors sélectionner l'enregistrement $i+1$; $j <= j+1$;
  sinon, passer l'enregistrement $i+1$;
  $i <= i+1$.

À chaque étape, j représente le nombre d'enregistrements déjà sélectionnés et i le nombre d'enregistrements passés (sélectionnés ou non). À chaque itération, on décide de la sélection de l'enregistrement $i+1$. Si cet enregistrement est sélectionné, il devient le $(j+1)$-ième de l'échantillon. Les coefficients b_i, $i = 0, \ldots, N-1$, sont des réels strictement positifs. Ces quantités doivent satisfaire à certaines conditions discutées plus bas pour que le plan

[1] Yves Tillé, Laboratoire de Méthodologie du Traitement des Données, C.P. 124, Université Libre de Bruxelles, Avenue Jeanne, 44, 1050 Bruxelles, Belgique, E-mail ytilleb@ulb.ac.be

soit de taille fixe ou pour que les unités soient sélectionnées à probabilités égales. Le choix de différentes valeurs pour les b_i, $i = 0, \ldots, N-1$, permettra de déterminer plusieurs cas particuliers de l'algorithme général.

Si les b_i sont des réels strictement positifs tel que $b_i \leq N - i$, alors la taille de l'échantillon est inférieure ou égale à n. En effet, supposons que l'on ait déjà prélevé n unités dans la population à l'étape i et que $b_i \leq N - i$, alors

$$\frac{(b_i + i)n/N - n}{b_i} = \frac{n}{N} - \frac{n}{b_i}\frac{N-i}{N} \leq \frac{n}{N} - \frac{n}{N-i}\frac{N-i}{N} = 0.$$

Il est dès lors impossible de prélever une unité supplémentaire. On supposera pour tout ce qui suit que $b_i \leq N - i$. Par ailleurs, si $b_i \leq N - i$, $i = 1, \ldots, N - n - 1$, et que $b_i = N - i$, $i = N - n, \ldots, N - 1$, alors la taille de l'échantillon est de taille fixe n. Soulignons que ces conditions pour obtenir un échantillon de taille fixe sont suffisantes mais non nécessaires.

Trois cas particuliers de l'algorithme seront étudiés plus loin. Ces trois cas sont définis par trois choix de coefficients b_i, $i = 0, \ldots, N-1$. Avant d'étudier ces choix particuliers, nous allons déterminer les probabilités d'inclusion d'ordre un et deux en toute généralité.

2.2 Probabilités d'inclusion d'ordre un

Notons n_i, le nombre d'unités sélectionnées après avoir passé i enregistrements. On voit directement que $n_1, \ldots, n_i, \ldots, n_N$ est une chaîne de Markov. En effet, il découle directement de l'algorithme que

$$\Pr[n_i = j \mid n_1, \ldots, n_{i-1}] = \Pr[n_i = j \mid n_{i-1}].$$

Les variables aléatoires

$$c_i = \frac{(b_i + i)n/N - n_i}{b_i}, \; i = 0, \ldots, N-1,$$

peuvent parfois prendre des valeurs supérieures à 1 ou inférieures à 0. Comme $\max(0, n - N + i) \leq n_i \leq \min(i, n)$, on a que $\Pr[0 \leq c_i \leq 1] = 1$ si

$$b_i \geq \begin{cases} \min\left(i\dfrac{N-n}{n}, N - i\right) & \text{si } n \leq N/2 \\ \min\left(i\dfrac{n}{N-n}, N - i\right) & \text{si } n > N/2 \end{cases}, \quad i = 0, \ldots, N-1. \quad (1)$$

Les conditions (1) sont à nouveau suffisantes mais non nécessaires. On peut donc construire des b_i qui ne satisfont pas à ces conditions mais qui fournissent des c_i dans $[0,1]$. Le cas traité en section 3.2 (stratification) en donne un exemple.

L'exemple suivant fournit également des c_i dans $[0,1]$ sans respecter la condition (1): prenons $N = 12$, $n = 4$ et $b_0 = b_1 = b_3 = b_4 = b_6 = 6$, $b_2 = b_5 = 7$, $b_i = N - i$, $i = 12 - i$, $i = 7, \ldots, 11$. On obtient $c_0 = 1/3$, $c_1 = (7 - 3n_1)/18$, $c_2 = (3 - n_2)/7$, $c_3 = (3 - n_3)/6$, $c_4 = (10 - 3n_4)/18$, $c_5 = (4 - n_5)/7$, $c_6 = (4 - n_6)/6$, $c_7 = (4 - n_7)/5$, $c_8 = (4 - n_8)/4$, $c_9 = (4 - n_9)/3$, $c_{10} = (4 - n_{10})/2$, $c_{11} = (4 - n_{11})$. Remarquons que $n_1 \leq 1$, $n_2 \leq 2$, $n_3 \leq 3$. Si $n_3 = 3$ alors $c_3 = 0$ et donc $n_4 \leq 3$. Ensuite, on obtient $n_5 \leq 4$ et si $n_5 = 4$ alors $c_5 = 0$ et donc $n_6 \leq 4$. Cette dernière remarque est vraie pour toutes les c_i qui suivent. On constate donc que tous les c_i sont dans $[0,1]$ alors que $b_4 = 6$ ne respecte pas la condition (1).

Pour simplifier les démonstrations qui suivent, on supposera par la suite que

$$\Pr[0 \leq c_i \leq 1] = 1, \; i = 0, \ldots, N-1.$$

Nous reviendrons sur le problème des c_i ayant des valeurs supérieures à 1 ou inférieures à 0 plus loin. Si

$$\Pr[0 \leq c_i \leq 1] = 1, \; i = 0, \ldots, N-1,$$

on a

$$E[I_{i+1} \mid n_1, \ldots, n_i] = E[I_{i+1} \mid n_i] = \frac{(b_i + i)n/N - n_i}{b_i}.$$

On montre aisément par récurrence que si $\Pr[0 \leq c_i \leq 1] = 1$, $i = 0, \ldots, N-1$, $E[n_i] = i\,n/N$, $i = 0, \ldots, N$. Donc,

$$\pi_i = E[I_i] = E[n_i] - E[n_{i-1}] = \frac{n}{N}. \quad (2)$$

2.3 Probabilités d'inclusion d'ordre deux

Quatre résultats donnés par les lemmes 1, 2 et 3 sont nécessaires pour déterminer les probabilités d'inclusion d'ordre deux.

Lemme 1 Si $\Pr[0 \leq c_i \leq 1] = 1$, $i = 0, \ldots, N-1$, alors

$$E[n_{i+k} \mid n_i] = (i + k)\frac{n}{N} + \left(n_i - i\frac{n}{N}\right)\prod_{t=i}^{i+k-1}\frac{b_t - 1}{b_t},$$

$$i = 1, \ldots, N-1, \; k = 1, \ldots, N - i.$$

Ce lemme se montre par récurrence en supposant qu'il est vrai pour $k - 1$. Par le lemme 1, on obtient aisément par différence le lemme suivant:

Lemme 2 Si $\Pr[0 \le c_i \le 1] = 1$, $i = 0, \ldots, N-1$, alors

$$E[I_{i+k} \mid n_i] = \frac{n}{N} - \left(n_i - i\frac{n}{N}\right)\frac{1}{b_{i+k-1}} \prod_{\ell=i}^{i+k-2} \frac{b_\ell - 1}{b_\ell},$$

$$i = 1, \ldots, N-1, k = 1, \ldots, N-i.$$

On suppose par convention qu'un produit vide vaut 1.

Lemme 3 Si $\Pr[0 \le c_i \le 1] = 1$, $i = 0, \ldots, N-1$, alors

$$\mathrm{Var}[n_i] = \frac{n}{N}\frac{N-n}{N} \sum_{j=1}^{i} \prod_{\ell=j}^{i-1} \frac{b_\ell - 2}{b_\ell}, i = 1, \ldots, N. \quad (3)$$

La démonstration est donnée en annexe.

Enfin, la probabilité d'inclusion d'ordre deux est donnée par la proposition suivante:

Proposition 1 Si $\Pr[0 \le c_i \le 1] = 1$, $i = 0, \ldots, N-1$, alors

$$E[I_{i+k}I_{i+1}] = \frac{n^2}{N^2} - \frac{n}{N}\frac{N-n}{N}\frac{1}{b_{i+k-1}} \times \left(1 - \frac{1}{b_i}\sum_{j=1}^{i}\prod_{\ell=j}^{i-1}\frac{b_\ell - 2}{b_\ell}\right)\prod_{\ell=i+1}^{i+k-2}\frac{b_\ell - 1}{b_\ell},$$

$$i = 0, \ldots, N-2, k = 2, \ldots, N-i. \quad (4)$$

La démonstration est donnée en annexe.

Corollaire 1 Si $\Pr[0 \le c_i \le 1] = 1$, $i = 0, \ldots, N-1$, alors

$$\pi_{ik} = \frac{n^2}{N^2} - \frac{n}{N}\frac{N-n}{N}\left(1 - \frac{1}{b_{i-1}}\sum_{j=1}^{i-1}\prod_{\ell=j}^{i-2}\frac{b_\ell - 2}{b_\ell}\right) \times \frac{1}{b_{k-1}}\prod_{\ell=i}^{k-2}\frac{b_\ell - 1}{b_\ell}, \; i = 1, \ldots, N-1, k > i.$$

2.4 L'estimateur de Horvitz-Thompson et sa variance

L'estimateur de Horvitz-Thompson est la moyenne échantillon simple puisque les probabilités d'inclusion d'ordre un sont toutes égales

$$\hat{\bar{y}}_\pi = \frac{1}{n}\sum_{i\in s} y_i.$$

Si le plan est de taille fixe, on peut utiliser la formule de la variance de Yates et Grundy (1953)

$$\mathrm{Var}[\hat{\bar{y}}_\pi] = \frac{1}{2N^2}\sum_{i\in U}\sum_{\substack{k\in U\\ k\neq i}}\left(\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k}\right)^2(\pi_i\pi_k - \pi_{ik}). \quad (5)$$

Comme $\pi_i = n/N$, $i = 1, \ldots, N$, et en posant

$$\gamma_{ik} = 1 - \pi_{ik}\frac{N^2}{n^2},$$

on peut écrire

$$\mathrm{Var}[\hat{\bar{y}}_\pi] = \frac{1}{N^2}\sum_{i\in U}\sum_{\substack{k\in U\\ k\neq i}}(y_i - y_k)^2\gamma_{ik}. \quad (6)$$

L'estimateur de la variance est donné par

$$\widehat{\mathrm{Var}}[\hat{\bar{y}}_\pi] = \frac{1}{2N^2}\sum_{i\in s}\sum_{\substack{k\in s\\ k\neq i}}\left(\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k}\right)^2\frac{\pi_i\pi_k - \pi_{ik}}{\pi_{ik}}. \quad (7)$$

Ce qui peut s'écrire ici

$$\widehat{\mathrm{Var}}[\hat{\bar{y}}_\pi] = \frac{1}{2n^2}\sum_{i\in s}\sum_{\substack{k\in s\\ k\neq i}}(y_i - y_k)^2\frac{\gamma_{ik}}{1-\gamma_{ik}}.$$

3. APPLICATION 1: TIRAGES ALÉATOIRES SIMPLE ET STRATIFIÉ

3.1 Plan simple

L'algorithme de tirage le plus simple, la méthode de sélection-rejet telle qu'elle est décrite dans Fan, Fuller et Rezucha (1962, méthode 1), Bebbington (1975) et dans Deville et Grosbras (1987, p. 210) est bien sûr un cas particulier de l'algorithme général. Il suffit de prendre

$$b_i = N - i, \quad i = 0, \ldots, N-1.$$

On a toujours $0 \le c_i \le 1$. Les probabilités d'inclusion d'ordre un valent toujours n/N. Le calcul des probabilités d'inclusion d'ordre deux découle de la proposition 1. En supposant $k > i$, par le corollaire 1, on peut retrouver les probabilités d'inclusion d'ordre deux du plan simple:

$$\pi_{ik} = \frac{n(n-1)}{N(N-1)}.$$

Rappelons également quelques résultats classiques concernant le plan simple que nous utiliserons plus loin. L'estimateur de $\bar{y}$ est donc la moyenne de l'échantillon

$$\hat{y}_{srs} = \frac{1}{n} \sum_{i \in s} y_i. \tag{8}$$

La variance de cet estimateur est donnée par

$$\mathrm{Var}[\hat{y}_{srs}] = \frac{\sigma_y^2}{n} \frac{N-n}{N-1} \tag{9}$$

où

$$\sigma_y^2 = \frac{1}{N} \sum_{i \in U} (y_i - \bar{y})^2. \tag{10}$$

Cette variance peut être estimée sans biais par

$$\widehat{\mathrm{Var}}[\hat{y}_{srs}] = \frac{s_y^2}{n} \frac{N-n}{N} \tag{11}$$

où

$$s_y^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \hat{y}_{srs})^2. \tag{12}$$

3.2 Plan stratifié

On peut définir également le plan stratifié au moyen de l'algorithme général. La variable de stratification est ici le numéro d'ordre de l'individu. Considérons le cas particulier d'un plan stratifié en H strates avec allocation proportionnelle où toutes les strates ont la même taille. Les strates sont constituées de telle manière que les individus d'une même strate soient contigus sur le fichier de données. On suppose, en outre, que N/H est entier. Ce plan stratifié s'obtient en prenant simplement

$$b_i = \left\{ (N - i - 1) \bmod \frac{N}{H} \right\} + 1, \; i = 0, \ldots, N-1.$$

4. APPLICATION 2: STRATIFICATION MOBILE

4.1 Le problème

Le fichier est supposé être ordonné selon une variable auxiliaire proche de la variable d'intérêt. Le problème est le suivant: Comment effectuer un tirage aléatoire qui donne une petite variance pour l'estimateur de Horvitz-Thompson d'une moyenne? En examinant la formulation de la variance de Yates-Grundy (5), on voit que l'on peut répondre à cette question de deux manières bien distinctes.

La première solution consiste à tirer à probabilités inégales avec des probabilités d'inclusion d'ordre un proportionnelles à la variable d'intérêt. S'il est possible d'effectuer un tel tirage, toutes les quantités

$$\left(\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k} \right)^2$$

seraient nulles et donc la variance serait nulle.

La seconde solution consiste à jouer sur les probabilités d'inclusion d'ordre deux. Un bon tirage pourrait être tel que les π_{ik} soient proches de $\pi_i \pi_k$ si y_i est très différent de y_k. D'autre part, si y_i est très proche de y_k, on peut choisir des probabilités d'inclusion d'ordre deux π_{ik} nettement inférieures à $\pi_i \pi_k$. Ainsi, quand les quantités

$$\left(\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k} \right)^2$$

seraient grandes (resp. petites), les quantités $\pi_i \pi_k - \pi_{ik}$ seraient petites (resp. grandes). On obtiendrait donc une petite variance.

La seconde solution que nous venons de décrire est en fait déjà abondamment utilisée. C'est l'idée de base de la stratification. Notre objectif est d'appliquer cette idée pour construire un algorithme séquentiel de tirage facile à mettre en oeuvre. Cet algorithme pourrait s'appliquer à n'importe quel fichier sans qu'il soit nécessaire de connaître autre chose que la taille de la population. Il serait donc applicable à de très gros fichiers. On profiterait ainsi de l'information apportée par cette variable auxiliaire comme pour une stratification, sans pour autant devoir se préoccuper d'un réel découpage en strates.

4.2 La méthode

On définit d'abord M, la longueur de la strate mobile dans la population. M représente en quelque sorte la taille de la strate dans la population et est telle que $N/n \leq M \leq N$. L'algorithme de la strate mobile est défini par

$$b_i = \min(M, N - i), \; i = 0, \ldots, N-1.$$

Un problème se pose cependant. Les quantités c_i définies par

$$c_i = \begin{cases} \dfrac{(M+i)n/N - n_i}{M} & \text{si} \quad i \leq N - M \\ \dfrac{n - n_i}{N - i} & \text{sinon,} \end{cases}$$

ne sont pas toujours dans $[0,1]$.

En effet, supposons que, avant la $(N-M)$-ième étape de l'algorithme, c_i soit positif et très proche de 0 et que par malchance l'unité i soit malgré cela choisie. Dans ce cas, c_{i+1} prendrait la valeur $c_i - (N-n)/(NM)$. c_{i+1} peut donc prendre une valeur négative mais cette valeur négative est toujours supérieure à $-(N-n)/(NM)$. En effet, dans le cas où un des c_i est déjà négatif, l'unité i n'est pas sélectionnée et donc c_{i+1} prend une valeur supérieure à c_i.

Supposons maintenant que avant la $(N-M)$-ième étape de l'algorithme, un c_i soit très légèrement inférieur à 1 et que malgré cela l'unité i ne soit pas sélectionnée. Dans ce cas, c_{i+1} prendrait la valeur $c_i + n/(NM)$. c_{i+1} peut donc prendre une valeur plus grande que 1 mais cette

valeur supérieure à 1 est cependant toujours inférieure à $1 + n/(NM)$. En effet, dans le cas où un des c_i est déjà supérieur à 1, l'unité i est toujours sélectionnée et donc c_{i+1} prend une valeur inférieure à c_i.

On obtient

$$\Pr\left[-\frac{N-n}{NM} < c_i < 1 + \frac{n}{NM}\right] = 1, i = 0, \ldots, N - M. \tag{13}$$

Le plan est cependant de taille fixe, ce résultat est donné par la proposition suivante:

Proposition 2 Si $b_i = \min(M, N - i)$, $(N/n < M < N)$, $0 = 1, \ldots, N - 1$, alors le plan est de taille fixe.

La démonstration est donnée en annexe.

Comme les c_i ne sont pas toujours dans l'intervalle $[0,1]$, nous avons effectué 50 simulations d'application de l'algorithme de la strate mobile pour différentes tailles de population et d'échantillon. Les tailles de population N choisies sont 100, 500, 2500, 12500, 62500, 312500. Les inverses des taux de sondages (N/n) sont 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096. Nous avons effectué plusieurs simulations en faisant varier la taille de la strate mobile ainsi: $M = N/n, 2N/n, 3N/n, \ldots$. Les simulations semblent indiquer que plus M est grand, plus la probabilité qu'un c_i soit en dehors de $[0,1]$ est faible. Dès que $M \geq 10N/n$, pour toutes les simulations que nous avons effectuées, le problème ne s'est plus jamais posé. Ce premier résultat n'implique pas que la probabilité qu'au moins un des c_i soit hors de $[0,1]$ est nulle quand $M \geq 10N/n$. Cependant on peut penser que, dans ce cas, cette probabilité est très faible.

4.3 Estimation de la moyenne et biais

En considérant les résultats donnés par l'expression (2) et la proposition 1, on obtient, en première approximation, que les probabilités d'inclusion d'ordre un valent approximativement $\pi_i \approx n/N$. Cette approximation des probabilités d'inclusion permet de construire un estimateur.

$$\hat{\bar{y}}_{sm} = \frac{1}{n} \sum_{i \in s} y_i.$$

Cet estimateur est légèrement biaisé car les c_i ne sont pas tous exactement dans l'intervalle $[0,1]$. Ce biais vaut

$$B\left[\hat{\bar{y}}_{sm}\right] = \frac{1}{N} \sum_{i \in U} \alpha_i y_i$$

où $\alpha_i = \pi_i N/n - 1$. Comme le plan est de taille fixe, $\sum_{i \in U} \alpha_i = 0$. On peut donc écrire le biais sous la forme d'une covariance: $B\left[\hat{\bar{y}}_{sm}\right] = \sigma_{y\alpha}$ où

$$\sigma_{y\alpha} = \frac{1}{N} \sum_{i \in U} \alpha_i (y_i - \bar{y}). \tag{14}$$

Comme la valeur absolue d'une covariance est toujours inférieure ou égale au produit des deux écart-types, on obtient un majorant pour la valeur absolue du biais

$$\left| B\left[\hat{\bar{y}}_{sm}\right] \right| \leq \sigma_y \sigma_\alpha$$

où σ_y est défini par (10) et

$$\sigma_\alpha^2 = \frac{1}{N} \sum_{i \in U} \alpha_i^2.$$

La variance de l'estimateur est d'un ordre de grandeur comparable (pour N et n fixé) à la variance de l'estimateur de la moyenne dans le plan simple sans remise. On peut donc écrire

$$\left| B\left[\hat{\bar{y}}_{sm}\right] \right| \leq C_\alpha \sqrt{\mathrm{Var}\left[\hat{\bar{y}}_{srs}\right]}$$

où $\mathrm{Var}\left[\hat{\bar{y}}_{srs}\right]$ est défini par (9) et

$$C_\alpha = \sigma_\alpha \sqrt{\frac{n(N-1)}{(N-n)}}.$$

Nous allons considérer que le biais est négligeable quand le majorant du biais de l'estimateur $\hat{\bar{y}}_{sm}$ est négligeable par rapport à $\mathrm{Var}\left[\hat{\bar{y}}_{srs}\right]^{1/2}$, c'est-à-dire quand C_α est petit.

On peut calculer récursivement la valeur exacte des $\Pr[n_i = j]$ car on a

$$\Pr[I_i = 1 \mid n_i] = \tilde{c}_i, i = 1, \ldots, N - M$$

où $\tilde{c}_i$ vaut 0 si $c_i < 0$, c_i si $0 \leq c_i \leq 1$ et 1 si $c_i > 1$. De ce résultat, on peut déduire la valeur exacte des probabilités d'inclusion d'ordre un.

Nous avons calculé (annexe, tableau 1) les valeurs de C_α pour diverses tailles de population (100 à 312500) et d'échantillon. Les valeurs de C_α sont données pour des tailles de strates mobiles M égales à N/n, $2N/n$, $3N/n$, $4N/n$ et $5N/n$. On voit que dès que la strate mobile vaut $2N/n$, C_α ne dépasse jamais 0.07. Quand $M = 3N/n$, le coefficient C_α s'exprime en millièmes. Selon Cochran (1977, pp. 13-14), le biais peut alors être négligé. Ce tableau montre donc que si $M \geq 3N/n$, le biais de l'estimateur sera négligeable au moins pour les tailles des populations et des échantillons spécifiés.

Ces résultats n'impliquent cependant pas que le biais de l'estimateur est grand quand M est très petit (par exemple $M = N/n$). Les C_α sont des majorants des biais. Par l'expression (14), on voit que le biais sera d'autant plus grand que la variable d'intérêt est corrélée aux probabilités d'inclusion exactes. Nous avons représenté (figure 1), les probabilités d'inclusion exactes (en ordonnée) pour les N individus (en abscisse) obtenues par l'algorithme de stratification mobile avec les paramètres $N = 51$, $n = 11$, $M = N/n$. Ce cas est évidemment très défavorable. Le résultat est intéressant. Dans ce cas, $n/N = 0.215686$. Les probabilités d'inclusion se répartissent de part et d'autre

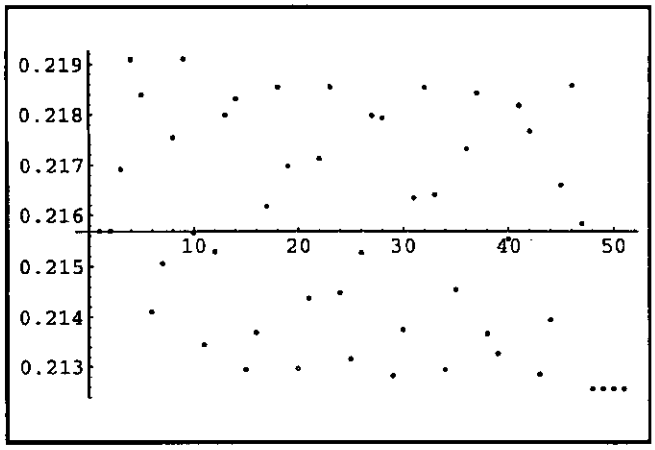

Figure 1. Probabilités d'inclusion.

de n/N sans qu'il y ait une tendance notable liée à l'ordre du fichier. En pratique, on peut considérer qu'il est très peu probable d'avoir une variable d'intérêt fortement corrélée aux probabilités d'inclusion exactes et le biais sera donc, le plus souvent, nettement inférieur au majorant donné.

On pourrait, bien sûr, utiliser les probabilités d'inclusion exactes pour procéder à une estimation. Nous pensons que cela n'en vaut pas la peine pour deux raisons:

- d'abord parce que le calcul des probabilités d'inclusion exactes nécessite un temps de calcul important,
- ensuite parce que les probabilités d'inclusion d'ordre un exactes sont telles que

$$\mathrm{Var}\left[\sum_{i\in s}\frac{1}{\pi_i}\right] \neq 0.$$

Dans ce cas, l'estimateur de Horvitz-Thompson d'une variable constante ($y_k = C$) est aléatoire. Pour pallier ce problème, on estime généralement la moyenne par le ratio de Hájek (1971). Cet estimateur est également biaisé.

4.4 Estimation de la variance de l'estimateur

En supposant que $\Pr(0 \leq c_i \leq 1) \approx 1$, on peut construire également une approximation des probabilités d'inclusion d'ordre deux à partir de corollaire 1. En considérant que b_i vaut M si $i \leq N - M$ et $N - i$ sinon, on obtient l'approximation suivante:

$$\pi_{ik} \approx \frac{n^2}{N^2}(1 - \theta_{ik})$$

où

$$\theta_{ik} = \frac{N-n}{2n}\frac{1}{M-1}\left\{1 + \left(\frac{M-2}{M}\right)^{\min(i-1,N-M)}\right\} \times \left(\frac{M-1}{M}\right)^{\max(0,\min(N-M-i+1,k-i))} \quad k > i.$$

En supposant que les probabilités d'inclusion d'ordre un valent n/N, on peut construire une approximation de la variance de $\hat{y}_{sm}$:

$$\mathrm{Var}_{app}[\hat{y}_{sm}] = \frac{1}{2N^2}\sum_{i\in U}\sum_{\substack{k\in U\\ k\neq i}}(y_i - y_k)^2\theta_{ik}. \qquad (15)$$

À partir de (15), on peut construire un estimateur de la variance de l'estimateur de moyenne:

$$\widehat{\mathrm{Var}}_{app}[\hat{y}_{sm}] = \frac{1}{2N^2}\sum_{i\in s}\sum_{\substack{k\in s\\ k\neq i}}(y_i - y_k)^2\frac{\theta_{ik}}{1-\theta_{ik}}. \qquad (16)$$

De nouveau, cet estimateur est biaisé. Afin de juger de l'ampleur de ce biais, nous avons effectué un ensemble de simulations. Les résultats sont donnés dans le tableau 2 en annexe. Nous avons généré des populations de taille $N = 400$. Les valeurs prises par les deux variables x et y ont été générées au moyen de nombres pseudo-aléatoires ayant une distribution normale bivariée avec un coefficient de corrélation ρ fixé. Ensuite, les populations ont été triées en fonction de la variable x. L'objectif est d'estimer $\bar{y}$.

Dans ces populations, des échantillons de taille 64 ont été sélectionnés au moyen de la méthode de la strate mobile (*sm*), d'un plan stratifié avec allocation proportionnelle où les tailles des strates sont toutes égales (*strat*) et d'un plan simple sans remise (*srs*). Ces trois méthodes sont des cas particuliers de l'algorithme général et ont été mises en oeuvre en utilisant les mêmes nombres aléatoires. Des simulations ont été effectuées pour différentes valeurs de la strate mobile M (cas: *sm*) et pour différents nombres de strates H (cas: *strat*). Nous donnerons plus loin une explication sur les choix de M et H. Pour chaque simulation, 200,000 échantillons ont été sélectionnés.

Pour chacune des simulations, on donne 3 résultats:

- les moyennes sur les simulations des estimateurs de variance de l'estimateur de la moyenne qui sont notées $E_{sim}\widehat{\mathrm{Var}}(\hat{\bar{y}})$. Ces estimateurs de variance sont donnés par les expressions (11) (*srs*) et (16) (*sm*).
- les erreurs quadratiques moyennes sur les simulations des estimateurs de moyenne. Ces quantités sont notées $EQM_{sim}(\hat{\bar{y}}) = E_{sim}(\hat{\bar{y}} - \bar{y})^2$.
- les variances des estimateurs de moyenne. Ces variances sont données par les expressions (9) (*srs*) et (15) (*sm*). Dans le cas de la stratification mobile, il s'agit bien sûr de l'approximation proposée.

Une lecture attentive des résultats semble indiquer que l'estimateur de variance proposé pour l'algorithme de la strate mobile ne souffre pas d'un biais systématique quelle que soit la valeur du coefficient de corrélation entre x et y. Les résultats semblent également indiquer que l'expression approximative donnée pour la variance de l'estimateur de moyenne pour la stratification mobile est une approximation valable.

4.5 Intérêt de l'algorithme

Dans la classe des algorithmes définis par l'algorithme général, on appelle horizon moyen d'un algorithme, la quantité

$$\bar{b} = \frac{1}{N} \sum_{i=0}^{N-1} b_i.$$

Pour le plan simple, on obtient $\bar{b}_{srs} = (N + 1)/2$. Pour l'algorithme de la strate mobile, on a

$$\bar{b}_{sm} = \frac{1}{N} \left\{ \sum_{i=0}^{N-M-1} M + \sum_{i=N-M}^{N-1} (N - i) \right\}$$

$$= \frac{M}{N} \left\{ N - \frac{M-1}{2} \right\}.$$

Supposons maintenant que nous sélectionnions, comme décrit en section 3.2, un échantillon au moyen d'un plan avec allocation proportionnelle où toutes les strates ont la même taille et où les tailles des H strates sont toutes égales. Dans un tel plan, l'horizon moyen vaut

$$\bar{b}_{strat} = \frac{1}{2} \left(\frac{N}{H} + 1 \right).$$

Un changement d'horizon moyen n'affecte pas fondamentalement les probabilités d'inclusion d'ordre un. Les probabilités d'inclusion d'ordre deux sont par contre fortement modifiées par un changement d'horizon. En effet, on voit aisément que plus l'horizon moyen est petit, plus la probabilité de sélectionner deux individus proches est faible. (On dit que deux individus sont proches si la valeur absolue de la différence de leur numéro d'ordre sur le fichier de données est petite.) Intuitivement, on peut donc s'attendre à ce que l'algorithme de la strate mobile ait un effet de stratification similaire à un plan stratifié avec allocation proportionnelle ayant le même horizon moyen, c'est-à-dire quand

$$\bar{b}_{strat} = \bar{b}_{sm},$$

ou autrement dit, quand

$$M = N + \frac{1}{2} - \sqrt{\frac{1}{4} + N^2 \frac{H-1}{H}}. \qquad (17)$$

Quand N est grand par rapport à M, on a approximativement

$$M \approx \frac{2N}{H}.$$

Pour chaque ensemble de simulations présenté en annexe (tableau 2), les tailles des strates mobiles (cas: *sm*) ont été fixées en fonction du nombre de strates (cas: *strat*) de manière à ce que les horizons moyens des deux plans soient identiques au sens de l'expression (17). On constate que, dans ce cas, le gain de précision (comparativement au plan simple) obtenu au moyen de l'algorithme de la strate mobile est du même ordre de grandeur qu'au moyen de la stratification.

5. REMARQUES

Les simulations effectuées montrent clairement que l'algorithme de stratification mobile donne un effet de stratification du même type qu'une stratification classique avec allocation proportionnelle. Cet algorithme permet d'éluder le délicat problème du découpage d'une variable continue en strates. Les estimateurs de moyenne proposés sont légèrement biaisés. Cependant, dès que $M \geq 10N/n$, les simulations montrent qu'il est extrêmement rare qu'au moins un des c_i soit en dehors de $[0,1]$. De plus, nous avons montré que même quand cette probabilité n'est pas nulle, le biais de l'estimateur que nous proposons est négligeable dès que $M \geq 3N/n$.

REMERCIEMENTS

L'auteur remercie Pierre Lavallée pour les conseils qu'il a exprimés sur des versions antérieures de cet article. L'auteur tient également à remercier un éditeur associé et un arbitre qui ont émis de nombreux commentaires constructifs qui ont permis d'améliorer considérablement cet article.

ANNEXE 1

Démonstration des lemmes et des propositions

Démonstration du lemme 3

$$\mathrm{Var}[n_{i+1}]$$

$$= \mathrm{Var}[n_i] + \mathrm{Var}[I_{i+1}]$$

$$+ 2E\left(E\left\{\left(n_i - i\frac{n}{N}\right)E\left[I_{i+1} - \frac{n}{N} \mid n_i\right]\right\}\right).$$

Comme

$$2E\left[E\left\{\left(n_i - i\frac{n}{N}\right)E\left[\left(I_{i+1} - \frac{n}{N}\right) \mid n_i\right]\right\}\right]$$

$$= 2E\left[\left(n_i - i\frac{n}{N}\right)\left(\frac{(b_i + i)n/N - n_i}{b_i} - \frac{n}{N}\right)\right]$$

$$= \frac{-2}{b_i}\mathrm{Var}[n_i],$$

on obtient

$$\mathrm{Var}[n_{i+1}] = \mathrm{Var}[n_i]\frac{b_i - 2}{b_i} + \frac{n}{N}\frac{N-n}{N},$$

$$i = 1, \ldots, N-1. \qquad (18)$$

On montre ensuite que (3) vérifie l'équation de récurrence (18) et la condition initiale donnée par

$$\mathrm{Var}(n_1) = \frac{n}{N}\frac{N-n}{N}.$$

Démonstration de la proposition 1

Cas 1: $i = 0$. Par le lemme 2, on a directement:

$$E[I_k I_1] = E[E[I_k \mid n_1]n_1]$$

$$= \frac{n^2}{N^2} - \frac{n}{N}\frac{N-n}{N}\frac{1}{b_{k-1}}\prod_{t=1}^{k-2}\frac{b_t - 1}{b_t}.$$

Cas 2: $i > 0$. En utilisant le lemme 2, on obtient:

$$E[I_{i+k}I_{i+1} \mid n_i = t]$$

$$= E[I_{i+k} \mid n_{i+1} = t+1]E[I_{i+1} \mid n_i = t]$$

$$= \left\{\frac{n}{N} - \left((t+1) - (i+1)\frac{n}{N}\right)\frac{1}{b_{i+k-1}}\prod_{t=i+1}^{i+k-2}\frac{b_t - 1}{b_t}\right\}$$

$$\times \left\{\frac{n}{N} - \left(t - i\frac{n}{N}\right)\frac{1}{b_i}\right\}.$$

Ce qui donne

$$E[E[I_{i+k}I_{i+1} \mid n_i]]$$

$$= E\left\{\frac{n}{N} - \left((n_i+1) - (i+1)\frac{n}{N}\right)\frac{1}{b_{i+k-1}}\prod_{t=i+1}^{i+k-2}\frac{b_t - 1}{b_t}\right\}$$

$$\times \left\{\frac{n}{N} - \left(n_i - i\frac{n}{N}\right)\frac{1}{b_i}\right\}$$

$$= \frac{n^2}{N^2} - \frac{1}{b_{i+k-1}}\left\{\frac{n}{N}\frac{N-n}{N} - \frac{\mathrm{Var}[n_i]}{b_i}\right\}\prod_{t=i+1}^{i+k-2}\frac{b_t - 1}{b_t}.$$

Or, le lemme 3 nous donne $\mathrm{Var}[n_i]$. On obtient directement (4).

Démonstration de la proposition 2

Par (13), on a

$$\Pr\left[n - M - \frac{n}{N} < n_{N-M} < \frac{N-n}{N} + n\right] = 1.$$

Donc,

$$\Pr[0 \leq n - n_{N-M} \leq M] = 1.$$

À partir de l'étape $N - M$, l'algorithme est un algorithme de sélection-rejet tel qu'il est décrit dans la section 3.1. Cet algorithme aboutit au prélèvement de exactement $n - n_{N-M}$ unités d'observation durant les M dernières étapes. Comme $n - n_{N-M} \leq M$, cette opération ne pose pas de problème et l'algorithme est donc de taille fixe n.

ANNEXE 2

Tableaux, majorants du biais et simulations

Tableau 1

Valeur des majorants du biais C_α

N	n	Valeur du coefficient C_α				
		$M=\frac{N}{n}$	$M=\frac{2N}{n}$	$M=\frac{3N}{n}$	$M=\frac{4N}{n}$	$M=\frac{5N}{n}$
100	50	0.000000	0.000000	0.000000	0.000000	0.000000
	25	0.057326	0.002610	0.000185	0.000015	0.000001
	12	0.041716	0.002604	0.000235	0.000023	0.000002
	6	0.032227	0.002029	0.000134	0.000005	0.000000
	3	0.023515	0.000645	0.000000		
500	250	0.000000	0.000000	0.000000	0.000000	0.000000
	125	0.129091	0.006002	0.000437	0.000038	0.000004
	62	0.090863	0.005664	0.000534	0.000059	0.000007
	31	0.066891	0.004666	0.000484	0.000059	0.000008
	15	0.048544	0.003586	0.000384	0.000046	0.000006
	7	0.035508	0.002552	0.000215	0.000015	0.000001
	3	0.024046	0.000699	0.000000		
2,500	1,250	0.000000	0.000000	0.000000	0.000000	0.000000
	625	0.289060	0.013495	0.000987	0.000086	0.000008
	312	0.202458	0.012607	0.001190	0.000133	0.000016
	156	0.147113	0.010234	0.001064	0.000130	0.000017
	78	0.105662	0.007742	0.000841	0.000107	0.000015
	39	0.075975	0.005719	0.000634	0.000082	0.000012
	19	0.054525	0.004174	0.000466	0.000060	0.000008
	9	0.039560	0.003014	0.000301	0.000029	0.000002
	4	0.028388	0.001451	0.000034	0.000000	
12,500	3,125	0.646539	0.030208	0.002211	0.000193	0.000018
	1,562	0.452450	0.028177	0.002661	0.000297	0.000036
	781	0.327879	0.022798	0.002371	0.000290	0.000039
	390	0.234114	0.017131	0.001863	0.000238	0.000033
	195	0.166626	0.012500	0.001388	0.000181	0.000026
	97	0.118357	0.008995	0.001009	0.000133	0.000019
	48	0.084217	0.006452	0.000727	0.000096	0.000014
	24	0.060797	0.004689	0.000529	0.000069	0.000010
	12	0.044677	0.003461	0.000377	0.000044	0.000005
	6	0.033727	0.002356	0.000173	0.000008	0.000000
	3	0.024172	0.000712	0.000000		
62,500	3 906	0.732684	0.050942	0.005299	0.000649	0.000087
	1,953	0.522918	0.038250	0.004159	0.000531	0.000074
	976	0.371301	0.027833	0.003092	0.000403	0.000057
	488	0.263300	0.019979	0.002243	0.000295	0.000042
	244	0.186736	0.014259	0.001609	0.000213	0.000031
	122	0.132653	0.010168	0.001150	0.000152	0.000022
	61	0.094601	0.007273	0.000823	0.000109	0.000016
	30	0.067467	0.005207	0.000590	0.000078	0.000011
	15	0.049227	0.003820	0.000427	0.000054	0.000007
	7	0.035847	0.002637	0.000227	0.000016	0.000001
	3	0.024176	0.000713	0.000000		
312,500	4,882	0.829762	0.062191	0.006909	0.000901	0.000128
	2,441	0.587909	0.044596	0.005006	0.000659	0.000095
	1,220	0.416165	0.031758	0.003583	0.000474	0.000068
	610	0.294647	0.022555	0.002551	0.000339	0.000049
	305	0.208743	0.016008	0.001813	0.000241	0.000035
	152	0.147877	0.011356	0.001287	0.000171	0.000025
	76	0.105272	0.008098	0.000918	0.000122	0.000018
	38	0.075422	0.005817	0.000659	0.000087	0.000013
	19	0.054695	0.004238	0.000479	0.000062	0.000009
	9	0.039644	0.003038	0.000305	0.000030	0.000002
	4	0.028427	0.001457	0.000034	0.000000	

Tableau 2

Résultats des simulations, plan simple, stratification et stratification mobile

ρ^2	Plan	Paramètres	$E_{sim}\widehat{\text{Var}}\,\bar{y}$	$\text{Var}\bar{y}$	$EQM_{sim}\bar{y}$
0.0	*sm*	$M = 18.83N/n$	0.01318	0.01317	0.01301
	srs		0.01317	0.01316	0.01296
	strat	$H = 2$	0.01319	0.01319	0.01318
0.2	*sm*	$M = 18.83N/n$	0.01210	0.01210	0.01187
	srs		0.01316	0.01316	0.01287
	strat	$H = 2$	0.01172	0.01188	0.01164
0.4	*sm*	$M = 18.83N/n$	0.01073	0.01073	0.01080
	srs		0.01316	0.01316	0.01320
	strat	$H = 2$	0.00943	0.00929	0.00946
0.6	*sm*	$M = 18.83N/n$	0.00957	0.00957	0.00954
	srs		0.01315	0.01316	0.01301
	strat	$H = 2$	0.00783	0.00778	0.00774
0.8	*sm*	$M = 18.83N/n$	0.00839	0.00839	0.00839
	srs		0.01315	0.01316	0.01322
	strat	$H = 2$	0.00630	0.00624	0.00622
1.0	*sm*	$M = 18.83N/n$	0.00757	0.00757	0.00760
	srs		0.01314	0.01316	0.01319
	strat	$H = 2$	0.00514	0.00508	0.00513
0.0	*sm*	$M = 8.65N/n$	0.01319	0.01319	0.01317
	srs		0.01317	0.01316	0.01296
	strat	$H = 4$	0.01320	0.01318	0.01316
0.2	*sm*	$M = 8.65N/n$	0.01107	0.01107	0.01084
	srs		0.01316	0.01316	0.01287
	strat	$H = 4$	0.01080	0.01076	0.01054
0.4	*sm*	$M = 8.65N/n$	0.00876	0.00876	0.00882
	srs		0.01316	0.01316	0.01320
	strat	$H = 4$	0.00811	0.00793	0.00796
0.6	*sm*	$M = 8.65N/n$	0.00695	0.00694	0.00688
	srs		0.01315	0.01316	0.01301
	strat	$H = 4$	0.00637	0.00639	0.00632
0.8	*sm*	$M = 8.65N/n$	0.00484	0.00484	0.00485
	srs		0.01315	0.01316	0.01322
	strat	$H = 4$	0.00402	0.00391	0.00390
1.0	*sm*	$M = 8.65N/n$	0.00312	0.00312	0.00313
	srs		0.01314	0.01316	0.01319
	strat	$H = 4$	0.00206	0.00197	0.00197
0.0	*sm*	$M = 4.21N/n$	0.01317	0.01317	0.01316
	srs		0.01317	0.01316	0.01296
	strat	$H = 8$	0.01321	0.01324	0.01325
0.2	*sm*	$M = 4.21N/n$	0.01067	0.01067	0.01046
	srs		0.01316	0.01316	0.01287
	strat	$H = 8$	0.01055	0.01047	0.01025
0.4	*sm*	$M = 4.21N/n$	0.00810	0.00809	0.00808
	srs		0.01316	0.01316	0.01320
	strat	$H = 8$	0.00794	0.00789	0.00789
0.6	*sm*	$M = 4.21N/n$	0.00592	0.00592	0.00588
	srs		0.01315	0.01316	0.01301
	strat	$H = 8$	0.00575	0.00564	0.00561
0.8	*sm*	$M = 4.21N/n$	0.00344	0.00344	0.00345
	srs		0.01315	0.01316	0.01322
	strat	$H = 8$	0.00315	0.00311	0.00308
1.0	*sm*	$M = 4.21N/n$	0.00124	0.00124	0.00125
	srs		0.01314	0.01316	0.01319
	strat	$H = 8$	0.00085	0.00079	0.00080

Tableau 2

Résultats des simulations, plan simple, stratification et stratification mobile - fin

ρ^2	Plan	Paramètres	$E_{sim}\widehat{\text{Var}}\,\hat{\bar{y}}$	$\text{Var}\,\hat{\bar{y}}$	$EQM_{sim}\hat{\bar{y}}$
0.0	*sm*	$M = 2.11N/n$	0.01319	0.01319	0.01328
	srs		0.01315	0.01316	0.01332
	strat	$H = 16$	0.01315	0.01308	0.01331
0.2	*sm*	$M = 2.11N/n$	0.01038	0.01036	0.01021
	srs		0.01317	0.01316	0.01334
	strat	$H = 16$	0.01034	0.01034	0.01025
0.4	*sm*	$M = 2.11N/n$	0.00796	0.00796	0.00792
	srs		0.01316	0.01316	0.01323
	strat	$H = 16$	0.00790	0.00801	0.00794
0.6	*sm*	$M = 2.11N/n$	0.00572	0.00573	0.00561
	srs		0.01315	0.01316	0.01299
	strat	$H = 16$	0.00568	0.00572	0.00563
0.8	*sm*	$M = 2.11N/n$	0.00295	0.00294	0.00290
	srs		0.01317	0.01316	0.01325
	strat	$H = 16$	0.00287	0.00288	0.00285
1.0	*sm*	$M = 2.11N/n$	0.00048	0.00048	0.00048
	srs		0.01317	0.01316	0.01335
	strat	$H = 16$	0.00037	0.00034	0.00034
0.0	*sm*	$M = 1.09N/n$	0.01325	0.01316	0.01310
	srs		0.01313	0.01316	0.01317
	strat	$H = 32$	0.01201	0.01239	0.01302
0.2	*sm*	$M = 1.09N/n$	0.01070	0.01062	0.01064
	srs		0.01313	0.01316	0.01316
	strat	$H = 32$	0.00972	0.01018	0.01083
0.4	*sm*	$M = 1.09N/n$	0.00807	0.00803	0.00811
	srs		0.01315	0.01316	0.01309
	strat	$H = 32$	0.00732	0.00751	0.00803
0.6	*sm*	$M = 1.09N/n$	0.00538	0.00534	0.00536
	srs		0.01315	0.01316	0.01310
	strat	$H = 32$	0.00484	0.00484	0.00543
0.8	*sm*	$M = 1.09N/n$	0.00283	0.00281	0.00276
	srs		0.01317	0.01316	0.01283
	strat	$H = 32$	0.00255	0.00276	0.00280
1.0	*sm*	$M = 1.09N/n$	0.00016	0.00016	0.00017
	srs		0.01317	0.01316	0.01304
	strat	$H = 32$	0.00012	0.00007	0.00011

BIBLIOGRAPHIE

BEBBINGTON, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 136.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: Wiley.

DEVILLE, J.-C., et GROSBRAS, J.-M. (1987). Algorithmes de tirage. Dans *Les sondages*. Droesbeke, J.-J., Fichet, B., et Tassi, P. (éds.). Paris: Economica, 209-233.

FAN, C.T., MULLER, M.E., et REZUCHA, I. (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computers. *Journal of the American Statistical Association*, 57, 387-402.

HÁJEK, J. (1971). Comment on an essay of D. Basu. Dans *Foundations of Statistical Inference*. Godambe V.P., et Sprott, D.A. (éds). Toronto: Holt, Rinehart et Winston.

McLEOD, A.I., et BELLHOUSE, D.R. (1983). A convenient algorithm for drawing a simple random sampling. *Applied Statistics*, 32, 182-184.

SUNTER, A.B. (1977). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26, 261-268.

SUNTER, A.B. (1986). Solutions to the problem of unequal probability sampling without replacement. *Revue Internationale de Statistique*, 54, 33-50.

YATES, F., et GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, B, 15, 235-261.

Techniques d'enquête, juin 1996
Vol. 22, n° 1, pp. 95–104
Statistique Canada

Aperçu du contrôle statistique de la divulgation des microdonnées

A.G. de WAAL et L.C.R.J. WILLENBORG[1]

RÉSUMÉ

Les problèmes que pose le contrôle statistique de la divulgation, lequel a pour but d'empêcher les utilisateurs des données de divulguer des renseignements sur des répondants particuliers, se sont multipliés rapidement au cours des dernières années. La situation est due principalement à l'augmentation de la demande de données détaillées provenant des bureaux de la statistique, elle-même causée par l'accroissement continuel de l'usage des ordinateurs. Auparavant, ces bureaux produisaient des tableaux contenant relativement peu d'information. Aujourd'hui, par contre, les utilisateurs de données demandent des tableaux beaucoup plus détaillés et, qui plus est, des microdonnées à analyser eux-mêmes. Or, l'augmentation du contenu informatique des données rend le contrôle statistique de la divulgation beaucoup plus difficile. Les auteurs se fondent sur l'expérience qu'ils ont acquise dans le domaine du contrôle statistique de la divulgation à Statistics Netherlands pour exposer les problèmes qu'il faut, selon eux, surmonter quand on essaie de protéger les microdonnées contre la divulgation.

MOTS CLÉS: Contrôle statistique de la divulgation; microdonnées; unicité.

1. INTRODUCTION

Le contrôle statistique de la divulgation (CSD) est un domaine qui acquiert de plus en plus d'importance en raison de l'intérêt croissant pour les renseignements produits par les bureaux de la statistique. Ces renseignements se subdivisent en deux grandes catégories, à savoir les données tabulaires et les microdonnées. Si la diffusion de tableaux est une activité de longue date des bureaux de la statistique, celle des ensembles de microdonnées, quant à elle, est relativement récente. Par le passé, les utilisateurs de données ne possédaient généralement pas les outils nécessaires pour analyser correctement les ensembles de microdonnées. Aujourd'hui, par contre, tout chercheur qui se respecte possède un ordinateur personnel puissant. L'analyse des microdonnées n'est donc plus le privilège des bureaux de la statistique puisque les utilisateurs de données sont désormais capables de l'effectuer eux-mêmes et en manifestent l'intention. Les problèmes que pose cette situation en matière de CSD sont loin d'être bénins.

Un des principaux obstacles auxquels se heurtent les théoriciens du CSD des microdonnées consiste à calculer la probabilité qu'un enregistrement inclus dans un ensemble diffusé de microdonnées soit reconnu. Diverses méthodes, dont les objectifs varient considérablement, ont été proposées pour estimer cette probabilité. Certains chercheurs tentent simplement d'estimer qualitativement la probabilité qu'un enregistrement indéterminé d'un ensemble de microdonnées soit reconnu, tandis que d'autres, dont les visées sont beaucoup plus ambitieuses, cherchent à déterminer la probabilité qu'un enregistrement précis le soit. Il s'agit évidemment de deux cas extrêmes. Le premier problème, quoique d'une certaine complexité, est assez facile à résoudre. Le second, plus difficile, pourrait ne pas avoir de solution.

Dans le présent article, nous donnons un aperçu des problèmes qu'on s'est efforcé de résoudre à Statistics Netherlands et dont la solution proposée a retenu notre attention. Nous nous limitons à décrire les problèmes et leur solution, en laissant de côté les aspects purement techniques. Le choix des problèmes et des solutions que nous examinons est en grande partie fonction des expériences vécues à Statistics Netherlands en ce qui concerne le CSD.

La suite de l'exposé se présente comme il suit. À la section 2, nous définissons les concepts de base. À la section 3, nous décrivons les préliminaires du CSD des microdonnées. À la section 4, nous exposons notre théorie de base sur le CSD des microdonnées. À la section 5, nous envisageons la situation idéale, c'est-à-dire celle où on peut calculer la probabilité de réidentification de chaque enregistrement. À la section 6, nous décrivons une situation un peu moins parfaite consistant à calculer la probabilité qu'un enregistrement indéterminé d'un ensemble de données soit reconnu. Puis, à la section 7, nous faisons face à la réalité – l'inexistence, à l'heure actuelle, d'un bon modèle du risque de divulgation qui oblige à se satisfaire d'arguments heuristiques. Enfin, à la section 8, nous présentons nos conclusions ainsi que des questions qui mériteraient d'être étudiées plus en détail.

2. CONCEPTS DE BASE

Dans la présente section, nous commençons par définir plusieurs concepts de base. Nous supposons, pour ce faire, que le bureau de la statistique souhaite diffuser un ensemble de microdonnées renfermant des enregistrements relatifs à un échantillon de la population. Chaque enregistrement contient des renseignements sur une entité particulière, qui

[1] A.G. deWaal et L.C.R.J. Willenborg, Statistics Netherlands, Division of Research and Development, Department of Statistical Methods, C.P. Box 4000, 2270 JM Voorburg, Pays-Bas (courrier électronique: TWAL@CBS.NL et LWLG@CBS.NL).

peut être une personne, un ménage ou une entreprise. Dans la suite de l'exposé, nous considérerons, en règle générale, que l'entité particulière est une personne, bien que cela ne soit pas essentiel.

Dans le domaine du CSD, les deux concepts les plus importants sont la réidentification et la divulgation. On dit qu'il y a réidentification quand un attaquant établit, avec un degré de confiance suffisant, un lien univoque entre un enregistrement de microdonnées et un individu cible. À l'exemple de Skinner (1992), nous distinguons deux types de divulgation. La divulgation par réidentification a lieu quand l'attaquant est capable de déduire la valeur que prend une variable délicate pour la personne cible après avoir reconnu cette dernière. La divulgation par prédiction (ou divulgation par attribut), quant à elle, a lieu si les microdonnées permettent à l'attaquant de prédire, avec un degré suffisant de confiance, la valeur d'une variable délicate pour une personne cible particulière. Dans le cas de la divulgation par prédiction, la réidentification n'est pas nécessaire. Jusqu'à présent, la plupart des travaux de recherche ont mis l'accent sur la divulgation par réidentification. Dans le présent article, sauf avis contraire, c'est à cette dernière que fait allusion le terme divulgation.

Définissons maintenant ce que nous entendons par variable identificatrice. Une variable est dite identificatrice si, seule ou combinée à d'autres variables, elle permet à un utilisateur de données de reconnaître certains répondants. À titre d'exemple, mentionnons le lieu de résidence, le sexe, la nationalité, l'âge, la profession et le niveau de scolarité. Un sous-ensemble de l'ensemble des variables identificatrices est celui des identificateurs directs (ou formels), dont le nom, l'adresse et le numéro d'identité sont des exemples. Les identificateurs directs doivent être supprimés d'un ensemble de microdonnées avant la diffusion de ce dernier, sinon la réidentification serait très facile. Dans la plupart des cas, seule la suppression de ces indicateurs est nécessaire. Une combinaison de variables identificatrices est appelée une clé. Les variables identificatrices qui, ensemble, constituent une clé sont aussi appelées variables clés. Une valeur clé est une combinaison de valeurs enregistrées pour les variables identificatrices qui, ensemble, constituent la clé.

En pratique, déterminer si une variable est ou non identificatrice est un problème qui ne peut être résolu qu'en faisant preuve de jugement. Il n'existe aucune liste limitative de variables identificatrices intrinsèques, ni, d'ailleurs, aucun ensemble non ambigu et bien défini de règles permettant de repérer de telles variables. La sélection d'un ensemble de variables identificatrices, donc, de clés, est généralement fondée sur des hypothèses subjectives au sujet de la population. Statistics Netherlands applique certains critères, comme la visibilité des catégories d'une variable, pour déterminer si celle-ci est identificatrice, mais ces critères ne permettent pas de résoudre catégoriquement la question dans tous les cas. Qu'une variable soit ou ne soit pas considérée identificatrice est essentiellement une question de jugement. Dans le reste de l'article, nous supposerons qu'un ensemble de clés a été déterminé.

Les variables délicates (ou confidentielles) sont la contrepartie des variables identificatrices. Une variable est dite délicate (ou confidentielle) si certaines de ses valeurs représentent des caractéristiques qu'un répondant ne souhaiterait pas qu'on dévoile. En théorie, Statistics Netherlands considère que toutes les variables sont délicates, mais, en pratique, juge certaines plus délicates que d'autres. Comme dans le cas des variables identificatrices, déterminer si une variable est ou non délicate est une question qu'on ne peut résoudre, en pratique, qu'en faisant preuve de jugement. Ainsi, on considère, en règle générale, que le comportement sexuel et les antécédents criminels sont des variables délicates. Par contre, dans d'autres cas, la décision dépend, par exemple, du contexte culturel. À titre d'exemple, Keller et Bethlehem (1992) mentionnent la variable de revenu, jugée délicate aux Pays-Bas mais non en Suède. En outre, certaines variables, comme l'appartenance ethnique, devraient être considérées à la fois identificatrices et délicates. Cependant, les auteurs qui traitent de ces questions émettent ordinairement l'hypothèse qu'il est possible de séparer les variables identificatrices et délicates pour former des ensembles disjoints. Dans la suite de l'exposé, nous supposerons qu'on a déterminé un ensemble de variables délicates qui est disjoint de l'ensemble de variables identificatrices.

En se servant de renseignements sur les variables identificatrices, un attaquant éventuel peut essayer de divulguer des renseignements sur des variables délicates. Il convient de noter que ce mode de divulgation n'est possible que si le lien entre les valeurs des variables identificatrices et des variables délicates n'a pas été perturbé par un bruit dans les données ou par une méthode telle que l'échange de données.

Pour conclure la présente section, définissons le contrôle statistique de la divulgation. Ce dernier a pour but de réduire à un niveau acceptable le risque que des renseignements confidentiels sur certaines personnes soient divulgués. La définition de l'acceptabilité est fonction de la politique du diffuseur de données. Afin de réduire le risque de divulgation, il serait utile de pouvoir estimer ce dernier, mais il ne s'agit pas d'une condition nécessaire (voir la section 7). Certains travaux de recherche ont été consacrés à la définition et à l'estimation du risque de divulgation.

3. PRÉLIMINAIRES DU CSD RELATIF AUX MICRODONNÉES

En tant que client d'un bureau de la statistique, l'utilisateur d'un ensemble de microdonnées doit être satisfait de la qualité de ce dernier. Ordinairement, il s'intéresse non pas à des enregistrements particuliers, mais aux résultats statistiques qu'il peut tirer de l'ensemble, comme les tableaux qu'il produit lui-même à partir de ce dernier.

Puisqu'un ensemble de microdonnées est destiné à l'analyse statistique, il n'est pas nécessaire que chaque enregistrement de l'ensemble soit correct. Le bureau de la statistique peut donc perturber les enregistrements, en ajoutant un bruit ou en échangeant certains éléments

d'enregistrement entre plusieurs de ceux-ci, par exemple, afin de diminuer le risque de réidentification. La perturbation des enregistrements diminue le risque, car, même si une réidentification a effectivement lieu, l'information divulguée n'est pas nécessairement correcte. En tout cas, l'attaquant ne peut être sûr que l'information divulguée est correcte. Le bureau de la statistique doit garantir «uniquement» que la qualité statistique des tableaux que l'utilisateur souhaite examiner, pour s'en tenir à l'exemple choisi, est suffisante. Néanmoins, en pratique, cela peut s'avérer difficile.

Bien que les méthodes de perturbation des données soient parfois utiles, Statistics Netherlands ne les utilise pas à l'heure actuelle. Pour protéger les ensembles de microdonnées, nous ne recourons qu'à la suppression locale et au recodage global. Dans le cas de la suppression locale, certaines valeurs des variables de certains enregistrements sont déclarées «manquantes», c.-à-d. supprimées de l'ensemble de microdonnées. Dans celui du recodage global, certaines variables sont catégorisées plus grossièrement. Nous essayons d'abord de protéger l'ensemble de microdonnées par recodage global. Cependant, si le recours à cette seule méthode entraîne une perte d'information considérable, nous effectuons également des suppressions locales afin d'éviter que la perte soit excessive. Nous tenons néanmoins à préciser que les suppressions locales sont effectuées parcimonieusement.

L'avantage de la suppression locale et du recodage global tient à ce que ces méthodes préservent l'intégrité des données. La suppression locale a toutefois pour inconvénient d'introduire un biais, car les valeurs extrêmes sont supprimées localement. Cependant, si les suppressions locales sont effectuées parcimonieusement, le biais est faible.

Du point de vue du CSD, l'utilisateur des données devrait aussi être considéré comme un attaquant éventuel. Par conséquent, il est utile d'imaginer de quelles façons la divulgation peut avoir lieu. Un attaquant essaye généralement d'apparier les enregistrements de l'ensemble de microdonnées à ceux d'un fichier d'identification ou à des personnes appartenant à son cercle de connaissances. Un fichier d'identification est un fichier dont les enregistrements contiennent des valeurs pour les identificateurs directs ainsi que pour d'autres identificateurs de l'ensemble de microdonnées. Les seconds identificateurs peuvent servir à apparier les enregistrements de l'ensemble de microdonnées diffusé à ceux du fichier d'identification. Après l'appariement, les identificateurs directs qui figurent dans le fichier d'identification peuvent être utilisés pour déterminer à quelle personne correspond l'enregistrement apparié, puis, les variables délicates de l'ensemble de microdonnées diffusé peuvent servir à divulguer l'information au sujet de la personne en question. Le cercle de connaissances de l'attaquant est l'ensemble des membres de la population pour lesquels il connaît les valeurs d'une clé particulière de l'ensemble de microdonnées. Donc, un cercle de connaissances pourrait effectivement jouer le rôle de fichier d'identification, et vice versa. Dans la suite de l'article, nous utiliserons indifféremment les expressions «fichier d'identification» et «cercle de connaissances».

Pour que l'attaquant puisse reconnaître la personne à laquelle correspond un enregistrement particulier, il faut que les conditions ci-après soient satisfaites:

C_1. La personne est unique en ce qui a trait à une valeur clé particulière K.
C_2. La personne est incluse dans un fichier d'identification ou appartient à un cercle de connaissances de l'attaquant.
C_3. La personne est un élément de l'échantillon.
C_4. L'attaquant sait que, en ce qui concerne la clé K, l'enregistrement est unique au sein de la population.
C_5. L'attaquant découvre l'enregistrement dans l'ensemble de microdonnées.
C_6. L'attaquant reconnaît l'enregistrement de la personne.

Si une des conditions C_1 à C_6 n'est pas remplie, la réidentification ne peut être accomplie de façon absolument certaine. Si la condition C_1 ou C_4 n'est pas vérifiée, un appariement peut être effectué, mais l'attaquant ne peut être certain qu'il aboutira à une réidentification correcte.

Il est manifeste, quand on examine la liste des conditions C_1 à C_6, qu'un «bon» modèle du risque de réidentification doit englober certaines caractéristiques non seulement des données, mais aussi de l'utilisateur. Si un ensemble de microdonnées d'origine hollandaise est utilisé, disons, en Chine, par quelqu'un qui, fondamentalement, ne connaît pas la population hollandaise, le risque de réidentification est négligeable. Afin de reconnaître une personne sur laquelle des données figurent dans un ensemble de microdonnées, il est nécessaire de bien connaître la population. L'effort qui doit être déployé pour acquérir cette connaissance est proportionnelle à la sécurité de l'ensemble de microdonnées.

4. UNE CONCEPTION DU CONTRÔLE STATISTIQUE DE LA DIVULGATION

L'attention d'un attaquant éventuel est vraisemblablement attirée par des combinaisons rares de variables identificatrices dans l'échantillon ou dans la population. Les combinaisons relativement fréquentes sont moins susceptibles de susciter sa curiosité. S'il tente délibérément d'apparier des enregistrements, il s'efforcera probablement de le faire pour des valeurs clés peu fréquentes. Il se peut aussi qu'il n'essaye pas délibérément d'apparier des enregistrements, mais que, sachant qu'une de ses connaissances possède une valeur clé rare, la découverte d'un enregistrement contenant cette valeur clé le pousse à essayer de déterminer s'il se rapporte à la personne en question. En outre, la probabilité qu'un appariement soit correct est d'autant plus élevée que le nombre de personnes pour lesquelles on enregistre la valeur clé à apparier est faible. Enfin, il est également fort probable que, parmi les personnes associées à une valeur clé rare, beaucoup deviennent uniques quand on ajoute une autre variable à la clé. Les enregistrements qui contiennent de telles combinaisons rares de variables identificatrices sont donc plus susceptibles d'être reconnus.

Plus précisément, les valeurs clés observées une seule fois dans la population, autrement dit uniques, peuvent mener à la réidentification. Auparavant on mettait presque exclusivement l'accent sur l'unicité. Toutefois, il convient de noter que cette dernière n'est une condition ni suffisante ni nécessaire de la réidentification. En effet, un membre de la population unique en ce qui a trait à certaines variables clés pourraient ne jamais être reconnu si personne ne se rend compte de cette unicité. Par contre, s'il n'est pas unique, mais qu'une seule autre personne est associée à la même clé, il peut, en principe, être reconnu par cette dernière. D'autre part, supposons qu'une personne n'est pas unique en regard d'une variable clé, mais qu'elle appartient à un petit groupe. Supposons, en outre, que l'attaquant possède à son sujet un renseignement qui, n'étant pas considéré identificateur par le bureau de la statistique, figure dans l'ensemble de microdonnées diffusé. Dans ces conditions, il se pourrait que la personne en question soit la seule associée à la clé combinée à la nouvelle information. Donc, il est possible de reconnaître une personne même si elle n'est pas la seule qui correspond aux valeurs clés des variables identificatrices au sein d'une population. Enfin, il peut y avoir divulgation par prédiction. Autrement dit, si une personne n'est pas unique dans la population, mais qu'elle appartient à un groupe dont les membres obtiennent (pratiquement) la même valeur pour une variable délicate particulière, l'information confidentielle au sujet de cette personne peut être divulguée sans qu'il y ait effectivement réidentification. La divulgation par prédiction n'est pas examinée davantage dans le présent article. Nous conseillons au lecteur qui souhaite obtenir plus de renseignements à ce sujet de consulter Skinner (1992), US Department of Commerce (1978), Duncan et Lambert (1986), et Cox (1986).

Le CSD devrait être axé sur les valeurs clés rares au sein de la population. La probabilité que soient divulgués des renseignements sur un répondant précis dont les données sont incluses dans un ensemble de microdonnées devraient refléter la «rareté» de la valeur clé qui figure dans l'enregistrement de ce répondant. La probabilité que soient divulgués les renseignements fournis pour tout répondant devrait refléter la «rareté globale» des enregistrements dans l'ensemble de données. Si un ensemble de microdonnées contient un grand nombre d'enregistrements dont la valeur clé est rare, la probabilité de divulgation associée à cet ensemble est, en principe, élevée. Dans les sections suivantes, nous examinons certains efforts en vue d'intégrer ces concepts à un modèle mathématique.

5. RISQUE DE RÉIDENTIFICATION PAR ENREGISTREMENT

Dans un monde idéal (en ce qui concerne le CSD), un diffuseur de microdonnées devrait être capable de déterminer le risque de réidentification que pose un enregistrement, c.-à-d. la probabilité que le répondant associé à cet enregistrement puisse être reconnu. Le calcul d'un tel risque par enregistrement nous permettrait d'adopter la stratégie que voici. En premier lieu, classer les enregistrements en fonction du risque qu'ils soient réidentifiés en regard d'une clé unique. En second lieu, déterminer le risque maximum que le bureau de la statistique est prêt à accepter. Enfin, modifier tous les enregistrements pour lesquels le risque de réidentification en regard de la clé choisie est trop élevé. Répéter la procédure pour chaque clé, s'il en existe plus d'une.

Malheureusement, à l'heure actuelle, ce monde idéal nous échappe. Cependant, Paass et Wauschkuhn (1985), et Fuller (1993) nous en rapprochent de quelque pas. Paass et Wauschkuhn (1985) supposent qu'un attaquant éventuel a à sa disposition un fichier de microdonnées diffusé par un bureau de la statistique, ainsi qu'un fichier d'identification. Bon nombre des données de ces deux fichiers peuvent être incompatibles. Les incompatibilités peuvent être le résultat d'erreurs de codage, de définitions différentes des catégories ou d'un «bruit» dans les données. Paass et Wauschkuhn émettent une hypothèse quant à la loi de distribution de ces données incompatibles et au mode de divulgation, et établissent un modèle complexe pour estimer la probabilité qu'un enregistrement particulier du fichier de microdonnées soit reconnu. Ils présument que l'attaquant connaît la loi de distribution des erreurs qui ont causé l'incompatibilité des données. Ils supposent en outre qu'il ne connaît pas la variance des erreurs. Un attaquant éventuel doit estimer cette variance, en se fondant sur sa connaissance (tenue pour acquise) de la méthode de production des données statistiques. Le modèle de Paass et Wauschkuhn repose essentiellement sur l'analyse discriminante et sur l'analyse typologique.

Paass et Wauschkuhn distinguent six scénarios qui correspondent chacun à un type particulier d'attaquant. Le nombre d'enregistrements dans le fichier d'identification et le contenu informatique de ce fichier dépendent du scénario choisi. Un exemple de scénario est celui du journaliste qui sélectionne des enregistrements où figurent des combinaisons d'attributs extrêmes pour réidentifier des répondants, dans le but de prouver que le bureau de la statistique ne réussit pas à sauvegarder la vie privée de ses répondants.

Paass et Wauschkuhn appliquent leur méthode en vue d'apparier des enregistrements du fichier d'identification à des enregistrements du fichier de microdonnées. Quand la probabilité qu'un enregistrement particulier du fichier d'identification correspond à un enregistrement particulier du fichier de microdonnées est suffisamment forte, les deux enregistrements sont appariés. Cette probabilité représente la probabilité de réidentification par enregistrement, sous réserve d'un scénario de divulgation particulier.

Müller, Blien, Knoche, Wirth et coll. (1991) et Blien, Wirth et Müller (1992) ont appliqué la méthode préconisée par Paass et Wauschkuhn (1985) à des données réelles. Cette méthode ne s'est pas avérée meilleure que le simple appariement, en vertu duquel un enregistrement est considéré réidentifié par un attaquant si ce dernier réussi à déceler dans le fichier de microdonnées un ensemble de valeurs uniques identique à un ensemble de valeurs figurant dans le fichier d'identification. Apparemment, le nombre d'enregistrements appariés correctement en appliquant la méthode de Paass et Wauschkuhn ne concordait pas avec la probabilité de réidentification par enregistrement.

En ce qui a trait aux méthodes de camouflage, c'est-à-dire les méthodes qui visent à limiter la divulgation en ajoutant un bruit aux microdonnées, Fuller (1993) propose une expression qui permet de calculer la probabilité qu'un enregistrement particulier de l'ensemble de microdonnées diffusé soit le même qu'un enregistrement cible d'un fichier d'identification, autrement dit la probabilité de réidentification par enregistrement. Il émet plusieurs hypothèses pour arriver à cette équation. En premier lieu, il suppose que la loi de distribution des données, du bruit et des erreurs est normale. En outre, il présume que l'attaquant connaît les matrices de covariances du bruit et des erreurs qui affectent les données. Enfin, il suppose que les données ont été obtenues par échantillonnage aléatoire simple. Ces hypothèses permettent à Fuller (1993) d'obtenir l'équation utilisée pour calculer la probabilité de réidentification grâce à des considérations théoriques au sujet des probabilités. Malheureusement, sa méthode n'a pas encore été éprouvée au moyen de données réelles. Donc, il est difficile de porter un jugement sur son applicabilité. Le lecteur trouvera dans Willenborg (1993) un commentaire de la méthode préconisée par Fuller.

Paass et Wauschkuhn (1985) et Fuller (1993) s'intéressent surtout aux effets qu'a sur le risque de divulgation le bruit qui est ajouté (involontairement et volontairement, respectivement) aux données. Un des points faibles de leurs méthodes respectives tient à l'hypothèse, implicite, que le nombre de dimensions de la clé est élevé. En effet, supposer que le nombre de dimensions de la clé est élevé sous-entend que (pratiquement) tous les membres de la population sont uniques. La probabilité qu'une combinaison ou une valeur clé se manifeste plus d'une fois au sein de la population est pratiquement nulle, situation qui rend le calcul de la probabilité de réidentification par enregistrement beaucoup plus facile. En revanche, si le nombre de dimensions des clés est faible, il n'est pas improbable que certaines valeurs clés se manifestent fréquemment dans la population. Calculer la probabilité de réidentification par enregistrement dans ces conditions est beaucoup plus difficile que quand le nombre de dimensions des clés est élevé et que la probabilité qu'il existe des paires statistiques au sein de la population est virtuellement nulle.

Les modèles du risque de réidentification par enregistrement proposés jusqu'à présent ne semblant pas satisfaisants, à la section 6, nous envisageons des modèles moins ambitieux, qui permettent de calculer le risque de réidentification par fichier.

6. RISQUE DE RÉIDENTIFICATION PAR FICHIER

Dans un monde un peu moins parfait, un diffuseur de microdonnées ne serait pas capable de déterminer le risque de réidentification pour chaque enregistrement, mais pourrait calculer le risque qu'un enregistrement indéterminé de l'ensemble de microdonnées soit reconnu. Dans ces conditions, le bureau de la statistique doit décider du risque maximal qu'il est disposé à courir quand il diffuse un ensemble de microdonnées. Ce dernier peut être diffusé si le risque réel est inférieur au risque maximal, mais doit être modifié si le risque dépasse la limite. Toutefois, déterminer quels enregistrements doivent être modifiés est un exercice qui demeure problématique.

Mokken, Pannekoek et Willenborg (1989) et Mokken, Kooiman, Pannekoek et Willenborg (1992) ont proposé un modèle de base pour calculer la probabilité qu'un enregistrement choisi au hasard dans un ensemble de microdonnées soit reconnu. Mokken et coll. (1989) envisagent uniquement le scénario comprenant un seul chercheur, une population non stratifiée et une clé unique. Ce modèle a été élargi afin d'inclure les cas de sous-populations, de chercheurs multiples et de clés multiples (consulter Willenborg, 1990a; Willenborg, 1990b; Mokken et coll., 1992). Le modèle de Mokken et coll. (1992) tient compte de trois probabilités. La première, f, est égale à la fraction de sondage. Autrement dit, f représente la probabilité qu'une personne choisie au hasard dans la population soit sélectionnée dans l'échantillon. La deuxième, f_a, représente la probabilité qu'un chercheur particulier ayant accès aux microdonnées connaisse les valeurs qui correspondent à une personne choisie au hasard au sein de la population pour une clé particulière. La troisième, f_u, est la probabilité qu'une personne choisie au hasard au sein de la population soit unique dans cette population en regard d'une clé particulière. En combinant ces trois probabilités, f, f_a et f_u, on peut calculer la probabilité qu'un des enregistrements d'un ensemble de microdonnées soit réidentifié.

Plusieurs variables sont mesurées pour chaque élément de l'échantillon. Les valeurs obtenues pour ces diverses mesures sont regroupées en enregistrements correspondant chacun à un élément particulier de l'échantillon. Mekken et ses collaborateurs supposent que les variables incluses dans la clé sont soit des variables nominales, soit des variables dont les mesures se classent dans un nombre fini de catégories. Regroupés, les enregistrements constituent un ensemble de données S qui est mis à la disposition d'un chercheur R. Rappelons que, chaque fois que nous utilisons le terme divulgation, nous entendons en fait divulgation par réidentification. Le modèle de Mokken et coll. (1989, 1992) ne tient pas compte de la divulgation par prédiction.

Aux termes du modèle de Paass et Wauschkuhn (1985), les probabilités f_a et f_u réunies reflètent l'*Informationsgehalt der Überschneidungsmerkmale*, c.-à-d. le contenu informatique des valeurs appariées. Les divers scénarios qu'ils envisagent se distinguent en ce qui a trait à f_a et à f_u. Plus précisément, la valeur de f_u dépend du nombre de variables qu'un attaquant a à sa disposition pour réidentifier un enregistrement ainsi que du contenu informatique, c.-à-d. la catégorisation, de ces variables. La valeur du paramètre f_a est fonction du nombre d'enregistrements que contient le fichier d'information.

En ce qui concerne le chercheur R et la clé K, il existe un cercle de connaissances A. Manifestement, A et sa dimension $|A|$ dépendent du chercheur R et de la clé K, ainsi que des variables enregistrées et codées dans l'ensemble de données.

Les auteurs présument que, si les conditions C_1, C_2 et C_3 de la liste de conditions de réidentification énumérées à la section 3 sont satisfaites, les conditions C_4, C_5 et C_6 sont remplies également. La condition C_4 est relativement astreignante, mais, par souci de commodité, peut-être introduite à titre d'hypothèse quand on formule un modèle du risque de divulgation. Il convient de noter qu'un tel modèle représente le pire scénario, en ce sens que les défauts de perception ou de mémoire, ou d'autres sources d'ignorance, de confusion et d'incertitude liées à un divulgateur éventuel sont exclues. Quand on prend simultanément pour hypothèse les conditions C_5 et C_6, on signifie que l'apparition de toute connaissance unique E de R dans l'ensemble de données S équivaut à la réidentification par R. Les auteurs supposent que la réidentification d'un enregistrement sous-entend la divulgation de renseignements confidentiels. Donc, ils peuvent considérer que la réidentification est équivalente à la divulgation. Implicitement, ils supposent que le lien entre les variables identificatrices et les variables délicates n'a pas été perturbé par une méthode telle que l'échange de données.

Qui plus est, contrairement à Paass et Wauschkuhn (1985), et à Fuller (1993), ils supposent que le chercheur R dispose de données tant identificatrices que confidentielles exemptes d'erreur ou de bruit. Manifestement, cette hypothèse est irréaliste pour la plupart des ensembles de microdonnées.

Le risque de divulgation D_R lié à un ensemble particulier de microdonnées S en ce qui a trait à un chercheur particulier R et à une clé particulière K se définit comme la probabilité que le chercheur divulgue au moins un des enregistrements de S en se basant sur K. Pour pouvoir appliquer un critère fondé sur le risque de divulgation, il est nécessaire de déterminer cette grandeur pour un ensemble de données particulier. En se basant sur un ensemble d'hypothèses, il est possible d'établir une équation permettant de calculer cette grandeur.

En plus des hypothèses C_1 à C_6 mentionnées précédemment, le modèle de Mokken et coll. se fonde sur les hypothèses suivantes:

A_1. Le cercle de connaissances A peut être considéré comme un échantillon aléatoire de la population.

A_2. L'ensemble de données S est un échantillon aléatoire de la population.

L'hypothèse A_1 sous-entend que l'expression de la probabilité qu'un élément choisi au hasard au sein de la population soit une connaissance de R est $f_a = |A|/N$, où N représente la taille de la population. Par conséquent, le nombre prévu d'éléments uniques dans A, $|U_a|$, est égal à $f_a |U| = |A| f_u$, où U représente l'ensemble de personnes uniques au sein de la population et $|U|$, sa taille. De toute évidence, l'hypothèse A_2 implique que la probabilité qu'un élément unique E particulier soit choisi dans l'échantillon est égal à f. Ces hypothèses permettent d'obtenir une expression très simple du risque D_R en fonction de f, f_a et f_u, à savoir

$$D_R = 1 - \exp(-Nff_af_u). \qquad (1)$$

Deux des paramètres du modèle de Mokken et coll. (1989, 1992), f_a et f_u, sont inconnus. On peut «estimer subjectivement» le paramètre f_a, c.-à-d. obtenir sa valeur par divination éclairée, en postulant divers scénarios qu'un attaquant est susceptible de suivre. Plusieurs de ces scénarios ont été décrits par Paass et Wauschkuhn (1985) et par Paass (1988). Néanmoins, il paraît difficile d'évaluer f_a. En ce qui concerne l'estimation de l'autre paramètre, f_u, plusieurs modèles ont été proposés. Ceux qui ont été préconisés pour estimer le nombre d'éléments uniques dans la population, donc f_u, incluent la distribution gamma de Poisson (Bethlehem, Keller et Pannekoek 1989; Mokken et coll., 1989; Willenborg, Mokken et Pannekoek 1990; De Jonge, 1990), le modèle de superpopulation à distribution binomiale négative (Skinner et coll., 1990), la distribution normale logarithmique de Poisson (Skinner et Holmes, 1992; Hoogland, 1994), des modèles fondés sur les classes d'équivalence (Greenberg et Zayatz, 1992) et des modèles fondés sur des fonctions gamma binomiales négatives modifiées (Crescenzi, 1992; Coccia, 1992). Comme nous l'avons mentionné à la section 4, non seulement le nombre de populations uniques est important, mais aussi celui de cellules comptant deux, trois, *etc.* personnes. Le modèle gamma de Poisson, le modèle normal logarithmique de Poisson et le modèle de superpopulation à distribution binomiale négative peuvent également servir à estimer le nombre de cellules comptant deux, trois, *etc.* personnes. Il semble que les autres modèles susmentionnés puissent être élargis afin d'estimer ces nombres. Un inconvénient important tient au fait que, dans de nombreux cas, les résultats ne sont pas très fiables.

Le modèle de Mokken et coll. (1989, 1992) montre clairement que le bureau de la statistique qui diffuse les données peut prendre des mesures pour modifier le risque de réidentification, et ce, essentiellement de deux façons. En premier lieu, il est possible de réduire la taille de l'ensemble de données, c.-à-d. la fraction de sondage f. Une diminution de f implique une diminution du risque. Cependant, il n'est généralement pas souhaitable de réduire f, car la réduction doit être considérable pour que la mesure soit efficace. Donc, le cas échéant, seule une faible proportion des données disponibles peuvent être diffusées. En second lieu, le bureau de la statistique peut modifier le risque de réidentification en réduisant le nombre d'éléments uniques dans la population, c.-à-d. en réduisant f_u. La fraction f_u dépend de l'information que fournissent les variables clés. La population compte d'autant moins d'éléments uniques que les variables clés fournissent moins de renseignements. Autrement dit, on peut réduire f_u en regroupant les catégories (recodage global) et en remplaçant certaines valeurs par l'indication «manquante» (suppression locale). Le regroupement des catégories est une mesure globale, car elle touche généralement de nombreux enregistrements; le remplacement de certaines valeurs par une lacune est une mesure locale, car elle n'en concerne que quelques-uns. Ordinairement, la perte d'information est nettement moins importante quand on réduit f_u que quand on réduit f. Par conséquent, le bureau de la statistique décide généralement de contrôler le risque de réidentification en réduisant f_u

plutôt que f. La troisième méthode de contrôle du risque de réidentification, c.-à-d. la réduction de f_a, n'est pas appliquée en pratique, car il est difficile de modéliser f_a.

Si le modèle de Mokken et coll. (1989, 1992) donne une idée des mesures qui peuvent être prises pour réduire le risque de divulgation, il peut difficilement servir de fondement à la protection des ensembles de microdonnées, car les deux paramètres du modèle, f_u et f_a, sont souvent difficiles à évaluer. Habituellement, on ne dispose pas de suffisamment de données pour les estimer avec précision. Nous concluons que même un modèle visant à calculer le risque de réidentification pour un ensemble complet de microdonnées est difficile à appliquer en pratique. À la section 7, nous faisons face à la réalité, à savoir le fait que nous ne disposons d'aucun modèle satisfaisant pour calculer le risque de réidentification par enregistrement ou pour un ensemble complet de microdonnées.

7. CONTRÔLE INTUITIF DU RISQUE DE RÉIDENTIFICATION

En réalité, nous sommes obligés de fonder le CSD sur des arguments heuristiques plutôt que sur des concepts théoriques solides. Les règles de CSD mentionnées dans la présente section permettent toutes de réduire le risque de réidentification. Néanmoins, il n'est pas possible d'évaluer la réduction. À Statistics Netherlands, les règles de CSD relatives aux microdonnées s'appuient sur des tests visant à déterminer si la fréquence de certaines valeurs clés est suffisamment élevée au sein de la population. Les problèmes qui se posent ici consistent à déterminer quelles clés il faut examiner et comment il faut estimer le nombre de membres de la population qui produisent une valeur clé particulière, à rendre opérationnelle la phrase «fréquence suffisamment élevée» en déterminant, par exemple, une ou plusieurs valeurs seuils, et enfin, à décider quelles sont les mesures de CSD appropriées.

Statistics Netherlands distingue deux types d'ensembles de microdonnées. Le premier est un fichier dit d'usage public. N'importe qui peut obtenir un tel fichier. Les clés qu'il convient d'examiner dans le cas d'un fichier public correspondent à toutes les combinaisons possibles de variables identificatrices prises deux à deux. Le nombre de variables identificatrices est limité et certaines, comme le lieu de résidence, ne sont pas incluses dans un fichier public. Il faut en outre examiner les poids d'échantillonnage avant de les inclure dans le fichier, car très souvent, ils peuvent fournir des renseignements supplémentaires (consulter De Waal et Willenborg, 1995a). Par exemple, une sous-population particulière qu'on a suréchantillonnée risque d'être reconnue à cause du faible poids attribué à ses membres dans l'échantillon. Les poids ne peuvent être publiés que s'ils ne fournissent aucun renseignement supplémentaire susceptible de faciliter la divulgation. Si on estime que les poids ne peuvent être publiés, il convient de prendre des mesures de CSD, telles que le sous-échantillonnage des unités dont le poids est faible, afin d'obtenir un sous-échantillon où toutes les unités ont à peu près le même poids. Puisque les poids sont presque égaux, émettre l'hypothèse qu'ils le sont parfaitement n'entraîne qu'une faible erreur. Le deuxième type d'ensemble de microdonnées est l'ensemble de microdonnées dit «de recherche». Un ensemble de microdonnées de recherche ne peut être obtenu que par des bureaux de recherche (statistique) dont la réputation est bien établie. Le contenu informatique d'un tel ensemble est beaucoup plus important que celui d'un fichier public. Le nombre de variables identificatrices n'y est pas limité et une variable identificatrice telle que le lieu de résidence peut y être incluse. En raison du contenu informatique détaillé de ces fichiers, les chercheurs doivent signer une déclaration précisant qu'ils s'engagent à protéger tout renseignement susceptible d'être divulgué sur des répondants particuliers. Les clés qu'il convient d'examiner pour un ensemble de microdonnées de recherche correspondent à des combinaisons triples de variables décrivant une région, de variables décrivant le sexe, le groupe ethnique ou la nationalité d'un répondant, et d'une variable identificatrice ordinaire.

Les règles qu'applique Statistics Netherlands en matière de CSD se fondent sur le principe selon lequel la diffusion d'une valeur clé, c'est-à-dire une combinaison de valeurs des variables identificatrices qui, ensemble, constituent la clé, est jugée sans danger si la fréquence à laquelle cette valeur clé se manifeste dans la population dépasse un certain seuil d_0. Cette valeur d_0 a été choisie après un examen minutieux et extensif, consistant à envisager un grand nombre de valeurs distinctes, et à comparer les enregistrements qui doivent être modifiés pour chaque valeur de d_0. La valeur qui produit l'ensemble «le plus probable» d'enregistrements qui doivent être modifiés a été choisie comme valeur de d_0. La sélection de l'ensemble «le plus probable» d'enregistrements à modifier est une question de jugement.

Quand nous appliquons les règles susmentionnées, nous nous heurtons généralement au problème de ne pas connaître le nombre de fois qu'une valeur clé survient dans la population. Nous ne disposons que de l'échantillon. La fréquence d'une valeur clé au sein de la population doit être estimée à partir de cet échantillon. Dans le cas de grandes régions, il est possible de se servir d'un estimateur d'intervalle pour vérifier si une valeur clé se manifeste suffisamment souvent dans une région donnée. Cet estimateur d'intervalle est défini en supposant que le nombre de fois qu'une valeur clé survient dans la population obéit à la loi de Poisson (consulter Pannekoek, 1995). Cependant, dans le cas des régions relativement petites, le nombre de répondants étant faible, la variance de l'estimateur est élevée, situation qui, à son tour, oblige à modifier un grand nombre d'enregistrements. Par conséquent, pour estimer le nombre de fois qu'une valeur clé se manifeste dans une petite région, nous proposons d'utiliser un estimateur ponctuel. Examinons maintenant certains de ces estimateurs.

L'estimateur ponctuel direct constitue un moyen simple d'estimer le nombre de fois qu'une valeur clé particulière est observée dans une région. On estime la fraction d'une valeur clé dans une région i en divisant la fréquence dans

l'échantillon de cette valeur clé dans la région i par le nombre de répondants dans la région i. Puis, on estime sa fréquence dans la population en multipliant la fraction estimée par le nombre d'habitants dans la région i. Quand le nombre de répondants dans la région i est faible, ce qui est souvent le cas, l'estimateur direct n'est pas fiable. Un autre estimateur ponctuel se fonde sur l'hypothèse selon laquelle les personnes qui produisent une valeur clé particulière sont réparties uniformément dans la population. Dans ce cas, on peut estimer la fraction de la valeur clé dans la région i en déterminant la fraction dans l'échantillon complet. L'avantage que présente cet estimateur dit «synthétique» tient au fait que sa variance est beaucoup plus faible que celle de l'estimateur direct. Malheureusement, l'hypothèse quant à l'homogénéité de la population n'étant généralement pas satisfaite, l'estimateur est biaisé. Cependant, on peut créer un estimateur combiné, dont tant la variance que le biais sont acceptables, par combinaison convexe de l'estimateur direct et de l'estimateur synthétique. Pannekoek et de Waal (1995) ont testé un estimateur combiné de ce type.

Le camouflage des valeurs extrêmes des variables continues (délicates) est un autre problème pratique qui mérite d'être examiné. Ces valeurs extrêmes peuvent mener à la réidentification, car elles sont rares au sein de la population. Pour le moment, à Statistics Netherlands, nous nous servons d'un estimateur d'intervalle pour vérifier si la population contient un nombre suffisant de personnes pour lesquelles on observe une valeur «comparable» de la variable continue (consulter Pannekoek, 1992), mais il est possible que nous appliquions un estimateur ponctuel à l'avenir. Si le nombre est suffisant, la valeur extrême peut être publiée. Sinon, elle doit être supprimée. Afin de mettre la méthode en pratique, il reste à préciser ce qu'on entend par «suffisant» et par «comparable».

D'importants problèmes d'ordre pratique surviennent lors de la détermination des mesures à prendre pour protéger un ensemble de microdonnées qui semble poser un risque. Le cas échéant, il convient de modifier l'ensemble de données original, de façon à réduire au minimum la perte d'information causée par les mesures de CSD tout en produisant un ensemble de données jugées sans risque. Un modèle permettant de déterminer les suppressions locales optimales est présenté par De Waal et Willenborg (1994a), et De Waal et Willenborg (1995b). La détermination du recodage global optimal est, par contre, beaucoup plus difficile. Comparer la perte d'information due au recodage global à celle qu'entraîne les suppressions locales pose déjà, en soi, un problème. De Waal et Willenborg (1995c) le résolvent en s'appuyant sur le concept d'entropie.

Statistics Netherlands est en train de mettre au point un progiciel d'usage général pour le CSD relatif aux microdonnées (consulter De Jong, 1992; De Waal et Willenborg, 1994b; Van Gelderen, 1995; Pieters et De Waal, 1995; De Waal et Pieters, 1995). Le progiciel, baptisé ARGUS, devrait permettre au bureau de la statistique d'analyser les données et d'exécuter les mesures de protection pertinentes. Il comprendra deux éléments distincts: μ-ARGUS, pour le CSD des microdonnées et τ-ARGUS, pour le CSD des données tabulaires. Le progiciel est conçu de façon à pouvoir préciser diverses règles de contrôle de la divulgation. Donc, ARGUS pourra être utilisé par d'autres bureaux de la statistique. Qui plus est, il sera possible de modifier assez facilement les règles incluses dans le progiciel.

8. CONCLUSIONS

Une des premières conclusions qui se dégagent du présent article est la suivante: malgré le nombre d'études déjà accomplies, il y a encore lieu d'effectuer de nombreux travaux de recherche dans le domaine du CSD. La théorie du CSD appliquée aux microdonnées présente plusieurs lacunes. Voici quelques-uns des problèmes techniques qui restent à résoudre. En ce qui concerne la diffusion de données sur des secteurs restreints, nous devons définir un estimateur acceptable de la fréquence à laquelle une valeur clé est observée dans ces régions. La construction d'un tel estimateur est difficile, mais les résultats préliminaires obtenus à Statistics Netherlands paraissent encourageants. En pratique, déterminer quels sont les recodages globaux et les suppressions locales appropriés constituent un problème important. La détermination du nombre d'éléments uniques ou, de façon plus générale, du nombre de fréquences rares dans la population en est un autre. Certains modèles proposés à la section 6 restent acceptables, mais pourraient sans doute être améliorés. Une autre méthode consiste à déterminer quels éléments de l'échantillon sont uniques dans la population. Verboon (1994) et Verboon et Willenborg (1995) ont examiné cette méthode. L'extension du modèle proposé par Mokken et coll. (1989, 1992) pour estimer le risque de réidentification d'un fichier est un autre problème non résolu. L'extension du modèle devrait tenir compte de ce que des erreurs de mesure ont été commises et que l'unicité de la population n'est pas toujours une condition nécessaire pour qu'ait lieu la divulgation d'information. Enfin, un modèle permettant d'estimer le risque de réidentification par enregistrement serait le bienvenu. En fait, un tel modèle fournirait un critère valable pour juger de la sécurité d'un ensemble de microdonnées. Ce critère servirait de guide pour produire des ensembles de microdonnées sûrs, grâce à l'application de mesures de CSD telles que le recodage global et la suppression locale.

Outre les problèmes techniques, des problèmes de politique restent à résoudre. Compte tenu de la politique qu'il se propose de suivre, le bureau de la statistique doit préciser quelles combinaisons de variables devrait être examinées et choisir des valeurs limites appropriées.

De nouveaux logiciels, plus performants, doivent être mis au point pour venir à bout plus facilement des calculs très longs. En ce qui concerne les microdonnées, il convient de mettre au point des logiciels qui, en appliquant une règle de divulgation particulière, indiqueront quels enregistrements et quelles variables doivent être modifiés, et de quelle manière. Au moment où nous rédigeons le présent

article, un projet international sur le CSD, auquel participeront l'Eindhoven University of Technology, l'University of Manchester, l'University of Leeds, l'Office of Population Censuses and Surveys (OPCS), l'Istituto Nazionale di Statistica (ISTAT), le Consortio Padova Ricerche (CPR) et Statistics Netherlands, est sur le point de débuter. Un des objectifs principaux du projet consiste à mettre au point un logiciel applicable au CSD des microdonnées (μ-ARGUS) et des données tabulaires (τ-ARGUS).

Enfin, des problèmes d'un ordre très pratique doivent être résolus, comme la détermination d'un ensemble de règles de sélection des variables identificatrices. Cet ensemble de règles constituerait un atout précieux. En effet, sans ces règles, les variables identificatrices sont choisies de façon subjective. L'élaboration d'un tel ensemble de règles est un autre objectif du projet de CSD susmentionné.

BIBLIOGRAPHIE

BETHLEHEM, J.A., KELLER, W.J., et PANNEKOEK, J. (1989). Disclosure control of microdata. *Journal of the American Statistical Association*, 85, 38-45.

BLIEN, U., WIRTH, H., et MÜLLER, M. (1992). Disclosure risk for microdata stemming from official statistics. *Statistica Neerlandica*, 46, 69-82.

COCCIA, G. (1992). Disclosure risk in Italian current population surveys. International Seminar on Statistical Confidentiality, Dublin.

COX, L.H. (1986). Commentaires sur Duncan et Lambert (1986). 19-21.

CRESCENZI, F. (1992). Estimating population uniques; methodological proposals and applications on Italian census data. International Seminar on Statistical Confidentiality, Dublin.

De JONG, W.A.M. (1992). ARGUS: An integrated system for data protection. International Seminar on Statistical Confidentiality, Dublin.

De JONGE, G. (1990). The estimation of population unicity from microdata files (en néerlandais), note interne, Statistics Netherlands, Voorburg.

De WAAL, A.G., et PIETERS, A.J. (1995). ARGUS user's guide. Rapport, Statistics Netherlands, Voorburg.

De WAAL, A.G., et WILLENBORG, L.C.R.J. (1994a). Minimizing the number of local suppressions in a microdata set. Rapport, Statistics Netherlands, Voorburg.

De WAAL, A.G., et WILLENBORG, L.C.R.J. (1994b). Development of ARGUS: past, present, future. Rapport, Statistics Netherlands, Voorburg.

De WAAL, A.G., et WILLENBORG, L.C.R.J. (1995a). Statistical disclosure control and sampling weights. Rapport, Statistics Netherlands, Voorburg.

De WAAL, A.G., et WILLENBORG, L.C.R.J. (1995b). Local suppression in statistical disclosure control and data editing. Rapport, Statistics Netherlands, Voorburg.

De WAAL, A.G., et WILLENBORG, L.C.R.J. (1995c). Optimal global recoding and local suppression. Rapport, Statistics Netherlands, Voorburg.

DUNCAN, G.T., et LAMBERT, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81, 10-28.

FULLER, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9, 383-406.

GREENBERG, B.V., et ZAYATZ, L.V. (1992). Strategies for measuring risk in public use microdata files. *Statistica Neerlandica*, 46, 33-48.

HOOGLAND, J. (1994). Protecting microdata sets against statistical disclosure by means of compound Poisson distributions (en néerlandais). Rapport. Statistics Netherlands, Voorburg.

KELLER, W.J., et BETHLEHEM, J.A. (1992). Disclosure protection of microdata: problems and solutions. *Statistica Neerlandica*, 46, 5-19.

MOKKEN, R.J., PANNEKOEK, J., et WILLENBORG, L.C.R.J. (1989). Microdata and disclosure risks, CBS Select 5, Statistical Essays, Staatsuitgeverij (La Haye), 181-200.

MOKKEN, R.J., KOOIMAN, P., PANNEKOEK, J., et WILLENBORG, L.C.R.J. (1992). Disclosure risks for microdata. *Statistica Neerlandica*, 46, 49-67.

MÜLLER, W., BLIEN, U., KNOCHE, P., WIRTH, H. et coll. (1991). *The Factual Anonymity of Microdata* (en allemand). Stuttgart: Metzler-Poeschel Verlag.

PAASS G., et WAUSCHKUHN, U. (1985). Data access, data protection and anonymization - analysis potential and identifiability of anonymized individual data (en allemand). Gesellschaft für Mathematik und Datenverarbeitung, Oldenbourg-Verlag, Munich.

PAASS, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Studies*, 6, 487-500.

PANNEKOEK, J. (1992). Disclosure control of extreme values of continuous identifiers (en néerlandais). Rapport, Statistics Netherlands, Voorburg.

PANNNEKOEK, J. (1995). Statistical methods for some simple disclosure limitation rules. Rapport, Statistics Netherlands, Voorburg.

PANNEKOEK, J., et de WAAL, A.G. (1995). Synthetic and combined estimators in statistical disclosure control. Rapport, Statistics Netherlands, Voorburg.

PIETERS, A.J., et De WAAL, A.G. (1995). A demonstration of ARGUS. Rapport, Statistics Netherlands, Voorburg.

SKINNER, S., MARSH, C., OPENSHAW, S., et WYMER, C. (1990). Disclosure avoidance for census microdata in Great Britain. *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 131-143.

SKINNER, C.J. (1992). On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica*, 46, 21-32.

SKINNER, C.J., et HOLMES, D.J. (1992). Modelling population uniqueness. International Seminar on Statistical Confidentiality, Dublin.

US DEPARTMENT OF COMMERCE (1978). Report on statistical disclosure and disclosure avoidance techniques. Statistical Policy Working Paper 2, Washington DC.

Van GELDEREN, R. (1995). ARGUS: Statistical disclosure control of survey data. Rapport, Statistics Netherlands, Voorburg.

VERBOON, P. (1994). Some ideas for a masking measure for statistical disclosure control. Rapport, Statistics Netherlands, Voorburg.

VERBOON, P., et WILLENBORG, L.C.R.J. (1995). Comparing two methods for recovering population uniques in a sample. Rapport. Statistics Netherlands, Voorburg.

WILLENBORG, L.C.R.J. (1990a). Remarks on disclosure control of microdata. Rapport, Statistics Netherlands, Voorburg.

WILLENBORG, L.C.R.J. (1990b). Disclosure risks for microdata sets: stratified populations and multiple investigators. Rapport, Statistics Netherlands, Voorburg.

WILLENBORG, L.C.R.J. (1993). Discussion statistical disclosure limitation. *Journal of Official Statistics*, 9, 469-474.

WILLENBORG, L.C.R.J., MOKKEN, R.J., et PANNEKOEK, J. (1990). Microdata and disclosure risks. *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, Washington DC, 167-180.

CONTENTS TABLE DES MATIÈRES

Volume 24, No. 1, March/mars 1996

Mikelis BICKIS, Susana BLEUER et Daniel KREWSKI
On the estimation of the proportion of positives in a sequence of screening experiments 1

Hugh CHIPMAN
Bayesian variable selection with related predictors 17

Lawrence JOSEPH, Alain C. VANDAL et David B. WOOLFSON
Estimation in the multi-path change-point problem 37

André Robert DABROWSKI et Herold DEHLING
Estimating conditional occupation time distributions for dependent sequences 55

Douglas P. WIENS
Robust sequential designs for approximately linear models 67

Norbert HENZE
Empirical distribution function goodness of fit tests for discrete models 81

Denna B. HAUNSPERGER
Paradoxes in nonparametric tests 95

Peter Yi-Shi SHAO et William E. STRAWDERMAN
Improving on truncated linear estimates of exponential and gamma scale variates 105

Samuel BAYOMOG, Xavier GUYON, Cécile HARDOUIN et Jianfeng YAO
Test de différence de contrastes et somme pondérée de khideux 115

Yoav BENJAMIN et Abba M. KRIEGER
Concepts and measures for skewness with data analytic implications 131

Volume 24, No. 2, June/juin 1996

N. REID
Likelihood and higher order approximations to tail areas: a review and annotated bibliography 141

Jiahua CHEN et J.D. KALBFLEISCH
Penalized minimum distance estimates in finite mixture models 167

Brajendra C. SUTRADHAR et J.N.K. RAO
Estimation of regression parameters in generalized linear models for clustered correlated data with measurement error 177

Bing LI et Ruben H. ZAMAR
M-estimates of regression when scale is unknown and the error distribution is possibly asymmetric: a minimax results 193

R. B. ARELLANO-VALLE, H. BOLFARINE et F. VILCA-LABRA
Ulstrastructural elliptical models 207

R.L. DYKSTRA, Hammou EL BARMI, James M. GUFFEY et F.T. WRIGHT
Nonhomogeneous Poisson processes as overhaul models 217

Edit GOMBAY
The weighted sequential likelihood ratio 229

Feng-Gin SUN, Jean-Yves LARAMÉE et John S. RAMBERG
On Spiring's normal loss function 241

Lee J. BAIN et Gaoxiong GAN
Conditional maxima and inferences for the truncated exponential distribution 251

Y. YANG, A.K. GUPTA et T.T. NGUYEN
Characterization theorems for some discrete distributions based on conditional structure 257

K. VIRASWAMI et N. REID
Higher order asymptotics under model misspecification 263

JOURNAL OF OFFICIAL STATISTICS

An International Quarterly Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey Methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 11, Number 4, 1995

Comparison Between Maximum Likelihood and Bayes Methods for Estimation of Binominal Probability with Sample Compositing
Yogendra P. Chaubey and Weiming Li 379

Editing Statistical Records by Neural Networks
Svein Nordbotten 391

Computer-assisted Personal Interviewing: An Experimental Evaluation of Data Quality and Costs
Reginald P. Baker, Norman M. Bradburn, and Robert A. Johnson 413

Questionnaire Effects on Measurements of Race and Spanish Origin
Nancy Bates, Elizabeth A. Martin, Theresa J. DeMaio, and Manuel de la Puente 433

A Content Analysis of Advance Letters from Expenditure Surveys of Seven Countries
Martin Luppes 461

In Other Journals 481

Special Note 483

Editorial Collaborators 485

Index to Volume 11, 1995 491

Lars Lyberg, U/LEDN, Statistics Sweden, S-115 81 Stockholm, Sweden

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points suivants:

1. Présentation

1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
1.4 Les remerciements doivent paraître à la fin du texte.
1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.
3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O, 0; 1, l).
3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).
5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.